Atti del IX Convegno Annuale
dell'Associazione per l'Informatica
Umanistica e la Cultura Digitale
(AIUCD)

LA SVOLTA INEVITABILE:
SFIDE E PROSPETTIVE PER L'INFORMATICA UMANISTICA

15 – 17 gennaio 2020
Milano
Università Cattolica del Sacro Cuore

A CURA DI:
Cristina Marras
Marco Passarotti
Greta Franzini
Eleonora Litta

ISBN: 978-88-942535-4-2

Cristina Marras, Marco Passarotti, Greta Franzini, Eleonora Litta (a cura di),
Atti del IX Convegno Annuale AIUCD. La svolta inevitabile:
sfide e prospettive per l'Informatica Umanistica.

Available online as a supplement of Umanistica Digitale: https://umanisticadigitale.unibo.it

Prefazione

La nona edizione del convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD 2020; Milano, 15-17 gennaio 2020) ha come tema "La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica", con lo specifico obiettivo di fornire un'occasione per riflettere sulle conseguenze della crescente diffusione dell'approccio computazionale al trattamento dei dati connessi all'ambito umanistico. Questo volume raccoglie gli articoli i cui contenuti sono stati presentati al convegno. A diversa stregua, essi affrontano il tema proposto da un punto di vista ora più teorico-metodologico, ora più empirico-pratico, presentando i risultati di lavori e progetti (conclusi o in corso) che considerino centrale il trattamento computazionale dei dati.

Dunque, la svolta inevitabile qui a tema va intesa innanzitutto come metodologica e, più nello specifico, computazionale. Ad essa la ricerca umanistica contemporanea assiste, con diversi gradi di accoglienza, critica addirittura rifiuto. La computabilità del dato empirico (anche) in area umanistica è, infatti, il tratto distintivo e il vero valore aggiunto che le innovazioni tecnologiche degli ultimi decenni hanno comportato in questo ambito. Nonostante negli anni il settore delle cosiddette Digital Humanities si sia voluto caratterizzare, anche a partire dalla propria denominazione, insistendo maggiormente sull'aspetto digitale che non su quello computazionale, i tempi sembrano ormai maturi perché il termine Computational Humanities, o il troppo precocemente accantonato Humanities Computing, (ri)prenda il posto oggi ancora occupato da Digital Humanities.[1] Digitale è, infatti, il formato dei dati con cui attualmente si ha in gran parte a che fare nel nostro settore: ma è computazionale l'uso che di questi dati si fa ed è un fatto che gran parte dei lavori prodotti nell'area delle Digital Humanities consista nel "fare conti" sui dati.[2]
Come tanti suoi predecessori, anche il formato digitale passerà; mentre il metodo, e la svolta che esso comporta, resterà, perché solidamente ancorato all'evidenza empirica del dato che è il punto di partenza e, quindi, il centro di analisi di molta ricerca umanistica. Per questa ragione, la svolta computazionale nelle scienze umanistiche è innanzitutto metodologica: a cambiare radicalmente non è tanto il formato dei dati, ma il modo con cui ad essi ci si approccia e l'uso che di essi si fa.
Non va negato un certo scetticismo reazionario che, ora esplicito, ora sottaciuto, parte del mondo della ricerca umanistica nutre nei confronti dei metodi e degli strumenti che la svolta computazionale ha messo a disposizione di noi ricercatori, che viviamo l'attuale scorcio di storia della scienza. Negli anni, tale scetticismo ha alimentato una irragionevole distinzione, e conseguente separazione, tra umanisti "tradizionali" e umanisti "digitali", quasi che si debbano identificare due aree al fine di evitare che gli uni infastidiscano troppo gli altri con le proprie ricerche, trascurando che esse trattano i medesimi oggetti e hanno quale fine comune la produzione di nuova conoscenza.
Siffatta separazione è dovuta a errori imputabili all'una e all'altra parte. Da un lato, certi umanisti "digitali" tendono a produrre ricerca che rischia di scadere nella superficialità, assumendo che l'alta quantità dei dati trattati possa compensarne l'eventuale bassa qualità e dimenticando, così, che le ricerche di area umanistica molto raramente lavorano su Big Data e non possono (anzi, non vogliono) accontentarsi di tendenze percentuali fondate su dati imprecisi. Dall'altro lato, i "tradizionali" sono spesso afflitti da un conservatorismo protezionista incompatibile con la natura stessa del lavoro di ricerca, che è in sé progressivo e in costante evoluzione. Ne consegue un dialogo interrotto tra le due parti: i "digitali" sono considerati dei tecnici (inteso in senso riduttivo) che brutalizzano il delicato dato umanistico, mentre i "tradizionali" vengono derubricati a dinosauri incartapecoriti che ormai non hanno più niente di nuovo da dire.
Ma la svolta computazionale non è né "digitale", né "tradizionale". Semplicemente, essa è inevitabile. Chi ne fa cattivo uso, come certo mondo "digitale", non sa valorizzarne la forza della portata; chi la rifiuta a priori, si pone fuori dalla realtà e, volutamente ignorando il nuovo, ferisce la ragione stessa del far ricerca.

[1] Una valida sintesi della questione relativa alla denominazione del settore, con una buona bibliografia a supporto, è riportata in un articolo di Leah Henrikson pubblicato su 3:AM Magazine (24 Ottobre 2019) e disponibile presso https://www.3ammagazine.com/3am/humanities-computing-digital-humanities-and-computational-humanities-whats-in-a-name/

[2] Da, Nan Z. "The computational case against computational literary studies." *Critical Inquiry* 45.3 (2019): 601-639.

Ma resta che la svolta è inevitabile: non si comprende perché sul tavolo dell'umanista del 2020 non possano trovarsi al contempo un'edizione critica cartacea e i risultati di un analizzatore morfologico automatico proiettati sullo schermo di un computer. Entrambi sono strumenti che diversamente trattano il comune oggetto d'interesse di tanta ricerca, ovvero i dati.

Ma di una svolta non solo metodologica questa edizione 2020 del convegno AIUCD vuole trattare e farsi carico, aspirando anzi a mettere in atto anche una piccola, ma sostanziale svolta organizzativa. Per la prima volta, la call for papers di un convegno dell'Associazione, ha richiesto l'invio non di abstract, ma di articoli completi della lunghezza di un massimo di 4 pagine (bibliografia esclusa). Di concerto con il Comitato Direttivo dell'Associazione, abbiamo deciso di orientarci in tal senso per due ragioni principali. Primo, crediamo che, giunto alla propria nona edizione, il convegno annuale della AIUCD sia ormai sufficientemente maturo per passare a una fase il cui obiettivo sia quello di accogliere nel programma del convegno proposte che nel formato dell'articolo completo consentissero ai revisori una valutazione piena e più accurata. Ciò si lega anche alla seconda ragione. Il nostro settore come è noto è molto veloce: i dati (e i risultati su di essi basati) tendono a cambiare nel giro di poco tempo. Ricevere articoli completi ci ha consentito di mettere i contenuti del presente volume nelle mani dei partecipanti (e più in generale della comunità tutta) il primo giorno del convegno, fornendo così una realistica fotografia dello stato dei lavori al gennaio 2020.

Tutti gli articoli selezionati per essere presentati al convegno hanno cittadinanza in questo volume. Anche questa è una svolta: diversamente dall'uso fino ad oggi adottato, gli articoli pubblicati non sono più il risultato di una selezione a posteriori rispetto al convegno, ma tutti quelli effettivamente apparsi nel programma di AIUCD 2020. In tal senso, una certa esclusività promossa a livello di selezione scientifica si fa inclusività in termini di pubblicazione e, dunque di visibilità dei lavori presentati. Ogni proposta è stata valutata da tre revisori; si è dovuto ricorrere a una quarta valutazione solo nel caso di due proposte su cui i tre revisori avevano espresso opinioni che rendevano difficile prendere una decisione in merito alla loro accettazione, o meno. Al proposito delle differenze tra i revisori, abbiamo constatato divergenze piuttosto frequenti e, in alcuni casi, nette tra coloro che provengono dall'area linguistico-computazionale e quanti, invece, sono a vario titolo legati ai diversi settori dell' "umanistica digitale". Mentre i linguisti computazionali sono tradizionalmente usi a valutare articoli completi e tendono a richiedere che i contenuti di essi descrivano motivazioni, metodi e risultati (preferibilmente replicabili) di lavori di ricerca in corso, o completati, i revisori di area umanistico-digitale sono disposti a valutare positivamente anche idee e proposte che ancora non si siano incarnate in una reale applicazione ai dati. La constatazione di tale diversità è il risultato della composizione volutamente inter- e trans-disciplinare del comitato dei revisori, a rappresentare la natura trasversale di AIUCD e, di riflesso, del suo convegno annuale. Nel prendere le decisioni in merito alle proposte, abbiamo cercato un equilibrio tra gli atteggiamenti delle due parti, favoriti dall'avere a disposizione un livello di dettaglio sul lavoro descritto. La richiesta di articoli completi ha avuto un impatto non molto rilevante sul numero delle proposte inviate, che sono state 71, di cui 67 sottoposte al processo di revisione, mentre 4 sono state escluse perché non confacenti ai criteri richiesti dalla call for papers (tra cui anonimato e originalità). Alla precedente edizione del convegno AIUCD (Udine, 23-25 gennaio 2019) erano state inviate 82 proposte, di cui 75 sottoposte a revisione. Conseguenze più sostanziali si sono, invece, riscontrate sulla percentuale delle proposte accettate e rifiutate. Delle 67 proposte valutate, 45 sono state accettate per apparire nel programma del convegno e, quindi, in questo volume, mentre 22 sono state rifiutate, risultando così in una percentuale di accettazione pari al 67.16%. All'edizione udinese, la percentuale si era attestata intorno all'84%. La contrazione del numero di proposte accettate è strettamente connessa alla richiesta di articoli completi invece che di abstract.

Il programma del convegno ha incluso due sessioni poster. Dei 45 contributi accettati, 21 sono stati giudicati adatti alla presentazione in modalità poster. Rispetto alle consuetudini del settore, che tende a relegare le proposte meno interessanti o più problematiche nelle sessioni poster, abbiamo deciso di assegnare la modalità di comunicazione in forma di poster non secondo la qualità, ma piuttosto in base alla tipologia della proposta. Dunque, tendenzialmente le proposte che presentano lavori che hanno portato a risultati pratici (come strumenti, risorse, o interfacce) sono state giudicate più adatte a una presentazione in formato poster, mentre le discussioni teoriche, disciplinari, o metodologiche hanno occupato le sessioni di comunicazioni orali. Resta che non sussiste differenza alcuna in termini di selezione qualitativa tra un articolo i cui contenuti sono stati

presentati al convegno in forma orale, o in forma di poster, come dimostra l'aver riservato il medesimo numero di pagine a tutti gli articoli presenti in questo volume.

I contenuti dei testi qui raccolti in ordine alfabetico testimoniano la varietà dei temi che usualmente sono trattati nei convegni della AIUCD. Essi spaziano da riflessioni generali sui settore di ricerca alla realizzazione di nuove risorse linguistiche e strumenti di analisi dei dati, da lavori di filologia ed editoria digitale a temi connessi alla digitalizzazione delle fonti in ambito bibliotecario. Oltre alla presentazione dei contenuti degli articoli di questo volume, il programma del convegno ha previsto tre relazioni su invito (una per ciascuno dei tre giorni della sua durata), che sono state rispettivamente presentate da Roberto Navigli (Sapienza, Università di Roma), Julianne Nyhan (University College London) e Steven Jones (University of South Florida).

Il contributo di Roberto Navigli, intitolato *Every time I hire a linguist my performance goes up (or: the quest for multilingual lexical knowledge in a deep (learning) world)*, è un esempio di ricerca che dice della ineludibilità del legame e, auspicabilmente, della collaborazione tra mondo scientifico e mondo umanistico e, nello specifico, tra la comunità che si riconosce nella AIUCD e quella della linguistica computazionale. Gli interventi di Julianne Nyhan (*Where does the history of the Digital Humanities fit in the longer history of the Humanities? Reflections on the historiography of the 'old' in the work of Fr Roberto Busa S.J.*) e Steven Jones (*Digging into CAAL: Father Roberto Busa's Center and the Prehistory of the Digital Humanities*) si posizionano nell'alveo della storia della disciplina, particolarmente riferendo in merito ai loro studi sulle attività di padre Roberto Busa. La figura di Busa è strettamente legata all'Università Cattolica del Sacro Cuore di Milano, dove a partire dalla fine degli anni settanta il gesuita tenne un corso di Linguistica Computazionale e Matematica e fondò un gruppo di ricerca che, nel 2009, fu trasformato in un Centro di Ricerca; quel CIRCSE che con l'AIUCD ha organizzato il convegno annuale dell'associazione di cui questo volume raccoglie gli Atti. Nel 2010, un anno prima di lasciarci, padre Busa volle donare alla Biblioteca della Cattolica il proprio archivio personale. Una ricchissima documentazione del lavoro di Busa e della sua diffusione, oltre che delle sue relazioni personali e professionali (ricostruibili attraverso il vasto epistolario), l'Archivio Busa è attualmente in fase di catalogazione e digitalizzazione da parte della Biblioteca d'Ateneo. Una selezione di materiale tratto dall'Archivio è stata resa direttamente accessibile ai partecipanti dell'edizione milanese del convegno AIUCD in una piccola mostra allestita nell'atrio dell'aula dei lavori congressuali. Le teche della mostra raccolgono fogli di lavoro, lettere, schede perforate, nastri e articoli di quotidiani che trattano del lavoro di padre Busa: una forma di ringraziamento che l'Università Cattolica, il CIRCSE e la comunità scientifica tutta vuole riservare a uno dei pionieri dell'analisi linguistica automatica.

I nostri ringraziamenti vanno innanzitutto alla Presidente di AIUCD Francesca Tomasi e a Fabio Ciotti, che in quel ruolo l'ha preceduta, per aver scelto Milano quale sede dell'edizione 2020 del convegno. Da loro è venuto il primo, fondamentale, sostegno alla "svolta organizzativa" di cui abbiamo voluto farci portatori. Ringraziamo altresì il Consiglio Direttivo dell'Associazione, il Comitato di Programma e tutti i revisori, che hanno lavorato alacremente per metterci nelle condizioni di definire il miglior programma possibile. La sede milanese dell'Università Cattolica del Sacro Cuore ci ha supportato a livello amministrativo e logistico; teniamo particolarmente a ringraziare l'Ufficio Formazione Permanente, nello specifico di Elisa Ballerini, la Biblioteca d'Ateneo, e specificatamente Paolo Senna, che ci ha messo a disposizione i materiali dell'Archivio Busa, l'Ufficio Eventi e la Direzione di Sede, che hanno fornito gli spazi per il convegno. Grazie soprattutto a chi ha inviato proposte, ai relatori e ai partecipanti tutti, perché sono loro i protagonisti essenziali dell'evento.

La nostra speranza è che il lavoro fatto sia utile ancora prima che apprezzato. E che i suoi risultati si mantengano nelle edizioni a venire, con l'obiettivo di migliorare sempre, guardando avanti; perché saper vedere le svolte e affrontarle è la ragione stessa della ricerca.

Cristina Marras
Marco Passarotti
Greta Franzini
Eleonora Litta

Chair e Comitati

General Chair

- Cristina Marras

Chair del comitato scientifico e di programma

- Marco Passarotti

Comitato scientifico e di programma

- Maristella Agosti
- Stefano Allegrezza
- Federica Bressan
- Cristiano Chesi
- Fabio Ciracì
- Greta Franzini
- Angelo Mario Del Grosso
- Eleonora Litta
- Pietro Maria Liuzzo
- Federico Meschini
- Johanna Monti
- Federico Nanni
- Marianna Nicolosi
- Dario Rodighiero
- Marco Rospocher
- Chiara Zuanni

Comitato Organizzatore

- Greta Franzini
- Eleonora Litta

Indice dei Contenuti

EcoDigit
Ecosistema Digitale per la fruizione e la valorizzazione dei beni e delle attività culturali del Lazio

Luigi Asprino
STLab, ISTC-CNR
luigi.asprino@istc.cnr.it

Antonio Budano
INFN Sezione di Roma Tre
antonio.budano@infn.it

Marco Canciani
RilTec, Università di RomaTre
marco.canciani@uniroma3.it

Luisa Carbone
Università della Tuscia
luisa.carbone@unitus.it

Miguel Ceriani
DIAG, Sapienza Università di Roma
ceriani@diag.uniroma1.it

Ludovica Marinucci
STLab, ISTC-CNR
ludovica.marinucci@istc.cnr.it

Massimo Mecella
DIAG, Sapienza Università di Roma
mecella@diag.uniroma1.it

Federico Meschini
Università della Tuscia
fmeschini@unitus.it

Marialuisa Mongelli
ENEA
marialuisa.mongelli@enea.it

Andrea Giovanni Nuzzolese
STLab, ISTC-CNR
andreagiovanni.nuzzolese@cnr.it

Valentina Presutti
LILEC, Università di Bologna STLab, ISTC-CNR
valentina.presutti@cnr.it

Marco Puccini
ENEA
marco.puccini@enea.it

Mauro Saccone
Università di RomaTre
mauro.saccone@uniroma3.it

Abstract

English. EcoDigit is one of the projects of the "Centro di Eccellenza DTC Lazio", which intends to aggregate and integrate expertises in the field of technologies applied to cultural heritage. EcoDigit aims at enriching the Anagrafe delle Competenze, another project of the DTC Lazio, with a middleware platform that is able to facilitate the integration of new data sources and to allow the publication and reuse of services for the enhancement and fruition of the cultural heritage in the Lazio region. The project is designed to integrate the collected data through open formats and semantic technologies based on Linked Data model.

Italiano. EcoDigit è uno dei progetti del "Centro di Eccellenza DTC Lazio", che ha l'obiettivo di aggregare e integrare varie competenze nell'ambito del patrimonio culturale digitale. EcoDigit intende aricchire il sistema dell'Anagrafe delle Competenze, altro progetto del DTC Lazio, con una piattaforma middleware che faciliti l'integrazione di nuove risorse di dati e permetta la pubblicazione e riuso di servizi per la valorizzazione e fruizione del patrimonio culturale nella regione Lazio. Il progetto è finalizzato all'integrazione dei dati censiti attraverso formati aperti e tecnologie semantiche basate sul modello Linked data.

1 Introduzione

Il progetto *Ecosistema Digitale per la fruizione e la valorizzazione dei beni e delle attività culturali del Lazio* (EcoDigit)[1] è una delle iniziative del *Centro di Eccellenza del Distretto Tecnologico per i beni e le attività Culturali* (DTC)[2], costituito dalle cinque università statali del Lazio - Sapienza, Tor Vergata, Roma Tre, Cassino e Tuscia - in rete con CNR, ENEA e INFN per aggregare e integrare competenze nel settore delle tecnologie per i beni e le attività culturali. Il progetto (luglio 2018 - gennaio 2020) si rivolge alle organizzazioni operanti nel settore della cultura le quali producono e mantengono basi di dati anche di grandi dimensioni archiviati usando formati, modelli e processi diversi. Questo fenomeno comporta la necessità di identificare processi, tecnologie e modelli di integrazione semplici e sostenibili che consentano la fruizione del patrimonio in modo globale e collegato.

Le tecnologie semantiche, e in particolare i Linked Open Data (LOD), sono state ampiamente sfruttate con successo nel campo del patrimonio culturale al fine di migliorare l'accesso e l'esperienza di fruizione dei beni culturali da parte dei cittadini, così come di facilitare la reperibilità, l'integrazione e l'arricchimento dei dati (Dijkshoorn et al., 2018; Daquino et al., 2017; Lodi et al., 2017). Infatti, il paradigma dei LOD è utilizzato per collegare dati provenienti da diverse istituzioni culturali, aumentando così la possibilità di raggiungere i dati culturali disponibili nel Web of Data. L'interconnessione dei contenuti delle organizzazioni collaboratrici ha anche contribuito ad arricchire le informazioni in modo efficace e finalizzato alla valorizzazione del patrimonio culturale (Hyvönen, 2009). La collaborazione tra organizzazioni culturali ha portato anche allo sviluppo collaborativo di ontologie che descrivono il patrimonio culturale a livello internazionale, ad esempio CIDOC-Conceptual Reference Model (CRM) (Doerr, 2003)[3], in modo tale che i requisiti di interoperabilità semantica potessero essere soddisfatti all'interno dei loro sistemi. Inoltre, l'uso di ontologie comuni ha facilitato lo scambio di dati e la creazione di enormi biblioteche digitali, ad esempio l'Europeana Data Model (EDM) (Isaac and Haslhofer, 2013)[4].

EcoDigit, inoltre, non solo ha preso ispirazione da esperienze già consolidate sia degli enti partner e non del progetto sia maturate nell'ambito delle Pubbliche Amministrazioni (PA), ma intende anche instaurare collaborazioni con tali realtà. Ad esempio, il progetto ReCAP[5] dell'Università Sapienza ha creato una rete di condivisione di conoscenze, strumenti e sperimentazioni che permette di elaborare linee guida e modelli per la costruzione di processi conservativi del patrimonio digitale. Inoltre, il *Sacher Project*[6] sta affrontando problematiche simili, ma non a livello di progettazione di una piattaforma regionale. Molte istituzioni regionali e nazionali che gestiscono il nostro patrimonio culturale stanno adottando il modello dei dati aperti seguendo le linee guida dell' *Agenzia per l'Italia Digitale* (AgID); tuttavia per le organizzazioni pubbliche e private non è semplice adeguarsi per la mancanza di una piattaforma che ne faciliti il compito. La missione del progetto *Data & Analytics Framework* (DAF)[7], di cui il network di ontologie OntoPia[8] è uno dei risultati, è analoga a quella di EcoDigit, ma insiste sull'intero territorio nazionale con focus sull'interoperabilità tra dati tra PA. In tale contesto, EcoDigit può contribuire a rendere efficace e coordinata l'integrazione con i sistemi nazionali.

2 Obiettivi

EcoDigit ha l'obiettivo di arricchire il sistema *Anagrafe delle Competenze*[9] con una piattaforma middleware che faciliti l'integrazione di nuove sorgenti di dati e consenta la pubblicazione e il riuso di servizi per la valorizzazione e la fruizione del patrimonio culturale del Lazio. Nello specifico, EcoDigit fornisce (i) l'architettura di riferimento per l'integrazione di servizi modulari e per la loro pubblicazione e il riuso; (ii) una componente software sviluppata in forma prototipale per l'orchestrazione dei servizi, l'integrazione

[1] http://ecodigit.dtclazio.it/
[2] https://dtclazio.it/
[3] http://cidoc-crm.org/
[4] https://pro.europeana.eu/page/edm-documentation
[5] http://digilab.uniroma1.it/attivit/recap
[6] http://www.sacherproject.com/progetto
[7] https://teamdigitale.governo.it/it/projects/daf.htm
[8] https://github.com/italia/daf-ontologie-vocabolari-controllati
[9] http://dtc.si.cnr.it/anagrafe-delle-competenze

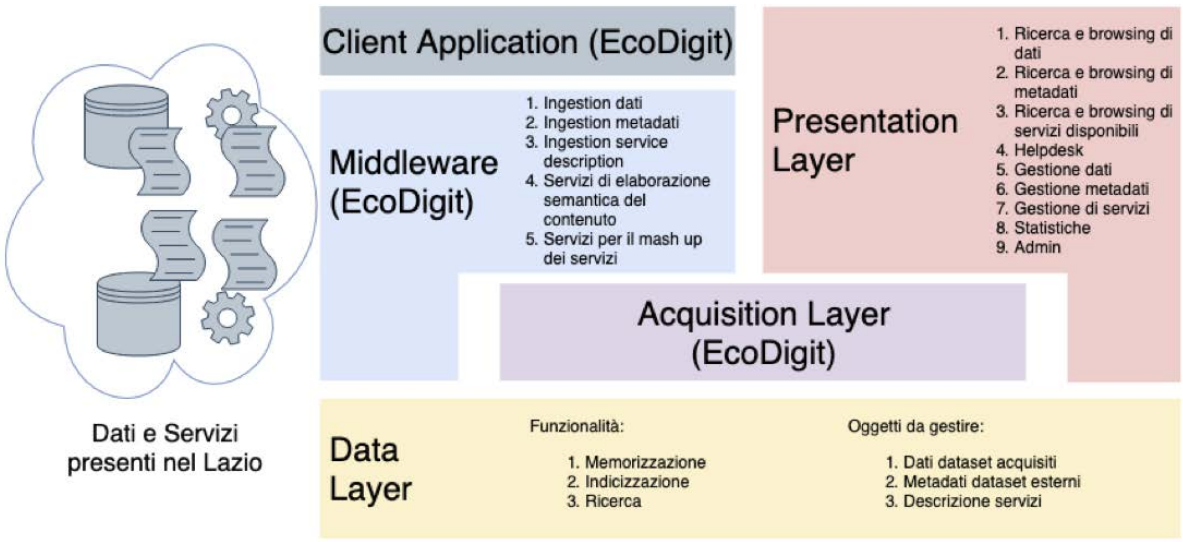


Figura 1: Schema architetturale in cui si inserisce il Middleware di EcoDigit.

e l'aggregazione delle interfacce e dei dati; (iii) la versione prototipale di servizi orientati alla fruizione e valorizzazione del patrimonio culturale.

Con queste caratteristiche EcoDigit si configura come un progetto di ricerca e trasferimento tecnologico collegato in maniera significativa agli altri progetti del DTC Lazio. Con Anagrafe è in relazione fornendo un livello software intermedio, detto middleware, che consente al sistema di aggregare nuove sorgenti di dati, servizi, strumenti innovativi sia industriali che accademici. In questa prospettiva EcoDigit estende il sistema Anagrafe e gli conferisce la capacità di evolvere ed essere esteso. Per gli altri progetti di ricerca, EcoDigit svolge il ruolo di mediatore nei confronti del sistema Anagrafe. Ciò significa che i risultati dei vari progetti potranno essere integrati grazie all'interfacciamento con il middleware di EcoDigit.

In Figura 1 è rappresentato uno schema architetturale di alto livello in cui si inserisce il middleware di competenza di EcoDigit. Sulla sinistra sono rappresentate le risorse, ovvero le banche dati (ciascuna col il proprio formato sintattico e modello concettuale di rappresentazione dei dati) e i servizi messi a disposizione dai vari enti operanti sul territorio che, però, mancano di un punto di accesso unico. Ciò ostacola potenziali utenti interessati a consultare le risorse o creare applicativi basati su esse.

Il sistema, a cui stanno lavorando congiuntamente i gruppi di EcoDigit e Anagrafe, ha l'obiettivo di superare questa frammentarietà. Esso è composto da vari livelli: subito al di sopra del *Data Layer*, c'è (i) l'*Acquisition Layer*, a cui lavorano congiuntamente Anagrafe ed EcoDigit, che si occupa di creare dei flussi di dati i quali saranno acquisiti, uniformati secondo uno schema comune e memorizzati dal sistema. Questi flussi dati verranno creati dal (ii) *Middleware*, che offre alle applicazioni client delle funzionalità per l'elaborazione semantica dei contenuti, facilities per il mashup dei servizi indicizzati dal sistema, oltre a definire le linee guida che servizi e dataset esterni dovranno seguire per garantire l'interoperabilità con il sistema. Nella parte più alta, (iii) il *Client Application* rappresenta una qualunque applicazione che intende utilizzare i dati acquisiti e offerti dal sistema. Essa potrà, attraverso il Middleware, interrogare il sistema per raccogliere dati, recuperare i metadati di dataset esterni o descrizioni dei servizi disponibili.

3 Metodologia

Gli obiettivi sovraesposti sono stati realizzati attraverso l'esecuzione di attività svolte in maniera collaborativa e sinergica tra i partner del progetto e sintetizzabili in tre fasi: 1. Censimento delle risorse relative ai beni culturali presenti nella regione Lazio (cf. Sezione 3.1); 2. Elaborazione del modello di integrazione delle sorgenti nel Middleware (cf. Sezione 3.2); 3. Realizzazione di prototipi volti a verificare la validità dei risultati e dell'approccio seguito (cf. Sezione 3.3).

Tali attività sono state svolte considerando una prospettiva almeno quinquennale di sostenibilità.

3.1 Attività di censimento delle risorse presenti nel Lazio

Fondamentali nel corso del progetto sono state tre tipi di attività preliminari di censimento consistenti in:

(a) un'analisi dei requisiti del sistema, con relativa ricognizione degli attori e dei casi d'uso, e dello stato dell'arte su tecnologie middleware e architetture per l'integrazione di servizi;

(b) l'individuazione delle sorgenti dei dati presenti nel Lazio, e dei modelli ontologici e tecniche per la loro integrazione e standardizzazione basata su formati aperti e semantici. In questo contesto, tramite la formalizzazione e disseminazione è stato possibile anche individuare potenziali stakeholder del progetto;

(c) una valutazione dei tool allo stato dell'arte per la creazione e la gestione di ambienti virtuali 2D/3D.

3.2 Modello di integrazione di una sorgente

Il modello di dati e metadati, che le sorgenti devono rispettare per poter essere acquisite dal middleware EcoDigit, si basa su metodi e tecniche di "metadatazione" e di Semantic Web, comprendendo anche tecnologie proprie delle openAPI e tutto ciò che viene comunemente classificato come Open Data.

Per la sua elaborazione, si è proceduto a un'analisi delle sorgenti censite nel Lazio al fine di evidenziare tanto i domini di conoscenza coperti dalle sorgenti quanto il dettaglio dei vari campi che il modello deve rappresentare. Per ognuno di essi sono stati ricercati gli schemi concettuali considerati standard di riferimento per la modellazione dei dati inerenti a un certo dominio, quali FOAF[10], DOAP[11], Org Ontology[12], OntoPia network, SPAR Ontologies[13], ArCo[14], ecc. Il modello deve essere pensato come l'unione di standard tecnologici, schemi concettuali e di metadatazione allo stato dell'arte per i vari domini di conoscenza, relativi in particolare all'ambito dei beni culturali.

Successivamente, quando i modelli esistenti allo stato dell'arte sono stati ritenuti non in grado di rappresentare semanticamente campi peculiari presenti nei dataset di input, è stata effettuata una modellazione ex novo, utilizzando una metodologia di ingegneria ontologica (Blomqvist et al., 2016), basata su un'estensione di *eXtreme Design* (XD) (Blomqvist et al., 2010). XD è un metodo di progettazione agile di ontologie che si basa sul riuso di *Ontology Design Patterns* (ODP) (Gangemi and Presutti, 2009) al fine di risolvere problemi di modellazione ontologica noti e ricorrenti.

Il workspace del progetto EcoDigit è disponibile su Github[15]. Di seguito, si elencano le ontologie definite ex-novo nel corso del progetto:

- *Ontologia delle Organizzazioni*[16] lo scopo di definire un vocabolario condiviso ditermini per la descrizione delle organizzazioni che partecipano al Centro di Eccellenza-DTC Lazio. Essa estende: OntoPiA-COV, FOAF, W3C's Organization Ontology.

- *Ontologia delle Valutazioni*[17] ha lo scopo di definire un vocabolario di termini per la descrizione di qualsiasi cosa abbiamo una valutazione associata che è espressa rispettouna certa scala.

- *Ontologia delle Esperienze e Competenze*[18] ha lo scopo di definire un vocabolario condiviso di termini per la descrizione delle esperienze e competenze di una persona. Estende: OntoPiA-CPV, OntoPiA-COV, FOAF, BIBO; importa l'Ontologia delle Valutazioni.

[10] http://xmlns.com/foaf/spec/
[11] http://usefulinc.com/ns/doap#
[12] https://www.w3.org/TR/vocab-org/
[13] http://www.sparontologies.net/ontologies
[14] http://w3id.org/arco
[15] https://github.com/ecodigit/workspace
[16] https://w3id.org/ecodigit/ontology/organization
[17] https://w3id.org/ecodigit/ontology/grade
[18] https://w3id.org/ecodigit/ontology/eas/

3.3 Prototipi

A seguito delle attività sovraesposte di censimento dei dataset, degli schemi concettuali e dei tool allo stato dell'arte per la fruizione del patrimonio culturale, sono in fase di elaborazione due prototipi:

- una *Proof-of-Concept*, volta a mostrare la validità del modello attraverso l'integrazione di alcune delle sorgenti identificate nel task di censimento. L'obiettivo principale è quello di creare una best practice che mostri sia la semplicità dell'approccio di integrazione, e la sua scalabità nel tempo, sia gli strumenti hardware e software che una sorgente deve adottare per aderire al modello di ingresso di EcoDigit;

- un prototipo di un servizio avanzato per la fruizione dei beni culturali nel dominio della formazione. Si sta lavorando ad una soluzione che permetta di raggiungere e fruire da una singola interfaccia i dati acquisiti dal sistema. I partner del progetto, sulla base delle loro expertise, hanno collaborato alla definizione di una tassonomia di categorie che popolino l'interfaccia utente a fini didattici. Anche questo prototipo fa uso di tecnologie semantiche per migliorare la ricerca delle risorse e collegarle ai contenuti di carte tematiche (GIS based) e ricostruzioni 3D per permettere la fruizione in ambiente virtuale dei beni mostrati dall'educatore. In particolare, sarà possibile gestire completamente modelli 3D attraverso l'implementazione ad hoc (cfr. il Workspace GitHub di EcoDigit[19]) del software 3DHOP[20] e di mostrare dati cartografici tematizzati grazie all'uso della libreria Openlayer[21].

I due prototipi utilizzano e rappresentano due viste diverse sugli stessi contenuti ai quali applicare le tecniche semantiche di integrazione dei dati, indicandone la provenienza: (i) il dataset della S&TDL-Science & Technology Digital Library del CNR; (ii) il dataset della Sapienza Digital Library; (iii) il modello 3D di Porta Latina (Mura Aureliane), implementato da Roma Tre; (iv) la mappatura GIS con modelli 2D e 3D di alcune chiese della città di Viterbo, curata dalla Tuscia; (v) la ricostruzione 3D del Trono Corsini e del busto in Terracotta di Alessandro VII Chigi, fornita da ENEA; (vi) i Linked Open Data del progetto Arco-Architettura della Conoscenza[22], progettati dal CNR-ISTC in collaborazione con il MiBAC-ICCD.

4 Avanzamento tecnologico e impatto del progetto

Il modello di integrazione delle sorgenti nel quale riveste un ruolo fondamentale lo sviluppo delle tecniche di interoperabilità semantica dei vari dataset censiti nel Lazio consente l'arricchimento delle informazioni dei dati esistenti secondo una fruizione integrata. I risultati di questo studio possono essere applicati a qualsiasi progetto che potenzialmente abbia necessità di integrare sorgenti eterogenee. Data la rilevante diversità dei tipi di dati rinvenuti finora, la riusabilità delle tecniche ideate e sperimentate in EcoDigit rappresenta uno degli aspetti principali di avanzamento tecnologico prodotto. Ciò comporta la possibile partecipazione di stakeholder eterogenei alla fruizione ed evoluzione di servizi e contenuti, dando supporto all'inclusione di patrimoni già esistenti e alla loro filiera di gestione.

L'utilizzo di formati e strumenti aperti garantisce tanto la sostenibilità nel tempo quanto la diffusione massiva dei contenuti. Inoltre, il progetto stimola la creazione di nuove figure professionali e la conseguente richiesta formativa sull'uso e sulla divulgazione delle tecnologie semantiche basate sul modello dei Linked Data per la fruizione e la valorizzazione del patrimonio culturale.

Ringraziamenti

Si ringrazia i seguenti enti per il loro supporto istituzionale: la Galleria Corsini, l'Associazione CIVITA, l'Istituto per il Catalogo e la Documentazione (ICCD) del MiBAC, la Sovrintendenza xx.

[19] https://github.com/ecodigit/3dhop-react
[20] http://vcg.isti.cnr.it/3dhop/
[21] https://openlayers.org/en/latest/doc/
[22] http://dati.beniculturali.it/arco/; http://dati.beniculturali.it/

Bibliografia

Eva Blomqvist, Karl Hammar, and Valentina Presutti. 2016. Engineering ontologies with patterns - the extreme design methodology. In Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, and Valentina Presutti, editors, *Ontology Engineering with Ontology Design Patterns*, IOS Press, volume 25 of *Studies on the Semantic Web*.

Eva Blomqvist, Valentina Presutti, Enrico Daga, and Aldo Gangemi. 2010. Experimenting with extreme design. In *Proc. of EKAW 2010*. Springer, volume 6317, pages 120–134.

Marilena Daquino, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 2017. Enhancing semantic expressivity in the cultural heritage domain: Exposing the zeri photo archive as linked open data. *JOCCH* 10(4):21:1–21:21.

Chris Dijkshoorn, Lora Aroyo, Jacco van Ossenbruggen, and Guus Schreiber. 2018. Modeling cultural heritage data for online publication. *Applied Ontology* 13(4):255–271.

Martin Doerr. 2003. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine* 24(3):75–92.

Aldo Gangemi and Valentina Presutti. 2009. Ontology design patterns. In (Staab and Studer, 2009), pages 221–243.

Eero Hyvönen. 2009. Semantic portals for cultural heritage. In (Staab and Studer, 2009), pages 757–778.

Antoine Isaac and Bernhard Haslhofer. 2013. Europeana linked open data - data.europeana.eu. *Semantic Web* 4(3):291–297.

Giorgia Lodi, Luigi Asprino, Andrea Giovanni Nuzzolese, Valentina Presutti, Aldo Gangemi, Diego Reforgiato Recupero, Chiara Veninata, and Annarita Orsini. 2017. *Semantic Web for Cultural Heritage Valorisation*, Springer, pages 3–37.

Steffen Staab and Rudi Studer, editors. 2009. *Handbook on Ontologies*, International Handbooks on Information Systems. Springer.

Encoding the Critical Apparatus by Domain Specific Languages: The Case of the Hebrew Book of Qohelet

Luigi Bambaci
Department of Cultural Heritage,
University of Bologna
luigibambaci@yahoo.it

Federico Boschetti
CNR-ILC of Pisa &
VeDPH, Ca' Foscari Venezia
federico.boschetti@ilc.cnr.it

Abstract

English. Manually encoding critical apparatuses by markup languages such as TEI-XML is a non-trivial, error-prone task. It requires a technology expertise which is not within the competence of most traditional philologists and may therefore be perceived as an obstacle, instead of an aid, to their research activities. We illustrate how an approach based on Domain Specific Languages (DSLs) may simplify the creation of digital apparatuses, the data interchange, and the cooperation between the communities of traditional and digital philologists. Our case studies are represented by a sample digital edition of the biblical Hebrew book of Qohelet and by the digitalization of the collation of Hebrew manuscripts of the same book performed by Benjamin Kennicott at the end of the XVIII century. Both have been annotated through the web application based on DSLs named Euporia.

Italiano. La codifica manuale di apparati critici attraverso linguaggi quali TEI-XML è un compito complesso e soggetto ad errore. Essa richiede una preparazione tecnica che non rientra in genere tra le competenze dei filologi tradizionali e rischia pertanto di essere percepita come un ostacolo alle rispettive attività di ricerca, anziché come una risorsa. Nel presente articolo illustriamo in che modo un approccio basato su Domain Specific Languages (DSLs) sia in grado di semplificare la creazione di apparati digitali, lo scambio dei dati e la cooperazione tra le due comunità di filologi tradizionali e filologi digitali. I casi di studio sono rappresentati da un esempio di edizione critica digitale del libro biblico di Qohelet e dalla digitalizzazione della collazione di manoscritti ebraici medievali dello stesso libro eseguita da Benjamin Kennicott alla fine del XVIII secolo. Entrambi sono stati annotati attraverso Euporia, un'applicazione web basata su DSLs.

1 Introduction[1]

The work of the textual philologist consists of two main parts: 1. the gathering and the systematic analysis of all the available documents (the *witnesses*) of a literary work (*recensio*); 2. the removing of all the errors due to the textual transmission process (*emendatio*) (Timpanaro 2005). During the *recensio* phase, the scholar proceeds to a comparison of the witnesses with the purpose of detecting textual differences (the *variant readings* or simply *variants*). This procedure, named *collatio*, is one of the most important and delicate phase within the workflow of the text-critical praxis and represents a preliminary step to the preparation of a critical edition. The variants are presented in the critical apparatus, an instrument devised to show the reader the results of both the *recensio* and the *emendatio* by means of a conventional and formalized language, specific for the domain of textual philology (Domain Specific Language, cf. section 3). One of the tasks of the digital philologist consists in encoding variant readings and conjectural emendations. The encoding enables the creation of dynamic critical apparatuses: as with databases, the user can decide which data to extract and present, to combine the result of different queries, to transform the philological data into numerical format suitable for quantitative analysis and, finally, to prepare a digital version of the work. Unlike traditional, printed critical apparatuses, where the information is

[1]Even if both authors contributed equally to this work, L. Bambaci is responsible for sections 1-3 and 5-6 and F. Boschetti is responsible for section 4.

stored in a predefined, static way, a digital apparatus allows to retrieve, from the same encoded file, different types of information according to different research purposes and needs (Driscoll and Pierazzo 2016). The guidelines provided by the Text Encoding Initiative (TEI)[2] are among the best practices in the domain of the digital philology. TEI markup schemes pursue interoperability and reusability, making available for digital philologists a common interchange language covering a large set of text-critical phenomena.[3] TEI digital framework, moreover, is flexible enough to enable the user to add new tags and attributes, thus allowing to shape customized encoding vocabularies suitable for specific text-critical problems and for different literary traditions. The verbosity and complexity of XML language, however, combined with the necessity of being adherent to standards, is at risk of distracting the traditional philologist from his or her critical activity. Goal of this paper is to show how it is possible to encode variant readings by exploiting the nature of DSL which characterizes the language of the critical apparatus, without requiring from the philologist to deal with TEI technicalities and with problems of conformity to standards. Our case studies are represented by a sample digital collation of one book of the Hebrew Bible, the book of Qohelet also known as Ecclesiastes, conventionally dated to V-III BC, and by the digital version of the collation of Hebrew Medieval manuscripts of the same book, carried out by Benjamin Kennicott at the end of the XVIII century (Kennicott 1776). Both the collations are part of a forthcoming doctoral dissertation, which aims to prepare a digital critical edition of the literary work.

2 Background

As we have already discussed in Bambaci et al. 2018, 2019, many tools for encoding critical apparatuses are already available or currently developing.[4] The strategies adopted by these tools to avoid or facilitate the encoding process are mainly based on *ad hoc* graphical user interfaces or on annotation systems through abbreviated markers. Tools that allow to create digital TEI apparatuses directly from printed, traditional ones are lacking or not fully developed, such as in the case of the Classical Text Editor (Hagel 2007). The methodology we propose is implemented on Euporia, a web annotation system based on DSLs developed at the CoPhiLab of the CNR-ILC. Euporia has been formerly used for interpretative tasks, such as the identification of ritual frames in the ancient Greek tragedies (Mugelli et al. 2016), and also for educational purposes, namely in teaching Ancient Greek both in secondary school (Liceo Classico) and with BA students in Classics (first year students) at the University of Pisa.

3 Methodology

As stated in the introduction, the critical apparatus is the part of a critical edition or collation devoted to the collection of textual variants and conjectural emendations. There are no fixed guidelines for compiling a critical apparatus. The different methodologies are indeed the result of different traditions of study and research practices, which depend not only on the literary domain (a critical edition of a classical text will necessarily be different from a critical edition of a medieval text, Varvaro 1970), but also on internal developments and trends within each discipline: even different editions of the same text may vary in the choice of vocabulary or typographical standards, according to the different scholarly orientations or the editor's critical insights. Despite this extreme degree of variability, the critical apparatuses, however different the shapes they may assume, all strive for the same goal: to overcome the verbosity of the natural language and present the information in a way which is as concise as possible. The result of this process of departure from the natural language is an *artificial* (or *planned*) language (Blanke 2011, Libert 2018), specific to the domain of the textual philology (Domain Specific Language, DSL). Unlike a General Purpose Language (GPL), which is applicable across domains, a DSL is a language of limited expresiveness optimized for a particular domain of knowledge or domain of

[2] https://tei-c.org/

[3] Cf. in particular chapter 12 of the TEI Guidelines, devoted to the encoding of the critical apparatus: TEI Consortium, eds. "12 Critical Apparatus." TEI P5: Guidelines for Electronic Text Encoding and Interchange. [3.5.0.]. [16th July 2019].

TEI Consortium. https://tei-c.org/release/doc/tei-p5-doc/en/html/TC.html ([29/11/2019]).

[4] Cf. https://wiki.tei-c.org/index.php/Category:Editing_tools and https://wiki.tei-c.org/index.php/Editors for a list of the main editing tools.

application (Fowler 2010).[5]

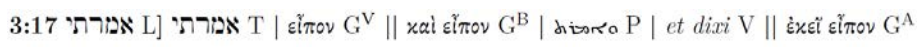


Figure 1: A sample apparatus entry from Qohelet 3:17

Let's take an example of apparatus entry from the third chapter of Qohelet, verse 17 (Fig. 1). Concepts such as "location", "lemma", "witness", "reading" and "variant reading" are encoded here by means of: 1. numbers (chapters and verses); 2. strings (the Latin *sigla* indicating the witnesses; the words of the readings); 3. characters (separators such as the square bracket "]", found after the lemma; a vertical line "|" which divides the readings and a double vertical line "||" which marks the end of a reading group). The function of the apparatus components is determined by their position within the sentence (the first reading group shows the readings supporting the lemma; the following groups contain the variants).[6] Similarly, in the apparatus entry of Qohelet 5:1 in Kennicott's collation (Fig. 2) we find integers for the

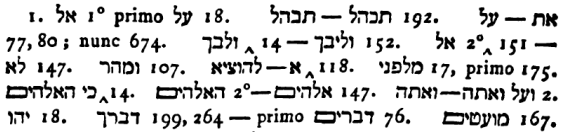


Figure 2: A sample apparatus entry of Qohelet 5:1 from Kennicott's collation

identification of the location ("verse 1"), of the manuscripts' *sigla* and of the word occurrence in the reference text (the numero sign following the roman number after the reading, e. g. "1°"); Hebrew words identifying lemmas and readings; special symbols for describing the variant typology (e. g. the symbol "^" standing for omission); annotations concerning the source description ("*primo*" for "first copyist's hand, "*nunc*" for "second copyist's hand") and finally special separators for discriminating reading groups ("—") and apparatus entries (".). A language of this sort, in which all the constituents are characterized in a concise, non-redundant and unambiguous way, is a formal language. From a computational point of view, a formal language is a language whose structure (its syntax) and meaning (its semantics) is clearly defined (Grishin 1989). A computer, therefore, is able to check that sentences are grammatically correct (well-formed) and to recognize their meaning and function (Reghizzi 2009).

4 Workflow

4.1 The Context Free Grammar

In order to allow the interpreter (or compiler) to parse the apparatus written in our DSL, we wrote the Context Free Grammar (CFG). A CFG is a formal grammar consisting of a set of rules describing a formal language (Parr 2010). An example of CFG suitable for analysing the lemma is shown in Fig. 3. Thanks to the parsing system provided by ANTLR software (Parr 2012), the CFG allows to tokenize and parse the whole apparatus entry, identifying the vocabulary symbols (token rules) and the syntactic structure (parser rules). The token rules (in the example, the rules in capital letters) allow to tokenize integers (`NUM`), alphabetic characters (`ALPHA_SEQ`), Hebrew words (`HEBW`) and separators (`R_BRACKET`). The parser rules allow to check the syntax: the lemma (`lem`), for example, is encoded as a sequence of Hebrew words (`w+`), witnesses sigla (`wit`) and separators (`LemSep`). The result of the parsing of the whole apparatus entry is an Abstract Syntax Tree (AST), as shown in Fig. 4 below. The AST attaches

[5]Examples of DSL can be considered the language of algebra for stating numerical relationships, the language of boolean logic for propositional calculus and, more in general, any kind of notation systems that allows, within a particular community of practice, the description of problems and solutions in a specific area of interest. In computer science, examples of DSLs are HTML for web pages, SQL for relational databases, LaTeX for text processing, XSLT for XML transformations and so forth. In software engineering, these languages can be called, more properly, Domain Specific Programming Language (DSPL), as opposed to General Purpose Programming Language (GPPL), such as Java, C, etc.

[6]Both vocabulary and structure of the critical apparatus of the sample edition have been shaped following the more recent critical edition of the book (Goldman 2004).

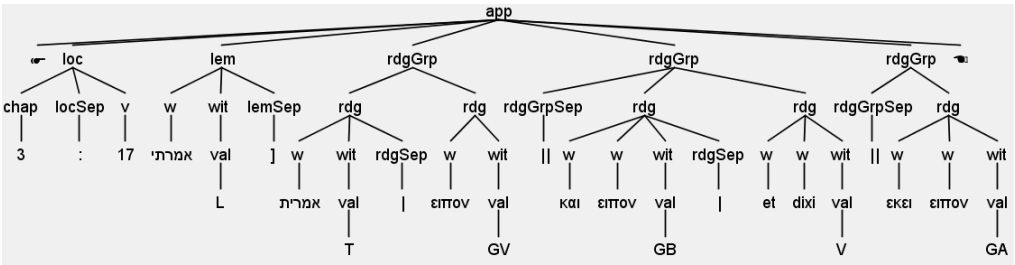
```
grammar QoheletEuporia;

app : loc lem;
lem : w+ wit lemSep;
loc : chap + locSep + v?;
chap : NUM;
v : NUM;
locSep : DOUBLE_POINT;
lemSep : R_BRACKET;
wit : ALPHA_SEQ;
w : HEBW ;

NUM : [0-9]+('.'[0-9]+)?;
ALPHA_SEQ : [a-zA-Z]+;
DOUBLE_POINT : ':';
R_BRACKET : ']';
HEBW : [\u0590-\u05ff]+;
```

(a) CFG for the sample edition of Qohelet

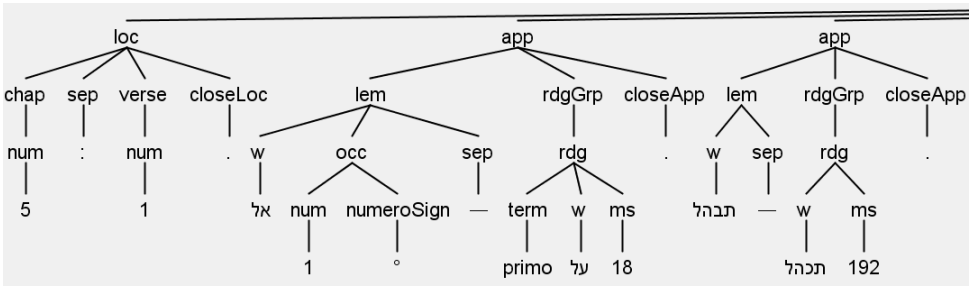
```
grammar kennicott;
all: listApp+;
listApp: loc app+ closeMainApp;
app: lem  rdgGrp+ closeApp | note? closeApp;
lem: (w+ occ? (missWord w+)? (sep)?) | (COMM NUM? (CONJ NUM)?);
rdgGrp: rdg+ (sep?);
rdg: ((w+)? (missWord w+)? (term+)? (w+)?) (w+|term+) ms+ note?;

w: HEBW rasura?;
missWord: MISSING+ (BRACKET_OP NUM BRACKET_CL MISSING+)?;
loc: chap sep verse closeLoc;
endVs: NUM END;
term: MAN_DESC;
ms: NUM ALPHA_SEQ? commaSep?| ALPHA_SEQ+ commaSep? ;
closeApp: END;
closeLoc: END;
```

(b) CFG for Kennicott's collation

Figure 3: Examples of Context Free Grammars



(a) AST from the sample edition



(b) AST from Kennicott's collation

Figure 4: Examples of syntactic trees (AST)

to the textual items (the nodes) a label which remind the function assumed in the context and shows the syntactic hierarchical relationships existing between them (the branches).

4.2 The Visitor

In order to convert any DSL in XML, we wrote a software component, named "AstToXmlVisitor", which generates an XML file structured on the AST. The Visitor passes through the tree nodes and slavishly translates the parser rules into XML markers (Fig. 5). The result is a structured, well-formed XML file, whose elements take the name from the parser rules and the hierarchical structure from the AST. The Visitor has been implemented in Java language, through the set of tools available in ANTLR4 software.

```
<app>
  <loc>
    <chap>3</chap>
    <locSep>׃</locSep>
    <v>17</v>
  </loc>
  <lem>
    <w>אמרתי</w>
    <wit>L</wit>
    <lemSep>]</lemSep>
  </lem>
  <rdgGrp>
    <rdg>
      <w>אמרתי</w>
      <wit>T</wit>
      <rdgSep>|</rdgSep>
    </rdg>
    <rdg>
      <w>εἶπον</w>
      <wit>GV</wit>
      <rdgSep>|</rdgSep>
    </rdg>
  <rdgGrp>
...
```

(a) XML output of Qoh. 3:17 from the sample edition

```
<listApp>
   <loc>
      <chap>
         <num>5</num>
      </chap>
      <sep>׃</sep>
      <verse>
         <num>1</num>
      </verse>
      <closeLoc>.</closeLoc>
   </loc>
   <app>
      <lem>
         <w>אל</w>
         <occ>
            <num>1</num>
            <numeroSign>˚</numeroSign>
            <sep>—</sep>
         </occ>
      </lem>
      <rdgGrp>
         <rdg>
            <term>primo</term>
            <w>על</w>
            <ms>18</ms>
         </rdg>
      </rdgGrp>
      <closeApp>.</closeApp>
   </app>...
</listApp>
```

(b) XML output of Qoh. 5:1 from Kennicott

Figure 5: Visitor's XML outputs

4.3 From XML to TEI-XML

A final conversion from XML code to a TEI compliant critical apparatus has been carried out through an XSLT stylesheet (Fig. 6). During the transformation phase from XML to TEI-XML, the philologist can choose which elements represent in the encoding and which to rule out (punctuation, separators and so forth). The encoding of both the apparatuses rely on the TEI model of critical apparatus: the apparatus of the digital edition of Qohelet has been encoded with the parallel segmentation method, while the encoding of Kennicott's collation follows the location-referenced method.

5 Results

So far, the first three out of twelve chapters of Qohelet have been collated. Kennicott's collation has been totally digitalized and automatically encoded. The critical apparatus of both the collations hosted in Euporia have been successfully converted into a compliant TEI file. Using XSLT stylesheets, it is possible to re-convert the TEI file back to our DSL, without loss of information. TEI encoding schemes and our DSL are therefore isomorphic. The implementation of AstToXmlVisitor in JavaScript language represents an important point of the work-flow. It allows to automatically create a well-formed XML file from any AST and is therefore applicable to different DSLs. It is written once and for all by the computer scientist and needs no further customization according to the input file. Such a division of tasks meets the needs and habits of digital philologists, who are generally more accustomed to manipulating XML code, rather than working with general-purpose programming languages.

```xml
<div type="chap" n="3">
  <ab n="17">
    <app>
      <lem wit="#L">אמרתי</lem>
      <rdgGrp>
          <rdg wit="#T">אמרית</rdg>
          <rdg wit="#GV">εἶπον</rdg>
      </rdgGrp>
      <rdgGrp>
          <rdg wit="#GB">καὶ εἶπον</rdg>
          <rdg wit="#P">ܐܡܪ̈ܬ</rdg>
          <rdg wit="#V">et dixi</rdg>
      </rdgGrp>
       <rdgGrp>
          <rdg wit="#GA">ἐκεῖ εἶπον</rdg>
      </rdgGrp>
    </app>
  </ab>
</div>
```

(a) TEI apparatus of Qoh. 3:17

```xml
<listApp>
   <app loc="5 1">
      <lem>
         <w>אל</w>
         <num>1</num>
         <pc>˚</pc>
      </lem>
      <rdgGrp>
         <rdg wit="#K18">
            <term>primo</term>
            <w>על</w>
         </rdg>
      </rdgGrp>
   </app>
   <app loc="5 1">
      <lem>
         <w>תבהל</w>
      </lem>
      <rdgGrp>
         <rdg wit="#K192">
            <w>תכהל</w>
         </rdg>
      </rdgGrp>
   </app> ...
</listApp>
```

(b) TEI apparatus of Qoh. 5:1

Figure 6: TEI compliant XML encoding

6 Discussion

There are several advantages in using a DSL for ecdotic purposes. First of all, the compactness of the DSL respect to TEI encoding. The annotation through a DSL is significantly less verbose than TEI annotation, as it can be seen by comparing the number of characters of the traditional apparatus shown in Fig. 1 and the TEI counterpart of Fig. 4 (on the right). Compactness is an important feature: the verbosity of XML language may compromise human readability and make the encoding difficult to handle, especially for traditional scholars not accustomed with long in-line encoded files. Manually encoding is a time-consuming task. Markup vocabularies require a long apprentice time to be mastered; once the encoding is complete, moreover, the encoder must check whether it is internally coherent, perfectly TEI conformant and in line with best practices. The cognitive stress derived from such a mixture of disciplinary content and cross-disciplinary formalism may stray away the world of humanistic academic research from the potentialities of computational technologies, thus contributing to increase the gap between the respective communities of scholars. A DSL-based approach, on the contrary, is entirely domain-centered: the scholar is not compelled to acquire skills which fall outside his or her cultural background, nor to make his or her research practices adhere to external, technology-conditioned standards. It is up to the digital philologist, who best knows how to organize the data according to standards, to create a perfectly conformant TEI encoding from the results of the XML general exporter. Finally, the DSL may represent a good way to exercise tighter control not only on transcriptional errors, but also on semantic errors. Thanks to the tokenization, indeed, the parser is able to assign a semantic value to each apparatus component directly from the data type to which it belongs: so, for example, tokens such as “omit” (abridgement for Latin “omittit”, non-capital character), “K1” (capital alphabetic character + integer), “McNeile(1904)” (string of alphabetic characters, integers and punctuation), will always be parsed differently and automatically assigned to different XML tags or attributes (In TEI encoding, respectively, `@ana`, `@wit`, `@resp`). In a manual encoding, on the other hand, the encoder must decide, each time, which tags or attributes are more suitable for expressing his or her interpretations of textual phenomena: this may often lead to an incoherent or erroneous choice of markers and increase the possibility of semantic errors, which are very difficult to be detected. In a DSL-based approach, on the contrary, the choice of markers is not entrusted to human decision, but it is determined by the form of the apparatus components and automatically performed by the Visitor.

References

Luigi Bambaci, Federico Boschetti, and Riccardo Del Gratta. 2018. Qohelet Euporia — A Domain Specific Language to Annotate Multilingual Variant Readings. In *Proceedings of the 5th International Congress on Information Science and Technology, Marrakech, Morocco, October 21-27, 2018*. Piscataway, NJ, pages 266–69.

Luigi Bambaci, Federico Boschetti, and Riccardo Del Gratta. 2019. Qohelet Euporia — A Domain Specific Language for the Encoding of the Critical Apparatus. *International Journal of Information Science and Technology* 3(5):26–37.

Detlev Blanke. 2011. Planned Languages — a Survey of some of the main Problems. In *Interlinguistics: Aspects of the Science of Planned Languages*, Walter de Gruyter, pages 63–87.

Matthew James Driscoll and Elena Pierazzo. 2016. *Digital Scholarly Editing: Theories and Practices*. Open Book Publishers, Cambridge.

Martin Fowler. 2010. *Domain-Specific Languages*. Pearson Education.

Yohanan A. P. Goldman. 2004. Qohelet. In *Biblia Hebraica Quinta: Megilloth: Ruth, Canticles, Qoheleth, Lamentations, Esther*, Deutsche Bibelgesellschaft, Stuttgart.

V. N. Grishin. 1989. Formalized language. In *Encyclopaedia of Mathematics*, Kluwer Academic Publishers, volume 4, pages 61–62.

Stefan Hagel. 2007. The Classical Text Editor. An attempt to provide for both printed and digital editions. In A. Ciula and F. Stella, editors, *Digital philology and medieval texts*, Pacini, University of Michigan, pages 77–84.

B. Kennicott. 1776. *Vetus Testamentum Hebraicum cum variis lectionibus*, volume 2. Clarendon, Oxford.

Alan Reed Libert. 2018. *Artificial Languages*. Oxford University Press.

Gloria Mugelli, Federico Boschetti, Riccardo Del Gratta, Angelo Mario Del Grosso, and Fahad Khan. 2016. A user-centred design to annotate ritual facts in ancient Greek tragedies. *BICS* 59(2):103–120.

Terence Parr. 2010. *Language Implementation Patterns: Create Your Own Domain-specific and General Programming Languages*. Pragmatic Bookshelf.

Terence Parr. 2012. *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf.

Stefano Crespi Reghizzi. 2009. *Formal Languages and Compilation*. Springer.

Sebastiano Timpanaro. 2005. *The Genesis of Lachmann's Method*. University of Chicago Press, Chicago / London.

Alberto Varvaro. 1970. *Critica dei Testi Classica e Romanza — Problemi Comuni ed Esperienze Diverse*. L'Arte Tipografica, Napoli.

600 maestri raccontano la loro vita professionale in video: un progetto di (fully searchable) open data

Gianfranco Bandini
Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia
Università di Firenze
gianfranco.bandini@unifi.it

Andrea Mangiatordi
Dipartimento di Scienze Umane per la Formazione "Riccardo Massa"
Università di Milano Bicocca
andrea.mangiatordi@unimib.it

Abstract[1]

English. Oral History – as in the collection, storage and interpretation of personal accounts – already provided interesting insights to the renovation of school practices, even though this effect is probably still limited. The project presented here is based on the experience of the "Memorie di Scuola" website (https://memoriediscuola.it), a collection of more than 600 interviews to teachers about their professional history. Using Open Source software and YouTube APIs, the authors were able to create a video repository where: 1) hundreds of hours of video interviews to teachers are organized and made accessible online for free, following a "public history" approach; 2) video content is fully searchable thanks to the use of automatic transcription and of a synchronization script.

Italiano. La storia orale – intesa come raccolta, archiviazione e interpretazione di testimonianze – ha dato un contributo interessante, anche se ancora limitato, al rinnovamento della storia della scuola. Il progetto qui presentato si basa su una piattaforma web (https://memoriediscuola.it) che raccoglie ad oggi oltre 600 interviste a maestri che raccontano in video la loro storia professionale. Attraverso l'uso strumenti software Open Source e delle API di YouTube, gli autori descrivono un modello di raccolta, archiviazione e codificazione del parlato che consente di raggiungere due importanti obiettivi: 1) rendere le centinaia di ore di filmato completamente disponibili on line (attraverso un approccio di public history); 2) consentire la libera esplorazione dei video attraverso l'indicizzazione di tutte le parole che sono state pronunciate durante le interviste.

1 Tematiche e obiettivi della ricerca

Il progetto "memorie di scuola" è basato sul Content Management System "WordPress" (https://wordpress.org), una piattaforma sulla quale si basa – secondo le stime dei suoi autori e sviluppatori – circa un terzo dei siti web mondiali. La nostra implementazione, utilizzabile attraverso qualsiasi web browser, è stata adattata a esigenze molto specifiche, in modo da consentire la costruzione di una memoria collettiva della vita scolastica nella scuola primaria in Italia. Al momento raccoglie oltre 600 interviste a maestre e maestri in pensione (o prossimi ad essa) che raccontano con molti dettagli e grande passione la loro vita professionale, a partire dagli anni '40.
Il nostro intento è quello di migliorare la piena disponibilità dei video (che costituiscono un ampio insieme di *open data*) e la loro completa fruibilità attraverso un sistema di indicizzazione analitico. I video, così come le attività formative e didattiche connesse, sono indirizzati in primo luogo agli insegnanti in servizio e in formazione, ma anche a un più ampio pubblico interessato agli aspetti educativi della nostra storia sociale. In questa sede presentiamo quindi l'implementazione di una particolare *feature* del sito web che consente di effettuare una ricerca full-text all'interno di tutte le parole pronunciate in tutti i video che compongono il progetto (e di collegarsi ad essi nel preciso istante durante il quale il soggetto pronuncia la parola cercata).

[1] Gli autori hanno lavorato alla stesura del testo confrontandosi e concordando ogni sua parte. Gianfranco Bandini si è dedicato in particolare alle sezioni 1, 2 e 4; Andrea Mangiatordi ha curato in particolare le sezioni 3 e 3.1.

2 Quadro teorico di riferimento

Nel corso del Novecento la storia orale ha affermato, non senza contrasti e opposizioni, la sua legittimità e utilità, soffermandosi soprattutto sulla storia dal basso, degli esclusi dalla storia tradizionale (Gardner, & LaPaglia, 2006). Il settore di studi che si occupa della storia della scuola, all'interno della storia dell'educazione (McCulloch, 2011), ha utilizzato con sempre maggiore convinzione fonti non testuali, come le fotografie o i dipinti. Una piccola parte di queste ricerche ha inoltre scoperto, anche se con un certo ritardo, l'utilizzo delle fonti orali, cioè la raccolta delle testimonianze (Gardner, 2003). La memoria personale ha consentito di spostare l'accento degli studi sulle percezioni e sui sentimenti delle persone, sugli aspetti interiori e comunitari della vita sociale.
La storia dell'educazione e la storia orale, in questa forma congiunta, hanno trovato un campo di nuova e eccezionale sperimentazione nell'ambito della *digital public history,* nel quale il presente progetto si colloca (Bandini, 2017). Nel contesto digitale, l'incrocio di queste diverse tradizioni di ricerca dà la possibilità di potenziare la comune aspirazione a un maggiore contatto tra il mondo accademico e la società, soprattutto per quanto riguarda il rapporto con le professioni educative e di cura (cfr. Depaepe, 2001; Linné, 2001; Vinovskis, 2015). La trasformazione delle classiche fonti storiche in *open data* risponde proprio a questo ambizioso obiettivo.
Il progetto consente, oltre alla piena e completa disponibilità *online* delle testimonianze, di superare una delle principali limitazioni che qualsiasi tradizionale ricerca che fa uso di storia orale si trova a affrontare (cfr. Ritchie, 2011): la numerosità delle testimonianze e la gestione della massa dei dati che possono esserne ricavati (intesi anche come momenti di titubanza del testimone, silenzi, salti temporali, ecc.).
A questo proposito, uno dei punti di discussione metodologica consiste, ad esempio, nel riflettere sulla questione del numero ottimale di testimonianze. In alcuni casi (cfr. l'ampio studio di Johnson & Reuband, 2008), in abbinamento a ricerche di tipo quantitativo, si è stabilito un vero e proprio piano di campionamento, come di consueto nelle ricerche statistiche. In generale, tuttavia, nelle ricerche storiche si sostiene che la ricerca di nuove testimonianze si può arrestare nel momento in cui si satura il campo concettuale che stiamo indagando, cioè quando nuove testimonianze non porterebbero più nulla di nuovo, o insignificanti dettagli, rispetto a quanto già detto.
Nel contesto della storia digitale il panorama risulta ampiamente modificato perché la raccolta delle testimonianze può essere espansa senza eccessiva fatica e scarso dispendio di risorse (cfr. Thomson, 2007). Questo aspetto ci consente di progettare una raccolta di testimonianze aperta, nel senso che può essere incrementata nel tempo e offrire sempre nuove opportunità di conoscenza e formazione. La cassetta degli attrezzi dello storico e l'insieme delle sue fonti vengono così trasportate da uno spazio privato a uno spazio pubblico.
Nell'ottica della *public history*, il contesto digitale può quindi essere progettato non soltanto per la collocazione online di fonti primarie (come sono le interviste), ma per consentire lo sviluppo delle interazioni tra gli utenti. Per superare l'impostazione tradizionale, che trasforma le fonti online in musei statici (per quanto digitali e più facilmente accessibili) c'è bisogno di alcuni strumenti di base, che sono volti a aumentare la significatività delle fonti. La trascrizione automatica dei video, in questo senso, rappresenta un tassello fondamentale della strategia comunicativa, a vantaggio della piena fruibilità da parte degli utenti.

3 Metodologia

La raccolta delle testimonianze video presenta delle particolari caratteristiche rispetto alla raccolta di documenti testuali (per esempio diari o autobiografie). Il *repository* "memorie di scuola" attualmente supera le 5.000 ore di filmato e rende oggettivamente molto difficile sia l'ascolto integrale, sia una completa attività di indicizzazione manuale basata sull'inserimento di *tags* (per esempio quelli contenuti in TESE, Thesaurus Europeo dei Sistemi Educativi). Nella prospettiva della *public history* il numero delle ore è inoltre destinato a crescere ancora: in questa situazione la messa a disposizione del pubblico, con una piccola serie di *tags* di indicizzazione, di fatto non consente l'accesso alla ricchezza documentaria contenuta nei video che diventa quasi casuale.
Mantenendo la prospettiva fin qui esposta, il progetto si è indirizzato a cercare di produrre un archivio video liberamente accessibile attraverso l'uso di risorse sostenibili – per lo più software Open Source – e l'automazione di una serie di operazioni che permettono di facilitare la ricerca all'interno della grande mole di contenuti raccolti. Il servizio online YouTube (https://www.youtube.com), famoso per essere tra i principali

repository gratuiti di contenuti video online, consente la trascrizione automatica del parlato dalla lingua italiana al fine di produrre sottotitoli. La quantità di errori di trascrizione è altamente variabile e dipendente da molteplici fattori, primo tra tutti la qualità della traccia audio (cfr. Alberti e altri, 2009). Tuttavia, un sistema in grado di estrarre i sottotitoli creati automaticamente dal servizio e di inserirli in un database permette di rendere le interviste esplorabili e ricercabili in modo simile a un corpus testuale. Questo sistema, pur non presentando un'innovazione dal punto di vista degli algoritmi e del software, consente di ottenere un significativo vantaggio rispetto ai sistemi attuali e offre un modello di funzionamento facilmente replicabile e applicabile a grandi masse di dati. Bisogna ricordare che quando sono in gioco grandi quantità di interviste, la trascrizione umana (ovviamente più accurata dei sistemi automatici) solo in rarissimi casi riesce a raggiungere il risultato: è il caso, quasi unico nel suo genere, della grande raccolta di testimonianze condotta dalla Shoah Foundation.[2]

3.1 Struttura del Software e flussi di lavoro

L'architettura della soluzione software messa in atto per sostenere il progetto Memorie di Scuola consta di servizi e applicazioni Open Source che si interfacciano con il servizio proprietario YouTube per l'hosting dei materiali video e per la produzione di trascrizioni rese disponibili nella forma di sottotitoli e in formati diversi. In particolare, lo *stack software* utilizzato è basato su:

- Un web server dotato in grado di eseguire codice scritto in linguaggio PHP – nel caso specifico del progetto è stato utilizzato il software NginX (https://www.nginx.com/), ma a questo livello non ci sono requisiti particolarmente stringenti;
- Un database MySQL (https://www.mysql.com/);
- Il CMS WordPress (https://wordpress.org);
- Il plugin Meks Video Importer per WordPress (https://wordpress.org/plugins/meks-video-importer/), che permette la ricerca di video disponibili pubblicamente su YouTube e l'importazione automatica del contenuto testuale della loro descrizione;
- Il plugin Advanced Custom Fields per WordPress (https://wordpress.org/plugins/advanced-custom-fields/), che facilita l'inserimento di metadati ai contenuti dei post;
- Il software YouTube Transcript/Subtitle API (https://github.com/jdepoix/youtube-transcript-api), uno script in grado di estrarre i sottotitoli da video YouTube pubblicamente disponibili;
- Lo script "YouTube transcript to WordPress" (https://github.com/andreamangia/youtube-transcript-to-wp), progettato specificamente per automatizzare le operazioni.

Il software si inserisce nel workflow descritto di seguito e lo sostiene, minimizzando la necessità di intervento umano ma rendendola comunque possibile in una fase successiva di revisione dei contenuti:

- L'intervistatore effettua l'intervista e esegue l'upload su YouTube, aggiungendo al video una semplice descrizione testuale;
- Un operatore del sito web www.memoriediscuola.it importa il video, nella forma di un nuovo post WordPress, con l'aggiunta di tag ed eventuali altri termini tassonomici, la verifica della congruenza tematica e l'indicazione di dati quali il nome dell'intervistatore, il luogo;
- L'operatore del sito web esegue lo script "YouTube transcript to WordPress" indicando il numero identificativo del post WordPress generato al punto precedente. Lo script si occupa di:
  - Estrarre in formato JSON la trascrizione dell'audio;

[2]La Shoah Foundation gestisce il Visual History Archive, disponibile all'indirizzo https://sfi.usc.edu/vha, con lo scopo di documentare e rinforzare l'empatia verso le memorie di persone che hanno vissuto genocidi e altri drammi.

    - Trasformare ciascun frammento della trascrizione (in genere, ma non sistematicamente, corrispondente a una breve frase) in un elemento HTML contenente l'indicazione del momento temporale in cui il testo viene pronunciato nel video;
    - Salvare l'intera trascrizione come valore per un campo personalizzato di WordPress associato al post contenente il video.
- L'utente effettua una ricerca libera all'interno del repository, che è stato opportunamente configurato per consentire la ricerca anche all'interno dei campi personalizzati, normalmente ignorati dal motore di ricerca interno di WordPress;
- Il sito web elenca tutti i video che contengono la parola ricercata. Accedendo alla pagina di ciascun video è possibile raggiungere il momento in cui una frase è pronunciata attraverso un click sulla porzione di trascrizione corrispondente.

La selezione degli strumenti operativi che rendono possibile il procedimento è dunque stata pensata per favorire la replicabilità totale dell'esperienza in qualunque sistema basato su WordPress, indipendentemente da elementi quali il tema grafico utilizzato o dalla necessità di modificare l'architettura del database sul quale poggia il sistema. Non è previsto supporto in questa fase per altri CMS, ma non è da escludere che le stesse funzionalità e la stessa logica di lavoro possano essere applicate anche altrove, data la presenza di diversi layer basati su tecnologie standard e interoperabili.

4 Risultati attesi e interventi futuri

Il progetto ha come obiettivo principale la costruzione di un set di *open data* costituiti da testimonianze video, liberamente accessibili sul web e esplorabili in profondità attraverso gli strumenti di ricerca del parlato sopra descritti.
Questo obiettivo, del resto, è propedeutico a molte azioni formative che possono essere svolte proprio grazie alla possibilità di trovare all'interno dei video esattamente ciò che stiamo cercando, siano esse parole generiche o identificatori di luogo o di persona.
L'insieme delle trascrizioni si presta inoltre a successive analisi e esplorazioni con software di data mining (per esempio T-Lab) o di categorizzazione dei testi (per esempio Nvivo), all'interno del paradigma di ricerca della Grounded Theory. Per quanto l'accuratezza delle trascrizioni possa essere migliorata, già allo stato attuale i testi dei video appaiono ben comprensibili e di grande aiuto per effettuare delle analisi approfondite. Nel caso di uno studio volto alla categorizzazione dei testi, un ulteriore passaggio manuale di correzione potrebbe portare a un corpus assestato e corretto in tempi ragionevolmente brevi. Tenendo conto che i software di analisi dei testi sono di fatto degli strumenti semi-automatici, questo tipo di operazione risulta essere di carattere ordinario. Bisogna inoltre considerare che il presente progetto, in modo del tutto automatico (attraverso il citato plugin Meks Video Importer per WordPress), si gioverà dei miglioramenti nel riconoscimento del parlato che verranno implementati nella piattaforma YouTube. Lo sviluppo dei sistemi di Intelligenza Artificiale ha dato prova, in questi ultimi anni, di consentire dei progressivi e tangibili miglioramenti nella comprensione del parlato, a partire dalla lingua inglese (riconosciuta nelle sue molte varianti di pronuncia).
L'insieme delle trascrizioni, infine, può costituire la base per una piattaforma partecipativa che permetta di dare avvio a progetti di crowdsourcing di correzioni e integrazioni alle trascrizioni automatiche, aumentando ulteriormente la sostenibilità del progetto; inoltre, in linea con l'approccio di public history fin qui adottato, sarà possibile dare la possibilità agli utenti di apporre commenti ai video, commenti che a loro volta verranno trascritti e inseriti in una mappa esplorabile per concetti chiave.
Dal punto di vista tecnico è possibile pensare in tempi brevi all'integrazione di altre due tecnologie Open Source. La prima di queste è la piattaforma di ricerca Apache Solr (https://lucene.apache.org/solr/): questa mette a disposizione un motore di indicizzazione più raffinata rispetto alla semplice ricerca testuale disponibile nel CMS WordPress e permetterebbe di avere risultati di ricerca rispondenti a query di tipo diverso (con operatori booleani, stemming, proximity search). La seconda tecnologia potenzialmente utilizzabile è Hypothes.is (https://web.hypothes.is/), software Open Source di annotazione di pagine web, che renderebbe possibile appunto l'annotazione dei contenuti, in modalità aperta e collaborativa da parte di ricercatori diversi.

Ringraziamenti

Si ringrazia sentitamente il prof. Gianfranco Crupi per gli utili consigli e suggerimenti; Inclusive Cloud S.r.l.s. per aver messo a disposizione competenze e infrastrutture informatiche; si ringraziano vivamente tutti i maestri che hanno cortesemente messo a disposizione il racconto della loro vita professionale e tutti gli studenti del corso di laurea in scienze della formazione primaria (università di Firenze) che hanno raccolto, con pazienza e serietà, le testimonianze in video.

Bibliografia

Agneta Linné. 2001. Myths in Teacher Education and the Use of History in Teacher Education Research. *European Journal of Teacher Education* 24(1): 35-45, DOI: 10.1080/02619760120055871

Christopher Alberti, Michiel Bacchiani, Ari Bezman, Ciprian Chelba, Anastassia Drofa, Hank Liao, Pedro Moreno, Ted Power, Arnaud Sahuguet, Maria Shugrina, and Olivier Siohan. 2009, April. An audio indexing system for election video material. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4873-4876. IEEE.

Gianfranco Bandini. 2017. *Educational Memories and Public History: A Necessary Meeting*. In: Cristina Yanes-Cabrera, Juri Meda, Antonio Viñao (eds.). *School Memories. New Trends in the History of Education*. 143-156. Springer International Publishing, Cham, Switzerland. ISBN:978-3-319-44062-0

Marc Depaepe. 2001. A Professionally Relevant History of Education for Teachers: Does it Exist? Reply to Jurgen Herbst's state of the art article. *Paedagogica Historica*. 37(3): 629-640, DOI: 10.1080/0030923010370305

Philip Gardner. 2003. Oral history in education: teacher's memory and teachers' history. *History of Education.* 32(2): 175-188.

James B. Gardner and Peter S. LaPaglia. 2006. *Public History: Essays from the Field*, 2nd edition. Krieger , Malabar, Florida. ISBN:978-157-524244-6.

Eric A. Johnson and Karl-Heinz Reuband. 2008. *La Germania sapeva. Terrore, genocidio, vita quotidiana. Una storia orale*. Mondadori, Milano. (orig. ed. 2005)

Gary McCulloch. 2011. *The Struggle for the History of Education*. Routledge, Abingdon, UK. ISBN:978-0-415-56535-6.

Simonetta Polenghi, Gianfranco Bandini. 2016. The history of education in its own light: signs of crisis. Potential for growth. *Espacio Tiempo y Educacion*. 3(1, January-July 2016): 3-20.

Donald A. Ritchie. 2011. *The Oxford handbook of oral history*. Oxford University Press.

Alistair Thomson. 2007. Four Paradigm Transformations in Oral History. *The Oral History Review*. 34(1): 49–70.

Vinovskis, Maris A. (2015). *Using Knowledge of the Past to Improve Education Today: US Education History and Policy-Making*. Paedagogica Historica, 51(1-2), 30-44, DOI: 10.1080/00309230.2014.9977

Ripensare i dati come risorse digitali: un processo difficile?

Nicola Barbuti
Università degli Studi di Bari Aldo Moro
nicola.barbuti@uniba.it

Abstract

English. The Art. 2 of the *EU Council Conclusions of 21 May 2014 on cultural heritage as a strategic resource for a sustainable Europe (2014 / C 183/08)* recognizes the existence of the new digital cultural heritage (*born digital* and *digitized*). Starting from this assumption, we need to rethink digital and digitization as social and cultural expressions of the contemporary age. Digital resources shall record and represent both digitization processes and themselves in their life cycle, they are no longer mere gateway to improve the access to reality. So, we have to define clear and homogeneous criteria to validate and certify them as a memory and sources of knowledge for future generations. In this regard, the present paper outlines a first proposal for identification of digital cultural resources, based on the expansion of the *R*: *Reusable* of the *FAIR Principles for the management of scientific metadata* in R^5: *Reusable, Readable, Relevant, Reliable and Resilient*.

Italiano. L'art. 2 delle *Conclusioni del Consiglio UE del 21 maggio 2014 sul patrimonio culturale come risorsa strategica per un'Europa sostenibile (2014 / C 183/08)* riconosce l'esistenza del nuovo patrimonio culturale digitale (*born digital* e *digitalizzato*). Partendo da questo assunto, si rende indispensabile ripensare il digitale e la digitalizzazione come espressioni sociali e culturali dell'età contemporanea. È necessario, perciò, riconsiderare le risorse digitali come registrazioni e rappresentazioni di processi, e non più come semplici mediatori atti a migliorare l'accesso alla realtà, definendo criteri chiari e omogenei per convalidarli e certificarli come memoria e fonti di conoscenza per le generazioni future. A riguardo, nel presente lavoro si delinea una prima proposta, basata sull'ampliamento della *R*: *Re-usable* dei *FAIR Principles* per la gestione dei metadati scientifici in R^5: *Reusable, Readable, Relevant, Reliable and Resilient*.

1 Introduzione

L'art. 2 delle *Conclusioni del Consiglio dell'UE del 21 maggio 2014 sul patrimonio culturale come risorsa strategica per un'Europa sostenibile (2014 / C 183/08)* recita[1]:
Il patrimonio culturale è costituito dalle risorse ereditate dal passato in tutte le sue forme e aspetti – tangibile, intangibile e digitale (nato digitale e digitalizzato), inclusi monumenti, siti, paesaggi, abilità, pratiche, conoscenze ed espressioni della creatività umana, nonché raccolte conservate e gestite da enti pubblici e privati come musei, biblioteche e archivi. Ha origine dall'interazione tra persone e luoghi nel tempo ed è in continua evoluzione. Queste risorse sono di grande valore per la società dal punto di vista culturale, ambientale, sociale ed economico e quindi la loro gestione sostenibile costituisce una scelta strategica per il 21° secolo.

Partendo da questo presupposto, dobbiamo necessariamente cambiare il nostro approccio al digitale e alla digitalizzazione iniziando a considerarli rappresentazioni qualificanti l'età contemporanea e la *digital transformation* che la connota. Ciò implica l'urgenza di individuare e classificare tra le risorse digitali prodotte fino a oggi e in produzione, siano esse singoli oggetti, o complesse *digital libraries*, o sistemi 3D, quelle che possono essere identificate come il nuovo *Digital Cultural Heritage* (DCH), distinguendole da quelle prodotte per mera semplificazione di processi gestionali o per la fruizione estemporanea e immediata di scadenti rappresentazioni relative a entità tangibili e intangibili.

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XG0614%2808%29

Da diversi anni, la digitalizzazione e la qualità e preservazione delle risorse digitali sono riconosciute tra le principali emergenze da affrontare in tutto il mondo. Nel 2012 l'UNESCO ha tenuto la sua conferenza a Vancouver con il significativo titolo *The Memory of the World in the Digital Age: Digitization and Preservation* (Duranti and Shaffer [ed. by], 2012), nel cui ambito è stata redatta ed emanata la *Vancouver Declaration on Digitisation and Preservation*[2], con l'IFLA e l'International Council of Archives (ICA) tra i principali responsabili.

Da allora, la situazione non sembra essere molto cambiata, nonostante gli sforzi intrapresi per accelerare l'elaborazione di soluzioni a criticità di una complessità che, forse, non hanno precedenti storici.

Gli attuali approcci ai processi di creazione delle risorse digitali, infatti, sembrano non recepire l'evoluzione che negli ultimi anni ha riguardato la digitalizzazione, ancora oggi associata semplicisticamente alla riproduzione fotografica, mentre, invece, è diventata un processo complesso guidato da regole definite e condivise. Anche l'importanza della qualità dei dati digitali è del tutto sottovalutata nel relegarne la funzione a meri strumenti di mediazione per la fruizione del il reale in forma virtuale, sebbene da più parti si riconosca che questi dovrebbero rispondere a requisiti di intellegibilità, affidabilità, pertinenza, persistenza, e registrare le trasformazioni delle funzioni legate al loro riutilizzo nel tempo. L'interpretazione strumentale, infatti, ancora orienta e condiziona negativamente soprattutto la strutturazione degli schemi di metadati con cui indicizzare gli oggetti digitali prodotti e la composizione delle descrizioni loro associate, formulate per essere meri codici funzionali esclusi-vamente alla ricerca e al recupero dei dati in rete.

Proprio i metadati e i contenuti descrittivi, invece, dovrebbero essere oggetto di particolare attenzione, in quanto sono la sola possibilità di registrare e rappresentare in modo intellegibile i processi di digitalizzazione, creazione e trasformazione che caratterizzano il ciclo di vita dei dati, e di conservare così le informazioni necessarie a conoscerli e a qualificarli come risorse digitali con funzioni anche culturali.

Il tema della funzione essenziale dei metadati nel management e nella fruizione dei dati digitali è il focus dei *FAIR Guiding Principles for Scientific Data Management and Stewardship*[3], le linee guida per il management dei dati scientifici pubblicate nel 2016, da qualche tempo uno dei temi di maggior interesse nell'ambito del più ampio dibattito sulle possibilità di applicare le metodologie del *data science* alla creazione e gestione dei *data humanities*[4].

A riguardo, nella CIDOC Conference 2018 si è tenuto un workshop sull'effettiva efficacia dei *FAIR Principles* rispetto agli scenari che oggi la digitalizzazione propone, e soprattutto rispetto a quelli che già si preannunciano imminenti[5].

Il presente lavoro sintetizza alcune riflessioni maturate da quel proficuo confronto, relative alla necessità di provvedere a un ampliamento del requisito *R*: *Reusable* nei requisiti *R*[5]: *Reusable, Readable, Relevant, Reliable* and *Resilient*, finalizzato a facilitare l'applicabilità dei *FAIR Principles* ai data humanities e, conseguentemente, l'identificazione e certificazione come DCH dei dati rispondenti a tali requisiti nell'informe magma digitale in cui oggi fluttuiamo.

2 Verso un ampliamento dei *FAIR* da *R* a *R*[5]

L'assunto di partenza per avviare la riflessione è che i dati digitali non possono più essere creati finalizzandoli alla mera funzione di strumenti di mediazione per una fruizione della realtà alternativa a quella fisica: è necessario ripensarli quali risorse digitali che si qualificano come *record*, entità dinamiche e diacroniche che registrano e conservano nelle descrizioni i processi di digitalizzazione che li hanno creati e quelli che hanno caratterizzato il loro successivo ciclo di vita.

[2] http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_ubc_vancouver_declaration_en.pdf

[3] https://www.go-fair.org/fair-principles/

[4] https://www.rd-alliance.org/open-consultation-fair-data-humanities-until-15th-july-2019; https://www.gofair.org/implementation-networks/overview/co-operas/; https://operas.hypotheses.org/

[5] http://www.cidoc2018.com/sites/default/files/CIDOC2018-BookOfAbstracts-Final-v-1-2.pdf

I metadati descrittivi diventano perciò fondamentali e inscindibili dagli oggetti digitali, in quanto sono proprio l'accuratezza e la qualità delle descrizioni a qualificarli come *record* e, quindi, a renderli risorse digitali pensate e strutturate per essere fruite diacronicamente dagli utenti del futuro, che devono comprendere cosa il dato rappresenta alla pari degli utenti contemporanei.

L'adozione dei *FAIR Principles* lascia aperte alcune questioni. Innanzitutto, non siamo del tutto persuasi che Ricercabilità (*Findable)*, Accessibilità (*Accessible*, che assolutamente non è identificabile con *Open*) e Interoperabilità (*Interoperable*) siano requisiti idonei a qualificare i dati come *record* e risorse digitali, conferendogli funzioni nuove e più evolute da quelle strumentali attualmente riconosciute.

Un dato che sia ricercabile, accessibile e interoperabile con altri non fornisce alcuna garanzia di qualità, sufficienza e affidabilità dei contenuti informativi che contiene. Inoltre, la ricercabilità e, di conseguenza, l'accessibilità e interoperabilità che sono a essa vincolate hanno senso nella misura in cui un dato sia oggetto di interesse da parte dei fruitori. E l'interesse per un dato è legato strettamente non alla sua mera funzione di chiave d'accesso a un oggetto digitale semplice o complesso, ma all'essere risorsa informativa e cognitiva per la quantità e qualità dei contenuti descrittivi che mette a disposizione dell'utente già in fase di lettura dei suoi metadati.

Siamo perciò del parere che il requisito che conferisce significato e senso ai primi tre, e dal quale questi dipendono indissolubilmente, sia la Riutilizzabilità (*Reusable*). L'utilizzo e, soprattutto il riutilizzo dei dati, infatti, sono secondo noi i fattori che ne garantiscono la sostenibilità nel tempo e, quindi, la sopravvivenza, in quanto requisiti caratterizzati da dinamismo e diacronia che, quasi sem-pre, implicano trasformazioni nelle funzioni delle entità che ne sono oggetto: per avere un'idea uti-lizzando un paradigma analogico, si pensi al Colosseo e al suo ciclo di vita.

Le registrazioni descrittive dei metadati sono perciò fondamentali per garantire qualità e persistenza delle risorse digitali, qualora siano improntate a equilibrate soluzioni quantitative/qualitative e rispondano a ulteriori requisiti che, secondo noi, sono altrettanto essenziali quanto la riutilizzabilità.

Anche la *Reusability*, infatti, di per sé non costituisce una garanzia di qualità del dato e del suo valore quale risorsa informativa e cognitiva. Anzi: proprio le variabili cui una risorsa è soggetta perché riutilizzabile possono essere fonte di distorsione e difformità dei contenuti, il cui valore informativo e cognitivo può perciò non essere più certificabile come affidabile.

La R di Reusable andrebbe perciò, secondo noi, ampliata in R5 con i seguenti requisiti:

- *Readability*: da intendersi non nell'accezione semantica di leggibilità, ma in quella concettuale di *intellegibilità* della risorsa digitale per tutte i possibili target di utenti interessati a fruirne; è requisito fondamentale per conferire ai metadati la funzione informativa e cognitiva necessaria a qualificarli come risorsa culturale, e si basa sull'equilibrato rapporto quantitativo/ qualitativo dei contenuti descrittivi e sull'accuratezza formale, stilistica e linguistica dei contenuti;
- *Relevance*: la *persistenza* nel tempo è legata all'interesse degli utenti per i contenuti informativi e cognitivi registrati nella risorsa; essa è strettamente legata al suo riutilizzo e alle possibili trasformazioni di funzione registrate nelle descrizioni; è, quindi, requisito indispensabile affinché la risorsa, di solito creata con funzioni e scopi non necessariamente culturali, possa essere identificabile e riconoscibile nella sua struttura formale e descrittiva anche se varia nel tempo le proprie funzioni evolvendosi in fonte di conoscenza sui processi che registra e, quindi, in risorsa culturale digitale;
- *Reliablility*: l'*affidabilità* è la certificazione e validazione della qualità della risorsa digitale rilevabili dalle registrazioni delle sue descrizioni durante tutto il suo ciclo di vita, in relazione a tutte le possibili trasformazioni ed evoluzioni funzionali cui può essere stata soggetta; è, dunque, strettamente connessa alla capacità dell'entità digitale di registrare e preservare gli elementi qualificanti la qualità informativa e cognitiva dei suoi contenuti descrittivi, anche nell'evoluzione delle funzioni e nelle variazioni di forme e funzioni nel tempo;
- *Resilience*: come l'intellegibilità, anche la *resilienza* applicata ai dati e, soprattutto, ai metadati è requisito fondamentale per conferire alle risorse digitali la nuova dimensione culturale;

chiosando la definizione comunemente in uso in ambito informatico[6], essa va intesa come *la capacità di una risorsa digitale di adattarsi alle condizioni di utilizzo e riutilizzo, di resistere all'usura, di essere duttile nelle trasformazioni e nell'evoluzione delle sue funzioni, al fine di garantire la disponibilità del proprio potenziale cognitivo e informativo nello spazio e nel tempo*; è, quindi, indispensabile per garantire la sostenibilità e il riutilizzo delle risorse digitali nel medio-lungo termine, provvedendo a preservare sia le informazioni utili a conoscere i processi della loro creazione, sia quelle sulla loro funzione originale, sia, infine, le registrazioni delle trasformazioni ed evoluzioni funzionali che ne hanno caratterizzato il ciclo di vita.

3 Conclusioni

Tirando le conclusioni su quanto sopra sinteticamente delineato, è nostra opinione che l'adozione dei requisiti *FAIR* con la *R* ampliata in R^5 sia prerequisito indispensabile nel processo di creazione dei dati digitali e, soprattutto, dei metadati che li descrivono, in quanto gli conferirebbero le funzioni di potenziale DCH, rendendoli sostenibili, permanenti, affidabili e, nel contempo, storicizzandoli come fonti di conoscenza dei processi e delle complessità che caratterizzano la rapidissima evoluzione della *digital transformation*.

Non il dato in sé, infatti, ma l'interesse degli utenti presenti e futuri per la fruizione del dato in quanto risorsa informativa e cognitiva deve diventare il prerequisito su cui fondare l'intero processo di creazione, pubblicazione e preservazione di risorse digitali. L'applicazione dei requisiti R^5, dun-que, deve diventare oggetto di attenzione fin dalla fase di analisi e progettazione dei processi sia di digitalizzazione che di creazione di qualsiasi schema di metadati con cui descrivere e gestire gli oggetti digitali in produzione.

Solo così si potrà dare un serio inizio, nel medio termine, all'individuazione di quanto possa essere identificato come DCH nella massa di dati che oggi sovrabbonda nel web e, nel contempo, si potranno definire linee guida omogenee e condivise che presiedano alla creazione di nuove risorse avendo chiaro fin dal principio se gli si voglia conferire il potenziale valore di entità culturali.

In questo modo, nel giro di pochi anni le *Conclusioni EU* del 2014 potranno finalmente essere sostanziate con un nuovo DCH ufficialmente riconosciuto. Diversamente, continueremo a considerare digitalizzazione e digitale solo come un modo diverso e accattivante di fruire il tangibile, perdendo di vista quanto invece tutto ciò sia già oggi l'*humus* identitario che, pur a livelli diversi, identifica l'era digitale contemporanea.

Bibliografia

https://www.go-fair.org/fair-principles/

https://www.go-fair.org/implementation-networks/overview/co-operas/

https://operas.hypotheses.org/

http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_ubc_vancouver_declaration_en.pdf

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XG0614%2808%29

http://www.interpares.org/

Agenzia per l'Italia Digitale (AgID), Presidenza del Consiglio dei Ministri, *Linee guide sulla conservazione dei documenti informatici,, Versione 1.0 – dicembre 2015*, pp. 45 ss. http://www.agid.gov.it/sites/default/files/linee_guida/la_conservazione_dei_documenti_informatici_rev_def.pdf

[6] https://it.wikipedia.org/wiki/Resilienza

Lila Bailey. 2015. *Digital Orphans: The Massive Cultural Black Hole On Our Horizon*, Techdirt, Oct 13th 2015 <https://www.techdirt.com/articles/20151009/17031332490/digitalorphans-massive-cultural-blackhole-our-horizon.shtml>.

Nicola Barbuti. 2016. *Le nuove entità culturali digitali tra Intangible Cultural Heritage e Patrimonio Culturale Immateriale,* in *The Creative Network – Conferenza GARR*, Firenze, 30 novembre-02 dicembre 2016 <https://www.eventi.garr.it/it/conf16/home/materiali-conferenza-2016/paper/19-conf2016-paper-barbuti/file>

Nicola Barbuti. 2017. *Dalla Digital Culture al Digital Cultural Heritage: l'evoluzione impossibile?*, in *AIUCD 2017 Conference – Book of Abstract. Il telescopio inverso: big data e distant reading nelle discipline uma-nistiche*, p. 14-17, AIUCD <http://aiucd2017.aiucd.it/wp-content/uploads/2017/01/book-of-abstract-AIUCD-2017.pdf>

Nicola Barbuti, and Ludovica Marinucci. 2018. *Dal Digital Cultural Heritage alla Digital Culture. Evoluzioni nelle Digital Humanities,* DH 2018 <https://dh2018.adho.org/dal-digital-cultural-heritage-alla-digital-culture-evoluzioni-nelle-digital-humanities/>

Enrico Daga, and Leif Isaksen. 2016. *Proceedings of the 1st Workshop on Humanities in the Semantic Web*, co-located with 13th ESWC Conference 2016 (ESWC 2016), Anissaras, Greece, May 29th, 2016 <http://ceur-ws.org/Vol-1608/paper-05.pdf>.

Luciana Duranti, and Elizabeth Shaeffer [ed. by]. 2012. *The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage*, UNESCO Conference Proceedings, 26-28 September 2012, Vancouver <http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf> .

Vincenzo Gambetta. 2009. *La conservazione della memoria digitale*, [Rubano] : Siav.

Pallab Ghosh. 2016. *Google's Vint Cerf warns of 'digital Dark Age'*, BBC News, Science & Environment, 13 February 2016 <http://www.bbc.com/news/science-environment-31450389>

Maria Guercio. 2008. *Gli archivi come depositi di memorie digitali*, «Digitalia», Anno III, n. 2, Rom : ICCU, pp. 37-53.

Maria Guercio. 2013. *Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali*, Roma-Bari : Laterza.

Joint Steering Committee for Development of RDA. 2015. *Resource Description and Access (RDA)* <http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2015/RDA_Traduzione_ICCU_5_Novembre _REV.pdf> .

Wouter Kool, Brian Lavoie, and Titia van der Werf. 2014. *Preservation Health Check: Monitoring Threats to Digital Repository Content*, Dublin (Ohio) : OCLC Research <http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-preservation-health-check-2014.pdf> .

Brian Lavoie, and Richard Gartner. 2013. *Preservation Metadata (2nd edition), DPC Technology Watch Report*, 03 May 2013, DPC Technology Watch Series <http://www.dpconline.org/docman/technology-watch-reports/894-dpctw13-03/file> .

Library of Congress. *PREMIS – Preservation Metadata: Implementation Strategies*, v. 3.0 <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf >.

Gilberto Marzano. 2011. *Conservare il digitale. Metodi, norme, tecnologie*, Milano : Editrice Bibliografica.

Mellon Foundation and Digital Preservation Coalition Sponsor Formation of Task Force for Email Archives, 1 November 2016 <https://mellon.org/resources/news/articles/mellon-foundation-and-digital-preservation-coalition-sponsor-formation-task-force-email-archives/ >.

OCLC. *PREMIS (PREservation Metadata: Implementation Strategies) Working Group*, 2005

<http://www.oclc.org/research/projects/pmwg/>.

Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (F. Berman and B. Lavoie, co-chairs), La Jolla, February 2010 <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf>.

Verso il riconoscimento delle Digital Humanities come Area Scientifica: il Catalogo online condiviso delle pubblicazioni dell'AIUCD.

Nicola Barbuti
Università degli Studi di Bari Aldo Moro
nicola.barbuti@uniba.it

Maurizio Lana
Università degli Studi del Piemonte Orientale
maurizio.lana@uniupo.it

Vittore Casarosa
ISTI-CNR
casarosa@isti.cnr.it

Abstract

English. The recognition of Digital Humanities (DH) as Scientific Sector is a first-level issue in the debate on innovation of scientific research in Italy. This issue is among the most recurrent even within the AIUCD. In this regard, during the 2019 Conference of the Association, the need emerged to launch activities aimed at keeping the attention of political decision makers to this claim. The Assembly opted for a "bottom-up" action: to produce an online digital catalogue of national sector publications for providing concrete elements to support the awareness that DH are no longer a nebula dotted with indistinct and doubled edged scientific entities, but are today a very well established and widespread reality in the panorama of Italian research.

Italiano. Il riconoscimento delle Digital Humanities (DH) quale Area Scientifica autonoma è uno dei temi di maggiore attualità nel dibattito sui nuovi orientamenti della ricerca scientifica in Italia. Anche in seno all'AIUCD, la questione di riconoscere alle DH la dignità di settore specifico di ricerca è tra i più ricorrenti. A riguardo, nel corso del Convegno 2019 dell'Associazione, è emersa la necessità di avviare attività orientate ad attirare l'attenzione dei decisori politici sul tema. Si è dunque optato per un'azione "dal basso": produrre un catalogo digitale on line delle pubblicazioni nazionali di settore, di modo da fornire elementi concreti a sostegno della consapevolezza che le DH non sono più una nebulosa costellata di entità scientifiche indistinte e ancipiti, ma sono ormai una realtà molto ben consolidata e diffusa nel panorama della ricerca italiana.

1 Introduzione

Nell'ambito della Conferenza AIUCD 2019, uno dei temi di maggior confronto, anche in seno all'Assemblea annuale dei soci, è stata la necessità ormai non più rinviabile di attivare iniziative finalizzate a riconoscere alle Digital Humanities – o Humanities Computing che si vogliano ridefinire – la dignità di Area Scientifica, emancipandole definitivamente dalla dimensione di come nebulosa in cui fluttuano in modo frammentario e caotico ricercatori e studiosi rinvenienti da una pletora di SSD tradizionali delle Humanities, in cui sono considerati delle entità scientifiche ancipiti.

L'analisi dello stato dell'arte della disciplina ha evidenziato la consistenza decisamente ampia di contributi scientifici che, pur riferendosi ad ambiti umanistici riconducibili a singoli SSD tradiziona-li, sono confluenti nella comune prospettiva di ricerca su digitale e computazionale applicati alle humanities e, perciò, assolutamente associabili in un un'unica area di contenimento.

Pertanto, al fine di sostanziare l'istanza di riconoscimento delle DH come Area con una sua digni-tà scientifica, si è scelto un indirizzo operativo "dal basso": produrre un catalogo digitale delle pub-blicazioni nazionali di settore che possa essere riconosciuto dall'AIUCD quale proprio riferimento ufficiale e, in prospettiva, possa diventare nodo di un più ampio e condiviso catalogo internazionale di pubblicazioni sulle DH.

Il Gruppo di Lavoro individuato per occuparsi di progettare e strutturare il catalogo è composto dai soci AIUCD Maurizio Lana, Vittore Casarosa e Nicola Barbuti.

Il GdL ha intrapreso le attività immediatamente dopo la chiusura del Convegno 2019 e attualmen-te sta provvedendo agli ultimi passaggi per la realizzazione esecutiva di quanto progettato e proposto all'Associazione agli inizi dell'estate 2019.

2 Il Catalogo delle DH AIUCD

Primo tema di riflessione è stato definire i limiti geografici e cronologici delle pubblicazioni da inse-rire in prima istanza. Si è deciso di limitare inizialmente l'inserimento agli autori italiani con priorità per i soci AIUCD, con un orizzonte temporale non superiore agli ultimi 10 anni. È stata ipotizzata la creazione dello spazio online con credenziali di accesso da condividere con tutti i soci in modo che, per quanto possibile, ognuno possa inserire da sé i record bibliografici relativi alle proprie pubblica-zioni, e di consentire anche l'associazione dei pdf ai record, ove legalmente disponibili. Tuttavia, per conferire al catalogo veste ufficiale, pur consentendo a ciascuno di inserire i dati direttamente, sem-bra opportuno stabilire regole definite e condivise, nominando un organismo deputato a eseguire un controllo annuo sulla coerenza delle nuove risorse bibliografiche caricate per evitare l'insorgere di situazioni caotiche.

Dal momento che una delle principali criticità connesse con la multiforme produzione scientifica delle DH è proprio l'elevata varietà di fonti e quindi di formati citazionali bibliografici, si è passati ad analizzare software open source per la gestione e la fruizione di record bibliografici digitali che avessero le caratteristiche necessarie a favorire un import di risorse di diversa tipologia e struttura.

Primi a essere presi in considerazione sono stati Zotero, in considerazione sia del fatto che l'AIUCD utilizzando questo software aveva già attivato un repository per l'allocazione di risorse di-gitali relative alle DH sia dell'ampio uso che se ne sta facendo per la creazione di bibliografie on line con pubblicazioni relative ad altri ambiti di ricerca scientifica, e Zenodo per il repository delle fonti non altrimenti online.

Zotero nacque proprio come strumento per la gestione di bibliografie di area storica (Cohen, 2008) ed è attualmente usato, per esempio, dall'associazione tedesca di Digital Humanities[1] per uno scopo simile a quello cui stiamo pensando anche per AIUCD o dalla *American School of Classical Studies at Athens* (ASCSA) per catalogare e gestire i metadati di tutte le pubblicazioni della Scuola stessa (libri e articoli)[2]. Tuttavia, quest'ultimo è risultato un open repository per "prodotti della ricerca" ge-nerici ed ha evidenziato il limite non secondario che ogni oggetto caricato deve essere linkato a ma-no al record corrispondente. Zotero invece è mirato a collezioni di articoli e bibliografie di varia na-tura e i dati caricati in uno suo spazio on line non sono soggetti al problema rilevato per Zenodo (O'Donnell, Manola, Manghi, Porter, Esau, Viejou, Rosselli Del Turco, and Singh, 2018; Peters, Kraeker, Lex, Gumpenberger and Gorriaz, 2017). È sufficiente, infatti, che il responsabile del cari-camento tagghi la pubblicazione con l'indicazione del/dei SSD in cui si colloca e con le keywords che egli ritiene ne identifichino correttamente il contenuto per rendere il record interrogabile anche in questo modo oltre che con i consueti criteri di autore, titolo, etc. Inoltre, altre caratteristiche interes-santi di Zotero sono la separazione dei (meta)dati dalla loro presentazione secondo uno stile citazio-nale piuttosto che un altro, la possibilità di esportare in RDF e altri formati open il database della bibliografia.

Altro punto a favore di Zotero è la possibilità di definire modalità di recupero delle informazioni bibliografiche incrociando i seguenti dati:

- dati dei titoli;
- dati delle keywords inserite ufficialmente nei lavori o, in mancanza, indicate espressamente dagli autori rilevando parole poi riscontrabili nel testo;
- dati che possono essere estratti dagli abstract o dai full text dei contributi inseriti;
- SSD degli autori.

Queste rilevazioni possono essere utilizzate per produrre report annui sullo stato dell'arte della ri-cerca scientifica sulle DH da ufficializzare per rimarcare la fertilità produttiva del settore.

Per iniziare a popolare il catalogo, si è concordato di proporre ai soci (e non, purché italiani) vari modi per inserire i dati nella bibliografia (Vahdati, Arndt, Auer and Lange, 2016):

[1] https://www.zotero.org/groups/372575/dhd_ag_publikationen [2] https://www.zotero.org/groups/80651
american_school_of_classical_studies_at_athens

- coloro che usano un Bibliographic Reference Software (BRS: Zotero, Mendeley, Endnote, Bibref, Refworks, etc.), possono esportare i (meta)dati citazionali delle loro pubblicazioni completi di URL a ciascuna pubblicazione, quindi inserirli in Zotero e dare l'accesso pubblico per consentire la fruizione diretta delle risorse caricate;
- coloro che possono utilizzare l'ISBN per le monografie o il DOI per gli articoli possono inseri-re direttamente nella bibliografia online, utilizzando Zotero, i dati delle loro pubblicazioni, completandoli con i tag che indicano il SSD e quant'altro può essere necessario al recupero del record.

Per tutte le forme di pubblicazione grigia (presentazioni, abstract, raccolte di dati, etc.) Zotero non può gestire in modo ottimale i dati perché gestisce per lo più risorse bibliografiche. Giunge utile a questo punto Zenodo, poiché assegna automaticamente un DOI alle pubblicazioni o alle fonti in ge-nere che non lo hanno già, e quindi permette di salvare nel suo repository aperto anche altri prodotti della ricerca, come dati e software (Potter and Smith, 2015), oltre alle classiche pubblicazioni.

Si è dunque concluso di creare il catalogo sfruttando al massimo le diverse opportunità offerte dai due software, creando una soluzione che, non comportando attività di sviluppo software ad hoc, dia ragionevoli prospettive di sostenibilità e permetta agevoli importazione, esportazione e migrazione dei dati.

Di seguito riportiamo l'articolazione dell'ipotesi progettuale, che si articola nelle seguenti fasi.

Base di partenza sarà il catalogo AIUCD già esistente in Zotero, sebbene esso presenti alcuni fisiologici punti di debolezza (duplicazioni, item incompleti, assenza di rimando con link alle fonti dove disponibili in OA, etc.).

I cataloghi Zotero sono accessibili in lettura-scrittura per gli editors, in sola lettura per tutti quelli che hanno il link.

Per facilitare l'acquisizione e l'inserimento dei dati si partirà dalle pubblicazioni che hanno già il DOI, quindi seguiranno, in ordine progressivo, quelle che hanno ISBN, poi quelle con ISSN, infine le pubblicazioni i cui dati devono essere raccolti e inseriti manualmente.

Le pubblicazioni cosiddette "grigie" prive di DOI (a es.: presentazioni) saranno invece preliminarmente caricate in Zenodo, di modo che il sistema le renda accessibili assegnandovi un DOI.

Scegliere di gestire le risorse digitali associando Zotero e Zenodo, oltre a presentare il vantaggio di rendere le pubblicazioni inserite facilmente ricercabili utilizzando chiavi di accesso uniformi, consente di collocare il catalogo in un contesto di ricerca aperta.

Ciascun socio/autore, quindi, potrà provvedere all'inserimento dei propri dati nel modo che segue:

- chi ha pubblicazioni in OA ma senza DOI le carica in Zenodo per ottenere il DOI;
- chi ha pubblicazioni già provviste di DOI comunica la lista dei DOI, uno per riga, mandando un messaggio all'indirizzo email dedicato pubblicazioni@aiucd.it; se la pubblicazione non ha keywords internamente, si possono indicare fino a un massimo di 5 keyword (parole o espressioni) accanto al DOI, separate da virgole; a riguardo, una buona chiave identificativa può essere il (o i) SSD in cui l'autore ritiene si collochi la sua pubblicazione nello spazio cloud di zotero;
- chi ha libri manda gli ISBN;
- chi ha articoli senza DOI manda l'URL.

Le pubblicazioni provviste di DOI saranno caricare in Zotero utilizzando il codice. Il DOI dovrà essere presente e ben visibile nel catalogo pubblico, in quanto è la chiave di accesso principale alla pubblicazione inserita.

Relativamente alle necessità di gestione del catalogo, si prospettano le seguenti soluzioni.

Sarà necessario ridefinire chi avrà accesso in scrittura al catalogo sia per inserire i DOI in Zotero e costruire l'elenco, sia per intervenire a correggere eventuali errori nei dati inseriti. Un accesso indiscriminato, infatti, creerebbe rischi di rumore notevole e soluzioni caotiche e difformi nell'organizzazione dei dati.

Ogni inizio d'anno, i soci saranno invitati a inviare all'indirizzo sopra indicato i DOI e gli ISBN delle nuove pubblicazioni dell'anno precedente.

Dal momento che esiste un problema concreto di autodefinizione sulla pertinenza delle pubblicazione al campo DH, al fine di evitare l'afflusso nel catalogo di pubblicazioni che poco hanno a che fare con il settore sarà necessario che il Direttivo AIUCD definisca delle linee guida sui temi di ricerca coerenti con esso.
Una volta analizzato l'impatto del catalogo nell'ambito della ricerca scientifica (Sample, 2011), si potranno prendere in considerazione altri aspetti che renderanno necessarie opportune implementazioni si pensi, nello specifico, alle pubblicazioni senza codici, che sono in linea di massima quelle più datate e richiedono notevole lavoro per essere inserite nel catalogo.
Resta da definire come agire operativamente, cioè chi si occuperà di inserire i DOI in Zotero per costruire l'elenco: potrebbe essere un'attività laboratoriale per studenti di biblioteconomia/scienze biblioteconomiche e dell'informazione o per i futuri digital librarians?

3 Conclusioni

Tirando le conclusioni sul catalogo progettato, siamo del parere che un'integrazione Zotero-Zenodo in un'unica soluzione on line, sfruttando al massimo l'assegnazione DOI e, in prospettiva, la funzione di repository del secondo, prende il massimo da due mondi open source. Ciò apre opportunità di ana-lisi della ricerca del settore prima impossibili da attuare (Winslow, Rains, Skripsky and Kelly, 2016), e si configura come un elemento qualificante il riconoscimento delle DH come settore di primo livel-lo nella ricerca italiana.

Bibliografia

https://www.zotero.org/groups/372575/dhd_ag_publikationen

https://www.zotero.org/groups/80651/american_school_of_classical_studies_at_athens

Daniel J. Cohen. 2008. *Creating Scholarly Tools and Resources for the Digital Ecosystem: Building Connec tions in the Zotero Project*, «First Monday 13», n. 8. https://doi.org/10.5210/fm.v13i8.2233

Daniel Paul O'Donnell, Natalia Manola, Paolo Manghi, Dot Porter, Paul Esau, Carey Viejou, Roberto Rosselli Del Turco and Gurpreet Singh. 2018. Using Zenodo as a Discovery and Publishing Platform. https://doi.org/10.5281/ zenodo.1297110

Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger and Juan Ignacio Gorraiz. 2017. *Zenodo in the spotlight of traditional and new metrics*, «Frontiers in Research Metrics and Analytics» 2: 13. https://doi.org/10.3389/frma.2017.00013

Megan Potter and Tim Smith. 2015. *Making code citable with Zenodo and GitHu*b, Software Sustainibility Institute. https://www.software.ac.uk/blog/2016-09-26-making-code-citable-zenodo-and-github

Rachel Rains Winslow, S. Skripsky and Savannah L. Kelly. 2016. *Not just for citations: Assessing Zotero while reassessing research.* In: *Information literacy: Research and collaboration across disciplines*, 299–316. Fort Collins, CO: WAC Clearinghouse and University Press of Colorado.

Mark Sample. 2011. *Sharing Research and Building Knowledge Through Zotero*. In: *Learning Through Digital Media Experiments in Technology and Pedagogy*, ed. by Trebor R. Scholz. New York: New School. http://mcpress.media-commons.org/artoflearning/sharing-research-and-building-knowledge-through-zotero/

Sahar Vahdati, Natanael Arndt, Sören Auer and Christoph Lange. 2016. *OpenResearch: Collaborative Management of Scholarly Communication Metadata. In: Knowledge Engineering and Knowledge Management,* ed. by Eva Blomqvist, Paolo Ciancarini, Francesco Poggi and Fabio Vitali, 10024:778–93. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-49004-5_50

Il trattamento automatico del linguaggio applicato all'italiano volgare. La redazione di un *formario* tratto dalle prime dieci *Lettere* di Alessandra M. Strozzi.

Ottavia Bersano
Università degli Studi di Firenze
Universität Bonn
ottavia.bersano@gmail.com

Nadezda Okinina
Eurac Research
Bolzano
nadezda.okinina@eurac.edu

Abstract

English. This paper aims to describe the strategies that have been experimented in the use of Natural Language Processing (NLP) on a Florentine text of the 15th century, that is the Epistolary of Alessandra Macinghi Strozzi. With the help of NLP, and in particular of POS tagging, it was possible not only to facilitate the linguistic analysis of the text by quickly obtaining the most interesting graphic, phonetic and morphological traits, but also to create, automatically, a collection of word forms, alphabetically ordered with the number of occurrences for each item.

Italiano. Questo contributo è teso a illustrare le strategie sperimentate nell'applicazione del Trattamento automatico del Linguaggio (TAL) a un testo di origine fiorentina risalente al sec. XV, l'Epistolario di Alessandra Macinghi Strozzi. Con l'ausilio del TAL, e in particolare con il processo di POS tagging, è stato possibile non solo agevolare l'analisi linguistica del testo ricavandone celermente i tratti grafici, fonetici e morfologici più interessanti, ma realizzarne anche, in modo automatico, un *formario*, ordinato su base alfabetica e comprensivo del numero di occorrenze per ciascun item.

1 Introduzione

Il presente contributo si propone di coniugare le tecniche offerte dal Trattamento automatico del Linguaggio (TAL) allo studio dell'italiano volgare, e nella fattispecie a testi di origine fiorentina del sec. XV. Esso nasce in seno ad alcune difficoltà emerse durante la selezione dei dati testuali di una tesi di dottorato – ancora in corso di stesura – tesa a fornire una nuova edizione dell'Epistolario di Alessandra Macinghi Strozzi (1406-1471): un documento assai conosciuto e apprezzato da chiunque si occupi della civiltà italiana del Quattrocento e che siamo ancora costretti a leggere nell'edizione curata da Cesare Guasti nel 1877 (Guasti, 1877). Una nuova edizione e una nuova analisi linguistico-interpretativa, condotta con gli strumenti più aggiornati, porterà indiscutibili vantaggi anche agli studi di ambito storico e letterario.

2 Le *Lettere* di Alessandra Macinghi Strozzi

Le *Lettere* – in tutto settantatré – furono scritte in un arco temporale compreso tra il 1447 e il 1470 e indirizzate, da Alessandra Macinghi – vedova di Matteo Strozzi – ai figli Filippo, Lorenzo e Matteo. Questi, raggiunta la maggiore età, ereditarono l'esilio paterno e furono costretti a lasciare Firenze: un provvedimento legislativo, infatti, stabiliva che tutti i figli maschi degli esuli, al compimento del diciottesimo anno, ne dovessero ereditare la condizione; lasciata dunque Firenze ed esercitando il mestiere della mercatura, Filippo, Lorenzo e Matteo viaggiarono per tutta Europa, trovandosi a soggiornare nei maggiori centri politico-commerciali del tempo, in particolare a Palermo, Napoli, Barcellona, Madrid, Londra e Bruges.

Il presente Epistolario – che nasce da una naturale quanto necessaria esigenza comunicativa, quella che intercorre tra una madre e i suoi figli – costituisce un raro esempio di scrittura femminile privata del sec. XV e rappresenta non solo una delle prime testimonianze in lingua volgare, ma anche una delle prime testimonianze scritte da una donna laica, le cui possibilità di scolarizzazione, com'è noto, erano al tempo limitatissime.

2.1 La rilevanza linguistica dell'Epistolario

La figura di Alessandra Macinghi Strozzi desta interesse non solo perché dimostra una certa dimestichezza nel padroneggiare la penna, ma anche perché riesce con disinvoltura a cimentarsi in partite e ragioni: un ambito lessicalmente e storicamente maschile. La prosa di Alessandra si contraddistingue per autenticità, schiettezza e onestà di valori pedagogici, tanto che Contini (1970) la definì la "paradigmatica" tra le prose domestiche del Quattrocento; la sua rilevanza linguistica è data anzitutto dal genere letterario cui va ascritta, ovverosia il genere epistolografico: «Le settantatré *Lettere* pervenuteci rispondono a un bisogno immediato, urgente di comunicazione, non hanno né preoccupazione né destinazione letteraria [...]. Sono scritti assolutamente privati, in cui la stessa grammatica è quella parlata; avvisi, massime, ricordi, notizie, resoconti [...] di fatti, avvenimenti, azioni, proposte relative al nucleo familiare, all'intimità della casa, segreti che non devono andare oltre il mittente e il destinatario [...]» (Doglio, 1984, p. 487).

Per quanto concerne il lessico delle *Lettere*, si intende redigere – in appendice e a completamento della nuova edizione – un *Glossario*, che recepirà e metterà prontamente a disposizione del LEI – *Lessico etimologico italiano* (Pfister, 1979-) – numerosi termini specialistici, fra cui primeggiano quelli relativi all'attività commerciale e finanziaria, che ebbe nell'Italia del tardo Medio Evo e del Rinascimento uno straordinario sviluppo. E poiché taluni di questi termini si sono trasmessi anche alle altre lingue europee (si pensi alla fortuna di termini come *banco*, *banchiere*, *capitale*, *credito*, *debito*, *polizza*, *assicurazione*, *lettera di cambio*, ecc.), ne deriverà un contributo molto utile alla storia della terminologia economica sovranazionale tuttora in uso.

Come messo in evidenza più volte dallo Stussi (2000), i documenti mercantili sono una fonte di inestimabile valore per la conoscenza della storia della lingua e offrono talvolta la possibilità di retrodatare parole o, addirittura, intere espressioni.

Una nuova edizione delle *Lettere* della Macinghi Strozzi metterà finalmente a disposizione degli studiosi un testo filologicamente corretto e linguisticamente affidabile, che potrà sostituire l'edizione curata da Cesare Guasti nel 1877, di cui più volte ne sono stati segnalati i limiti (Trifone, 1989). Verrà così per la prima volta alla luce la grafia realmente utilizzata dall'autrice, sistematicamente modernizzata dal Guasti, e si potranno recuperare i diversi caratteri relativi ai suoni e alle forme, che non risultano nella precedente edizione. Le *Lettere* si prestano inoltre a interessanti rilievi di ordine sociolinguistico, essendo la scrivente una donna di ceto mercantile, assimilabile alla classe dei cosiddetti "semicolti".

3 L'applicazione del TAL per la redazione del *formario*

Per il presente contributo è stato realizzato un primo modello di *formario* tratto dalle prime dieci *Lettere* di Alessandra M. Strozzi, teso a mettere in luce i tratti grafici, fonetici e morfologici caratterizzanti la lingua della scrivente, autentica espressione del fiorentino *argenteo* (cfr. Manni, 1979); trattandosi di un archetipo, lo studio si è limitato all'analisi delle prime dieci lettere dell'*Epistolario*: ciò ha consentito di eseguire un lavoro accurato, ma soprattutto di riflettere sui benefici e sulle criticità date dall'utilizzo degli strumenti offerti dal TAL.

Alcune rese grafiche dell'originale (del tipo *x* per *s*, *y*/*j* per *i*, *cha* per *ca*, *cho* per *co*, *chu* per *cu*) sono state anzitutto normalizzate conformemente alla grafia moderna; è stata successivamente realizzata una tokenizzazione attraverso l'impiego del modello TreeTagger per l'italiano contemporaneo elaborato da Achim Stein (Schmid, 1994). Si è quindi proceduto con l'assegnazione di un POS per ogni forma delle prime dieci *Lettere*, il cui *corpus* consta, in tutto, di 13.782 occorrenze. Il tagset impiegato, elaborato sulla base del tagset ELRA, è stato semplificato e adattato alle esigenze dell'analisi linguistica eseguita, della quale vengono esposti gli esiti al § 4; il tagset adottato è illustrato nella tabella sottostante, la quale pone inoltre in evidenza le corrispondenze tra le etichette impiegate da chi scrive e quelle proprie degli analizzatori adoperati.

Come si evince dalla tabella, è stata operata una distinzione tra articolo determinativo e indeterminativo; per quel che riguarda le congiunzioni, invece, si è preferito non introdurre alcuna distinzione tra congiunzione subordinante e coordinante, dato che nella lingua del tempo la congiunzione assume una funzione che sovente non è possibile classificare con certezza, giacché polivalente. Le preposizioni, ripartite in semplici e articolate, constano di una terza etichetta, "prep.", sotto la quale sono state fatte confluire le preposizioni improprie.

SPIEGAZIONE	ETICHETTA	CORRISPONDENZE ELRA	CORRISPONDENZE TREETAGGER	CORRISPONDENZE TINT
aggettivo	agg.	AS, AP,AN, DP, DN, DS	ADJ	A, AP, DI, PI
antroponimo	antrop.	SPN	NPR	SP
articolo determinativo femminile	art.det.f.	RS, RP	DET:def	RD
articolo determinativo maschile	art.det.m.	RS, RP	DET:def	RD
articolo indeterminativo femminile	art.indet.f.	RS, RP	DET:indef	RI
articolo indeterminativo maschile	art.indet.m.	RS, RP	DET:indef	RI
avverbio	avv.	B	ADV	B
congiunzione	cong.	C*	CON	CC, CS
interiezione	int.	I	INT	I
non verbale	x	X*	SENT, SYM, PON, FW, LS	FB, FF, FS
numerale	num.	N	NUM	NO
preposizione	prep.	E	PRE	E
preposizione articolata	prep.art.	ES, EP	PRE:det	E+RD
preposizione semplice	prep.sempl.	E	PRE	E
pronome	pron.	P*, Q*	PRO*	PC, PD, PE, DQ
pronome femminile	pron.f.	P*, Q*	PRO*	PC, PD, PE, DQ
pronome maschile	pron.m.	P*, Q*	PRO*	PC, PD, PE, DQ
sostantivo femminile	s.f.	SS, SP, SN	NOM	S
sostantivo maschile	s.m.	SS, SP, SN	NOM	S
toponimo	top.	SPN	NPR	SP
verbo	v.	V*	VER*	V*

Tabella 1. Corrispondenze delle etichette grammaticali.
Si segnalano mediante l'asterisco (*) le etichette trascritte nella forma base, senza i dettagli offerti dai tools.
L'etichetta TreeTagger "PRO:demo", per esempio, è stata riportata "PRO*".

Dal momento che nessun POS tagger dispone di un modello per l'italiano volgare, il processo di POS tagging è risultato piuttosto articolato e, per ovviare all'assenza del modello, sono state sperimentate due diverse strategie: la prima è consistita nell'applicare il modello POS tagger addestrato sul *corpus* D(h)ante (Basile e Sangati, 2016), la seconda nel normalizzare il *corpus* delle *Lettere*, così da poter impiegare un POS tagger predisposto per l'italiano contemporaneo. Al fine di eseguire un valido confronto fra le due strategie, sono state manualmente etichettate le prime 1.000 parole del *corpus* delle *Lettere*, le cui assegnazioni sono state comparate con gli esiti dati dai due processi sopra descritti. Impiegando dunque TreeTagger e il parser Stanford CoreNLP – addestrati da A. Basile e F. Sangati sul *corpus* D(h)ante – sono state assegnate le parti del discorso. TreeTagger ha dato un risultato migliore rispetto al parser Stanford: il primo ha dato infatti il 59% dei tag corretti contro il 54% del secondo. Prima di approdare alla seconda strategia, il testo delle *Lettere* è stato normalizzato utilizzando il correttore ortografico GNU Aspell (http://aspell.net) e quindi nuovamente etichettato mediante il modello TreeTagger elaborato da Achim Stein e addestrato sull'italiano contemporaneo, il cui impiego ha consentito di approdare a una percentuale di tag corretti pari al 69%. I migliori risultati di POS tagging, tuttavia, sono stati ottenuti utilizzando il tagger per l'italiano contemporaneo Tint (Aprosio e Moretti, 2018) che, una volta applicato al testo precedentemente normalizzato, ha restituito il 72% delle etichette corrette.

Programma	Stanford D(h)ante	TreeTagger D(h)ante	Aspell+TreeTagger (Stein)	Aspell + Tint
Accuratezza	54%	59%	69%	72%

Tabella 2. Descrizione dell'accuratezza dei programmi impiegati nell'assegnazione dei tag alle prime 1000 parole del *corpus* delle *Lettere*.

Nell'attribuzione delle etichette le parti del discorso che hanno presentato maggiori difficoltà sono state aggettivi e pronomi, indipendentemente dal programma impiegato. L'etichettatore Stanford è inoltre risultato particolarmente debole e impreciso nel riconoscimento dei verbi, attribuendo erroneamente tale etichetta a molte altre parti del discorso.

Al fine di incrementare la percentuale di tag corretti, il sistema è stato perfezionato attraverso l'impiego dei dati derivanti dal dizionario TLIO (*Tesoro della Lingua italiana delle Origini*, http://tlio.ovi.cnr.it/TLIO/), che ha consentito di attribuire le etichette grammaticali a 5.194 forme e grazie al quale è stato possibile ricavare, inoltre, una puntuale distinzione di genere per 606 sostantivi. Per i sostantivi restanti, che non hanno trovato riscontro all'interno del dizionario TLIO, sono state elaborate alcune regole, basate su una serie chiusa di articoli e aggettivi, a seconda che questi accompagnino sostantivi femminili o maschili; tali regole – per l'elaborazione delle quali sono state manualmente redatte delle liste, comprensive di serie di articoli e aggettivi – hanno consentito di assegnare ad altri 120 sostantivi la distinzione di genere, precedentemente mancante.

Prossimamente si ritiene opportuno impiegare un analizzatore morfologico per l'identificazione automatica del genere delle parole, così da verificarne il grado di correttezza; si intende inoltre sperimentare una terza strategia – simile a quella applicata per il POS tagging del *corpus* MIDIA (Iacobini et al., 2014) –, finalizzata a perfezionare il lessico di TreeTagger per l'italiano contemporaneo attraverso le voci provenienti dal dizionario TLIO.

Per favorire il riconoscimento di antroponimi e toponimi, inoltre, sono stati impiegati i dati ricavati dal *Glossario* del *Libro dei debitori, creditori e ricordi* che Alessandra Macinghi Strozzi tenne tra il 1453 e il 1473, un testo dunque coevo all'Epistolario oggetto del presente studio e di mano della stessa Alessandra Macinghi Strozzi (Bersano, 2015-16, pp. 271-294).

Dal *corpus* annotato è stato tratto automaticamente un *formario*, strutturato in ordine alfabetico; sulla base di queste due fonti è stata eseguita una breve quanto esaustiva analisi linguistica, di cui si riportano gli esiti più significativi nel paragrafo successivo.

4 Esiti

Per garantire la completa affidabilità dei risultati, è stato necessario confrontarsi costantemente con la trascrizione originale – specie nella fase iniziale del TAL – onde evitare errori nel processo iniziale di trasmissione dei dati – *input* –, e poter essere certi dell'attendibilità dei dati in uscita, *output*.

Il *formario* tratto dal testo delle *Lettere* e realizzato grazie al processo sinora descritto, ha consentito a colpo d'occhio di cogliere i fenomeni e i tratti grafici, fonetici e morfologici tipici del fiorentino *argenteo*, così come di porre in luce quelli che ne esulano, presentandosi inaspettatamente 'controcorrente'. Qui di séguito una breve illustrazione degli esiti linguistici grafici e morfologici più interessanti ricavati dal *corpus* annotato e dal *formario*.

GRAFIA: si rileva un'oscillazione costante per la resa della *l* palatale, con il primeggiare del tipo *gl*, anche dinanzi a *i* (seppure risultino consistenti le rese grafiche *lgl*, rare invece quelle in *li*); meno incerta sembrerebbe la resa grafica della *n* palatale, per la quale primeggia il tipo *ngn* (38 occorrenze), più sparute le rese *ngni*, *gn*; rarissima *gni* (con 3 sole occorrenze); dinanzi a *i* risulta nuovamente schiacciante il tipo *ngn* (16 occorrenze) rispetto a *gn* (2 sole occorrenze).

Ancora, è da evidenziare l'uso della grafia *k* per l'occlusiva velare sorda dinanzi ad *a* nella voce *Karisimo*; tale resa grafica per l'occlusiva velare è fatto notevole: essa risulta infatti del tutto assente nel *Libro dei debitori, creditori e ricordi di Alessandra Macinghi Strozzi* (Bersano, 2015-16, p. 166). Scorrendo gli item presenti alla lettera *h*, spicca la scrizione etimologica *homo*, impiegata due volte in queste prime dieci *Lettere* (8 attestazioni in tutto, invece, per *uomo*, conforme alla grafia moderna).

MORFOLOGIA: notevoli sono i plurali in *-(l)gli* < *-li*; alcuni esempi: *begli* (*-lgli*) con 3 occorrenze e *fanciugli* (*-lgli*) con 2 occorrenze, senza esiti contrari.

Per l'art. det. masch. sing. prevale la forma *il*, sebbene risulti considerevole anche la presenza della variante *el* (si contano in tutto, rispettivamente, 88 e 13 occorrenze), penetrata nel fiorentino intorno alla seconda metà del sec. XIV per influsso dei dialetti occidentali e meridionali (Manni,1979). Per il plurale dell'art. det. è attestata la forma *e* in luogo di *i*, il sistema ha tuttavia etichettato tutte le *e* presenti (550 occorrenze) come *congiunzioni*; occorrerà senz'altro ovviare all'errore, insegnando alla macchina che *e* può essere anche articolo se posta dinanzi a un sostantivo masch. pl., così che il sistema offra la possibilità di scegliere fra le due etichette: *cong.* oppure *art.det.m.*

Per l'uso di *mie*, *tuo*, *suo* invariabili (del tipo, *mie bisongni*) e *mia*, *tua*, *sua* pl. masch. e femm. (del tipo, *mia figluoli*), è risultato maggiormente utile consultare il file contenente il testo delle *Lettere* etichettato automaticamente dal sistema piuttosto che il *formario*, poiché quest'ultimo è privo dei contesti, essenziali al fine di verificare a quale sostantivo pronomi e aggettivi si accordino e dunque comprendere, come nel caso presente, l'uso dei possessivi sopraccitati.

Molto agevole è stata la ricerca delle occorrenze per i numerali *duo* (prevalente) e *dua* in luogo di *due*: le 12 occorrenze di *duo* sono state erroneamente etichettate come *sostantivi*; le 3 occorrenze di *dua* sono state correttamente etichettate come *num.* (numerale) secondo il tagset elaborato da chi scrive; le 7 occorrenze di *due* sono state invece etichettate come *aggettivi.*

VERBI: è attestata una volta soltanto la forma *sete* per *siete*; anche in questo caso, è stato essenziale consultare il file contenente il testo delle *Lettere* etichettato automaticamente dal sistema piuttosto che il *formario*, così da verificare il contesto e sciogliere ogni perplessità rispetto al valore semantico della parola (*sete* sostantivo vs. *sete* verbo).

Schiacciante è la presenza dei tipi *arò*, *arei* per *avrò*, *avrei*, che non presentano esempi contrari; tutte le attestazioni, inoltre, sono state correttamente etichettate come *verbi*, a eccezione dei tipi *aresti* (4 occorrenze in tutto) e *arebe* (un'occorrenza) classificati erroneamente come *aggettivi*.

Non sono presenti esempi contrari ai tipi *dia* e *stia*; il primo, che consta di 18 occorrenze in tutto, risulta due sole volte erroneamente etichettato come *sostantivo*; il tipo *stia* occorre una volta soltanto ed etichettato regolarmente.

Infine, a riprova della conformità della lingua di Alessandra al fiorentino *argenteo*, compaiono senza esempi contrari i tipi *fussi* per *fossi* e *fusti* per *fosti*.

Per quel che concerne le desinenze verbali, non è possibile in questa sede offrirne una panoramica esaustiva; basti dire, tuttavia, che grazie al supporto informatico, l'analisi linguistica tesa a individuare le forme nonché il numero di occorrenze per ciascuna desinenza verbale è stata di facile esecuzione, oltre che rapida e accurata. Si segnala una desinenza atipica per la lingua di Alessandra, riscontrata nel *formario* e verificata nell'originale: il tipo *preghiamo* per la 1° pers. pl. del pres. ind.; tale occorrenza è un *unicum* in tutto il testo delle *Lettere*, poiché la scrivente adotta uniformemente la desinenza *-no* per la 1° pers. pl. (del tipo *noi laviano*), tanto nelle *Lettere* quanto nel *Libro dei debitori, creditori e ricordi di Alessandra Macinghi Strozzi* (cfr. Bersano, 2015-16, p. 233).

Si riscontrano 3 attestazioni per la forma metatetica *drento* (Manni, 1979) – di cui una con raddoppiamento dell'occlusiva postconsonantica: *drentto* – in luogo di *dentro* e un'attestazione per la forma metatetica *grillanda*, tipiche del fiorentino *argenteo*. Sono inoltre attestate le forme metatetiche *adrieto*, *adrietro* (con mancata dissimilazione *r-r* in *r-ø*), *indrieto* e, più rara per questo periodo, *dirieto*, derivanti dall'influsso esercitato da altri dialetti toscani (Manni, 1979, pp. 167-168).

Significativa è ancora l'attestazione di *sun* in luogo di *su* nel tipo *in sun un*, con l'inserzione della *n* eufonica in *sun*.[1]

Per quel che concerne antroponimi e toponimi, ne sono stati riconosciuti in tutto 346, grazie ai dati ricavati dal *Glossario* del *Libro di debitori, creditori e ricordi* di Alessandra (Bersano, 2015-16); 127, invece, non sono stati riconosciuti secondo le etichette prestabilite ('antrop.' e 'top.'); di questi 127, tuttavia, 43 sono stati riconosciuti ed etichettati come *nomi propri*; 112 sono stati invece etichettati più genericamente come *sostantivi* e solo 4, erroneamente, come *verbi*.

[1] Non si hanno esempi, in queste prime dieci *Lettere*, per il tipo *sur*, che trae origine da *sun* per dissimilazione (*in sun un* > *in sur un*).

5 Conclusioni

Il *corpus* annotato, da cui è stato tratto il *formario* impiegato per l'analisi linguistica, costituisce non solo una proficua base di lavoro per la redazione del futuro *Glossario* delle *Lettere* di Alessandra Macinghi Strozzi, ma anche un primo quanto fondamentale strumento di analisi; esso ha infatti consentito di selezionare celermente i dati più significativi che andranno opportunamente registrati nel *Glossario* finale, che sarà posto in appendice alla nuova edizione delle *Lettere* di Alessandra Macinghi Strozzi.

Il lavoro di revisione effettuato a mano è stato senz'altro ingente; si ritiene tuttavia che il supporto informatico sia stato essenziale al fine di porre in evidenza taluni tratti, così come la quantità di occorrenze per ogni forma riscontrata: fattore indispensabile, quest'ultimo, per chiarire quanto un fenomeno, o un tratto, fosse frequente nella lingua dell'epoca.

Bibliografia

Alessio Palmero Aprosio, Giovanni Moretti. 2018. *Tint 2.0: an All-inclusive Suite for NLP in Italian*, in «Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)».

Angelo Basile, Federico Sangati. 2016. *D(h)ante: A New Set of Tools for XIII Century Italian*. Edito online sul sito della Fondazione Bruno Kessler e accessibile al seguente indirizzo: https://dh.fbk.eu/D%28h%29ante

Ottavia Bersano. Tesi di laurea, a.a. 2015-16. *Il libro dei debitori, creditori e ricordi di Alessandra Macinghi Strozzi (1453-1473). Analisi linguistica*. Tesi di laurea, Università degli Studi di Firenze.

Gianfranco Contini. 1970. *Letteratura italiana del Quattrocento*. Firenze, Sansoni.

Maria Luisa Doglio. 1984. *Scrivere come donna: fenomenologia delle 'Lettere' familiari di A. Macinghi Strozzi*, in «Lettere italiane», XXXVI, pp. 484-97.

Progetto GNU (a cura di). 2000-. *GNU Aspell* (http://aspell.net).

Cesare Guasti (a cura di). 1877. *Lettere di una gentildonna fiorentina del sec. xv ai figliuoli esuli*, Firenze, Sansoni.

Claudio Iacobini, Aurelio De Rosa, Giovanna Schirato. 2014. *Part-of-Speech Tagging Strategy for MIDIA: a Diachronic Corpus of the Italian Language,* in «Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)», pp. 213-218.

Paola Manni. 1979. *Ricerche sui tratti fonetici e morfologici del fiorentino quattrocentesco*, in «Studi di Grammatica italiana», VIII, pp. 115-171.

Max Pfister (a cura di). 1979-. *LEI - Lessico etimologico italiano*, Wiesbaden, L. Reichert, edito per incarico della Commissione per la Filologia romanza da Max Pfister.

Helmut Schmid. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*, in «Proceedings of International Conference on New Methods in Language Processing», Manchester, UK.

Alfredo Stussi. 2000. *Studi di filologia e letteratura italiana in onore di Gianvito Resta*, 2 voll., Salerno Editrice, Roma.

Opera del Vocabolario italiano – OVI (a cura di). 1997-. *Tesoro della Lingua italiana delle Origini (TLIO)*, pubblicazione periodica online consultabile al seguente link: http://tlio.ovi.cnr.it/TLIO/

Pietro Trifone. 1989. *Sul testo e sulla lingua delle lettere di Alessandra Macinghi Strozzi*, in «Studi linguistici italiani», xv, pp. 65-99.

Annotazione semantica e visualizzazione di un corpus di corrispondenze di guerra

Beatrice Dal Bo, Francesca Frontini, Giancarlo Luxardo
Praxiling UMR 5267
Université Paul-Valéry Montpellier 3 - CNRS
France
name.surname@univ-montp3.fr

Abstract

English. This paper introduces Corpus 14, a corpus of correspondences between French soldiers and their relatives during the Great War. We describe the digital edition and its encoding in TEI, as well as the ongoing activities related to indexing and referencing of places, persons and other named entities, undertaken in order to represent the network of correspondences by means of a geovisualisation.

Italiano. Questo articolo presenta Corpus 14, un corpus di corrispondenze tra i soldati francesi della Grande Guerra e le loro famiglie. Descriviamo l'edizione digitale e la sua codifica in TEI, nonché i lavori attualmente in corso per indicizzare e referenziare luoghi, persone ed altre entità nominate, al fine di poter rappresentare la rete di scambi epistolari attraverso una visualizzazione grafica di tipo spaziale.

1 Introduzione

Il progetto Corpus 14, iniziato in concomitanza con il centenario della Grande Guerra, nasce dalla volontà di studiare la lingua delle persone comuni all'inizio del XX secolo, ed in particolare degli scriventi *peu lettrés*, che potremmo tradurre, seguendo la letteratura italiana, con *semicolti* (D'Achille, 1994). In questo contesto si tratta dei *Poilus*, soldati francesi della Grande Guerra, spesso provenienti dalle campagne e da contesti rurali, ancora in parte dialettofoni, che si confrontano spesso per la prima volta con il testo scritto. Se lo studio delle competenze linguistiche e pragmalinguistiche è alla base della raccolta delle loro corrispondenze, tali documenti si dimostrano essere una fonte interessante anche per altre discipline, con interessanti informazioni di carattere storico, geografico e culturale. In particolare l'interesse si è focalizzato su due ambiti:

- lo studio della Grande Guerra e della sua eredità in termini di memoria sociale, e delle trasformazioni da essa prodotte,
- l'evoluzione degli usi linguistici, in particolare per quanto riguarda l'influenza delle varietà regionali (in particolare per le zone come il Sud della Francia o la Bretagna, caratterizzate da diglossia), o lo sviluppo di un socioletto comune, il cosiddetto *argot des poilus*.

Utilizzando materiale proveniente da archivi pubblici, nonché documenti donati da eredi al progetto, Corpus 14 si compone ad oggi (versione 2.0[1]) di 37 scriventi, provenienti da 11 regioni diverse, per un totale di 1.797 lettere e circa 500.000 parole. I criteri di selezione del corpus sono stati i seguenti:

- la selezione di scriventi che non hanno completato la formazione elementare,
- la preferenza per le corrispondenze complete, o che per lo meno permettessero di seguire gli scriventi su un arco temporale lungo, e che potessero dunque dare luogo a reti di corrispondenze complesse. Al momento Corpus 14 è costituito di 11 reti di corrispondenze, raggruppate per zona geografica e nominate secondo i luoghi di origine (si veda Figura 1)

[1] https://hdl.handle.net/11403/corpus14

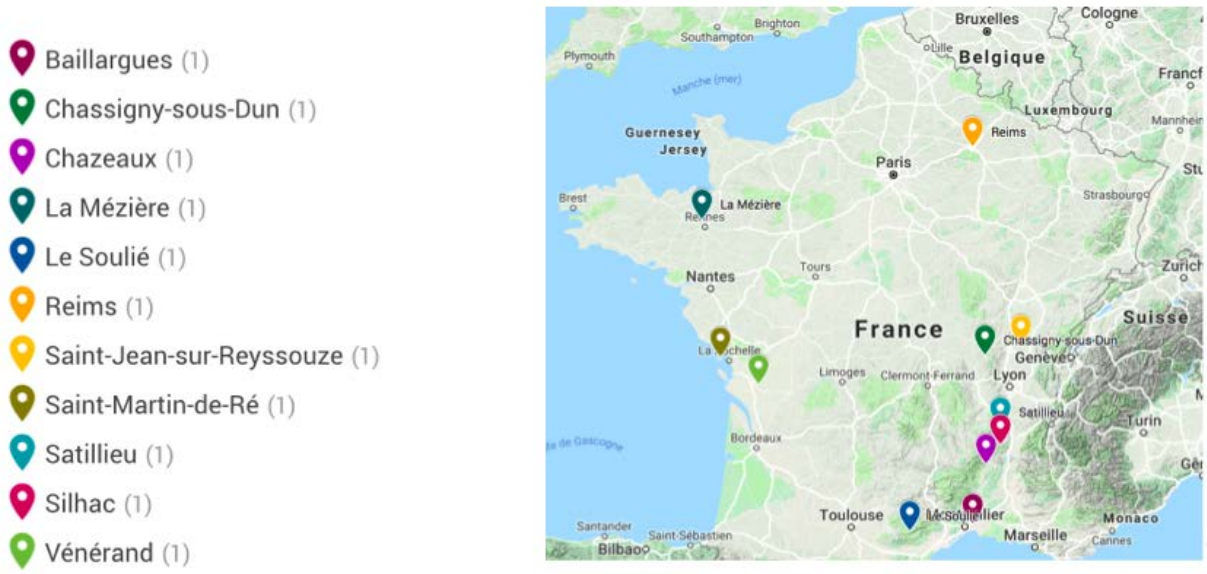


Figura 1: Localizzazione dei luoghi di origine dei soldati di Corpus 14.

Tali criteri di selezione fanno dei fondi di Corpus 14 una collezione unica nel suo genere. Tuttavia la sua realizzazione si ispira anche a progetti faro nella comunità delle edizioni digitali di corrispondenze, molti dei quali dedicati agli epistolari di personaggi illustri[2], con poche eccezioni, come: "Digitising experiences of migration: the development of interconnected letter collections" di Moreton e Nesi, 2013-2014[3]. Inoltre, per la tematica il progetto può essere accostato ad altri omologhi sviluppati in diversi paesi europei in occasione del centenario della Prima Guerra Mondiale, come l'italiano "Voci della Grande Guerra"[4] ed il britannico "Letters from the First World War"[5].

2 L'edizione digitale

L'edizione digitale si è avvalsa di pratiche già ben stabilite, come la trascrizione diplomatica del testo, l'allineamento tra i facsimile delle cartoline o delle lettere e la loro codifica precisa (con precisazioni sulla leggibilità del testo).

Per quanto riguarda la codifica, si è fatto appello allo standard della Text Encoding Initiative (TEI[6]). In particolare le trascrizioni sono state effettuate in modo da permettere la descrizione della struttura logica del testo, nonché delle caratteristiche di leggibilità del supporto fisico. Per ogni lettera sono state realizzate due versioni (fedele e normalizzata all'ortografia corrente).

L'applicazione di questo schema di annotazione XML alla tipologia testuale in oggetto è stato facilitato dall'esistenza di un gruppo di lavoro sulle corrispondenze in seno alla comunità TEI[7]. In particolare si è fatto ricorso agli elementi *TEIheader*, *correspDesc* e *CorrespAction* (introdotti nella versione 2.8.0 delle specifiche TEI P5).

Per quanto riguarda la distribuzione, Corpus 14 è reso disponibile in diverse modalità di accesso che garantiscono la fruizione da parte di tipologie di utenti diverse. Da una parte si è voluto fornire un'interfaccia di esplorazione[8] ed analisi del testo attraverso la piattaforma di testometria TXM (si

[2]Uno dei progetti più noti in questo senso è *Mapping the Republic of Letters*, http://republicofletters.stanford. edu/; per una ricognizione più completa di tali progetti si veda (Stadler et al., 2016)

[3]http://lettersofmigration.blogspot.com; per ulteriori informazioni si veda (Moreton et al., 2014; Moreton, 2016)

[4]http://www.vocidellagrandeguerra.it/

[5]https://www.nationalarchives.gov.uk/education/resources/letters-first-world-war-1915/

[6]https://www.tei-c.org

[7]*Special Interest Group della TEI* sulle corrispondenze https://tei-c.org/activities/sig/correspondence/

[8]http://textometrie.univ-montp3.fr/

veda la Figura 2)[9]. Allo stesso tempo i sorgenti TEI sono scaricabili dalla piattaforma Ortolang[10], che garantisce l'interoperabilità dei dati, la loro preservazione e la loro reperibilità (tramite il protocollo OAI-PMH).

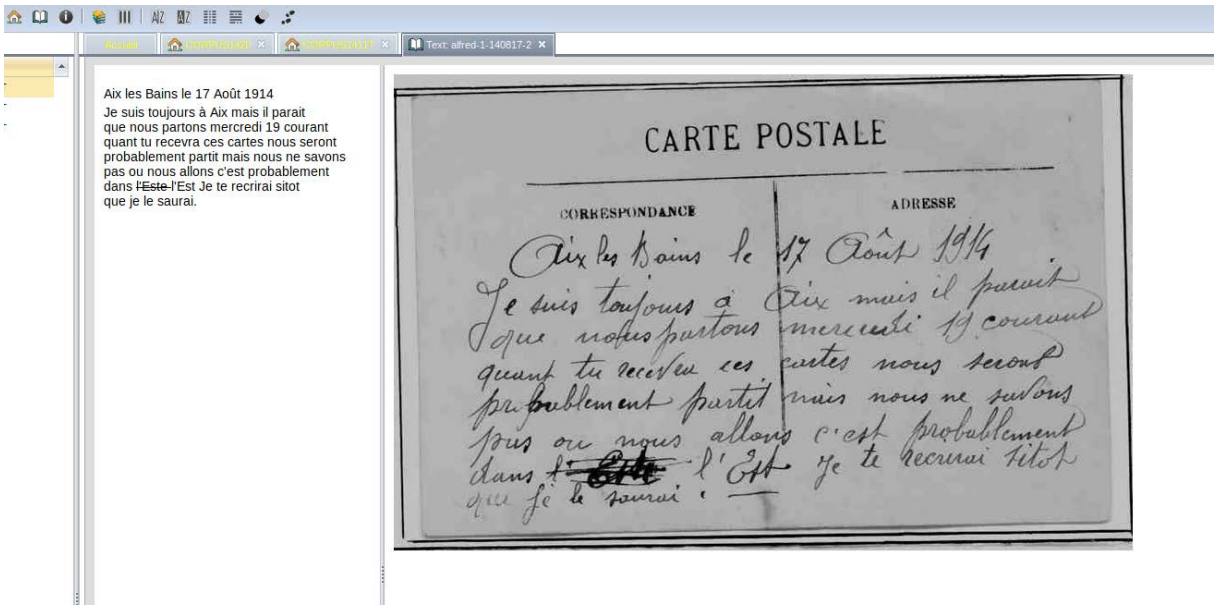


Figura 2: L'interfaccia di esplorazione del corpus TXM.

3 L'indicizzazione semantica dei testi

Una volta realizzata la prima versione dell'edizione digitale, si è posto il problema di arricchire e indicizzare i testi, ed in particolare di creare indici a persone, luoghi, organizzazioni citate. Tali indici, collegati ai riferimenti nel testo, dovranno poi essere arricchiti e collegati con l'informazione corrispondente disponibile.

L'indicizzazione dei testi, che per ora esiste solo su due reti di corrispondenze (*Chazeaux* e *Le Soulié*), è stata condotta secondo le buone pratiche della codifica in TEI, che si sono delineate anche nel contesto di gruppi di lavoro francesi facenti riferimento al consorzio CAHIER[11], come il progetto *Testaments de poilus*[12].

In particolare le menzioni di luoghi, persone e organizzazioni sono state dapprima annotate nel testo di ogni lettera (sia nei metadati della corrispondenza che nel corpo della lettera), utilizzando gli elementi TEI *persName*, *placeName*, *orgName*. Si è inoltre scelto di annotare oltre ai nomi propri anche stringhe di testo aventi nel contesto della lettera dei referenti univoci, usando l'elemento *rs*. L'annotazione è stata effettuata in maniera ricorsiva, dunque un'espressione come "les cousins de Cicignan" è stata annotata come una *rs*, contenente un *placeName*.

In seguito ogni menzione è stata referenziata con l'attributo *ref* e un codice univoco. Tale codice rinvia a tre indici, file separati contenenti delle liste di persone, luoghi, organizzazioni (*listPerson*, *listPlace*, *listOrg*). Per il referenziamento a DBpedia si è utilizzato il sistema di riconoscimento automatico di entità nominate REDEN Online (Résolution et Désambiguïsation d'Entités Nommées) (Frontini et al., 2016), con postcorrezione manuale.

Infine, tali liste sono state dove possibile arricchite con informazioni addizionali in nostro possesso (come le date e i luoghi di nascita e morte delle persone, scriventi o solo menzionate, il loro grado di

[9]TXM è uno strumento per l'esplorazione e l'analisi statistica di corpora testuali, sviluppato dall'ENS di Lione. Permette tra le altre cose l'import di testi annotati in TEI. Si veda http://textometrie.ens-lyon.fr.

[10]ORTOLANG, Outils et Ressources pour un Traitement Optimisé de la LANGue è la piattaforma francese per la pubblicazione delle risorse linguistiche, ora integrata all'infrastruttura CLARIN ERIC. https://www.ortolang.fr/

[11]CAHIER, Corpus d'Autueur pour les Humanités Numériques, è un consorzio di progetti aliati all'infrastruttura Huma-Num, che si occupa di edizioni digitali principalmente in TEI. Si veda https://cahier.hypotheses.org/

[12]https://testaments-de-poilus.huma-num.fr/

parentela, ecc.). Per quanto riguarda i luoghi si è fatto ricorso alla georeferenziazione e all'aggiunta di link al database geografico esterno GeoNames, oltre a quanto già referenziato su DBpedia. In alcuni casi, toponimi non presenti nelle basi sono stati individuati e localizzati. In alcuni casi, toponimi non presenti nelle basi sono stati individuati e localizzati.

4 Visualizzazione

Attualmente in corso è lo sviluppo di una piattaforma di visualizzazione, che permetterà di esplorare le corrispondenze in maniera geolocalizzata[13]. Come si può vedere dalla Figura 3, l'interfaccia permette di selezionare gli scambi epistolari di una stessa rete familiare per data, proiettando sulla carta ad esempio la lettera di un soldato e la risposta della moglie. Nella visualizzazione i segnaposto indicano il luogo di invio della lettera, mentre le bandierine indicano i luoghi citati nella lettera. La visualizzazione è realizzata in modo da sfruttare al massimo lo standard TEI recuperando i *placeName* con interrogazioni basate su XQuery (sia all'interno dei metadati corrispDesc che nel corpo della lettera) e utilizzando la geocodifica degli indici. In questo modo, una volta terminata, la piattaforma potrà essere riutilizzata come base per altri progetti con lo stesso formato di annotazione[14].

Figura 3: L'interfaccia di visualizzazione e geolocalizzazione delle corrispondenze.

5 Analisi

Numerose analisi sono state condotte su Corpus 14, si cita in particolare il volume collettivo curato da Agnès Steuckardt (Steuckardt, 2015a), nel quale sono analizzati vari aspetti linguistici di queste corrispondenze, fra cui la punteggiatura (Steuckardt, 2015b), l'ortografia (Pellat, 2015), il lessico (Luxardo, 2015) e la lingua regionale (Géa, 2015). Ricordiamo inoltre altri studi riguardanti aspetti morfosintattici (Steuckardt and Dal Bo, 2018) o discorsivi (Dal Bo and Wionet, 2018). Una tesi di dottorato è attualmente in corso di completamento (Dal Bo, 2019).

[13]La rappresentazione delle reti di corrispondenze attraverso visualizzazioni è una componente tipica di questo tipo di progetti. Si vedano ad esempio (O'Leary and Moreton, 2017); *Visual Correspondence*, http://www.correspondence.ie; *Mapping the Republic of Letters*, http://republicofletters.stanford.edu/; *Early Modern Letters Online*, http://emlo.bodleian.ox.ac.uk/home; *Clavius on the Web*, http://claviusontheweb.it/.

[14]Il progetto di interfaccia è stato realizzato da studenti del corso di laurea specialistica in informatica dell'Università di Genova, sotto la supervisione della prof. Marina Ribaudo.

Per quanto riguarda gli aspetti spaziali, l'analisi dei luoghi citati nelle corrispondenze dei soldati ha permesso di mettere in evidenza il fatto che questi evochino nelle lettere in maniera prevalente luoghi legati alla loro vita familiare, alla casa e agli affetti, e molto meno luoghi legati alla guerra e al fronte (come già messo in evidenza da (Dal Bo and Wionet, 2018; Gibelli, 2016)). La proiezione su una carta geografica delle informazioni geografiche e temporali delle corrispondenze permette inoltre di seguire gli spostamenti dei soldati al fronte e delle donne rimaste all'interno del Paese. Gli spostamenti di quest'ultime, più raramente studiati, potranno essere inoltre paragonati a quelli delle donne appartenenti a classi sociali superiori durante lo stesso periodo storico.

Bibliografia

Paolo D'Achille. 1994. L'italiano dei semicolti. In *Storia Della Lingua Italiana*, Einaudi, Torino, volume 2, pages 41–79.

Beatrice Dal Bo. 2019. *Aux Frontières de La Norme : Usages Linguistiques de Scripteurs Peu Lettrés Dans Des Correspondances de La Grande Guerre*. Ph.D. thesis, Université Paul-Valéry Montpellier 3.

Beatrice Dal Bo and Chantal Wionet. 2018. Alleviare l'assenza : La modalità ingiuntiva in alcune lettere di donne peu-lettrées durante la Grande Guerra. In Fabio Caffarena and Nancy Murzilli, editors, *In Guerra Con Le Parole. Il Primo Conflitto Mondiale Dalle Testimonianze Scritte Alla Memoria Multimediale*, Fondazione Museo Storico del Trentino, Trento, pages 187–201.

Francesca Frontini, Carmen Brando, and Jean Gabriel Ganascia. 2016. REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pages 193–197.

Jean-Michel Géa. 2015. Le dialecte dans l'écriture de la Guerre : la parte absente ? In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 53–65.

Antonio Gibelli. 2016. *La guerra grande: Storie di gente comune*. Laterza, Bari.

Giancarlo Luxardo. 2015. Fréquences des colis et marmites : comment mesurer la languitude ? In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 113–123.

Emma Moreton, Niall O'Leary, and Patrick O'Sullivan. 2014. Visualising the Emigrant Letter. *Revue européenne des migrations internationales* 30(vol. 30 - n°3 et 4):49–69. https://doi.org/10.4000/remi.7081

Emma Louise Moreton. 2016. *The Emigrant Letter Digitised: Markup and Analysis*. D_ph, University of Birmingham.

Niall O'Leary and Emma Moreton. 2017. The Migrant Letter Digitised: Visualising Metadata. *Journal of Cultural Analytics* https://doi.org/10.22148/16.013

Jean-Christophe Pellat. 2015. Les graphies des Poilus, loin des canons orthographiques. In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 67–77.

Peter Stadler, Marcel Illetschko, and Sabine Seifert. 2016. Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>. Journal of the Text Encoding Initiative (Issue 9). https://doi.org/10.4000/jtei.1433

Agnès Steuckardt, editor. 2015a. *Entre village et tranchées: l'écriture de poilus ordinaires*. Inclinaison, Uzès.

Agnès Steuckardt. 2015b. Sans point ni virgule. In Agnès Steuckardt, editor, *Entre village et tranchées: l'écriture de poilus ordinaires*, Inclinaison, Uzès, pages 91–100.

Agnès Steuckardt and Beatrice Dal Bo. 2018. Avoir été ou être allé ? Évolution d'une concurrence, d'après des corpus lettrés et peu lettré. In Peter Blumenthal and Denis Vigier, editors, *Études Diachroniques Du Français et Perspectives Sociétales*, Peter Lang, Berlin, page 295.

The use of parallel Corpora for a contrastive (Russian-Italian) description of discourse markers: new instruments compared to traditional lexicography[1]

Anna Bonola
Università Cattolica del Sacro Cuore
Milan
anna.bonola@unicatt.it

Valentina Noseda
Università Cattolica del Sacro Cuore
Milan
valentina.noseda@unicatt.it

Abstract

English. This paper relates to Corpus Linguistics and in particular to parallel corpus linguistics (Borin, 2002), which promotes the use of parallel corpora for studying languages. After briefly presenting the features of an Italian-Russian parallel corpus designed and compiled by the authors of this paper, and after having clarified the reasons why parallel corpora are such a valid aid, compared to traditional lexicography, especially to investigate linguistic structures characterized by a high pragmatic and language-specific content, such as discourse markers, we propose to test the efficacy of the Italian-Russian parallel corpus by presenting two case studies: the Italian discourse marker *allora* and the Russian particle *ved'*.

Italiano. Questo lavoro si inserisce nel filone della *Corpus Linguistics* e in particolare in quello della *parallel corpus linguistics* (Borin, 2002), che promuove l'uso dei corpora paralleli nello studio delle lingue. Dopo aver presentato in breve le caratteristiche di un corpus parallelo italiano-russo progettato e compilato dagli autori del presente contributo, e dopo aver chiarito le ragioni per cui i corpora paralleli, in confronto alla lessicografia tradizionale, sono un ausilio molto più valido per indagare strutture linguistiche ad alto contenuto pragmatico e linguospecifiche, come i segnali discorsivi, ci proponiamo di attestare la validità del corpus parallelo italiano-russo presentando due *case study*: il primo in direzione italiano-russo (il segnale discorsivo *allora*), e il secondo in direzione russo-italiano (la particella *ved'*).

1 Parallel corpora[2] and linguistic research

Despite the skepticism of early corpus linguists, who refused to use translated texts to draw conclusions about the functioning of a language[3], nowadays the scientific community has produced countless works that demonstrate how the use of parallel corpora (PC) can have a greater impact in several areas[4]:
1) in linguistic research (contrastive, but not only) PC provide a rather solid empirical basis for comparing two or more languages (Johansson, 2003); moreover, the 'translation method' allows to deepen the semantics and functions of a given linguistic structure (Noël, 2003)[5];
2) in *Translation Studies*, since Baker's work (1993), PC have become a fundamental tool for the study of translated texts, treated as a linguistic variety in its own right, worthy of analysis;
3) finally, PC have allowed computational linguistics to make progress in the programming of translation software and, more generally, they have favored the development of NLP (Calzolari and Lenci, 2004).

However, these 3 points must be integrated with a further aspect: PC are in fact very useful for the heuristic phase of a contrastive analysis on polyfunctional linguistic elements that are strongly influenced by the context.

1 This paper is the result of a research in which the authors have equally contributed; however, Valentina Noseda is the author of sections 1, 1.1, 3.1 and 4, Anna Bonola of sections 2, 3 and 3.2.

2 A parallel corpus consists of texts in a language A, aligned (usually at the sentence level) with the corresponding translations in a language B. If bidirectional, the corpus will also contain language B originals alongside translations in language A.

3 The reasons for this skepticism can be traced to the generally recognized existence of the so-called universals of translations, to the influence that the source text often exerts on translators and their final product, and to the freedom with which a translator can interpret the source text while transferring its contents into a target text (Olohan, 2004; von Waldenfels, 2012; Zanettin, 2012).

4 Another field where parallel corpora have proved to be useful is the teaching of second languages, since bilingual corpora, first of all, allow students to grasp equivalences and differences between L1 and L2, thus acquiring greater awareness of the structures of a studied language (Granger, 2003), and secondly help them learn unknown words (Bernardini, 2004).

5 Noël (2003) was among the first to promote the use of PC not only for contrastive analysis but also to deepen the semantic investigation of one of the two aligned languages. In Russian studies, many works have been carried out following Noël's example, including (Zaliznjak, 2015; Levontina and Denissova, 2017; Zaliznjak, Denissova and Mikeljan, 2018).

In this paper we will show an example of such further use of PC, applying it to the contrastive study of two DMs.

1.1 The Italian-Russian parallel corpus of the Russian National Corpus

In Russian studies, the active use of language corpora fell slightly behind the spread of Corpus Linguistics around the world and it has been directly linked to the creation of the Russian National Corpus (*Nacional'nyj korpus russkogo jazyka*, from now on: NKRJa) in 2004. With its 500 million words, its numerous specialized sub-corpora and a highly sophisticated search engine, NKRJa has quickly become an essential tool for the study of Russian[6].

In 2005 NKRJa already presented a section dedicated to PC, although for the Russian-Italian pair there was only a small pilot corpus, not very balanced and almost useless for any type of research. A first expanded version – resulting from the collaboration between Catholic University of Milan (Università Cattolica del Sacro Cuore di Milano), the University of Bologna (Università di Bologna) and the Russian Language Institute in Moscow (*Institut Russkogo Jazyka imeni V.V. Vinogradova*) – became available in 2015. Now the Italian-Russian PC (it-ru PC) exceeds 4 million words and has become a sufficiently large tool allowing to conduct scientifically valid and statistically relevant research.

The corpus, compiled according to precise criteria, has the following features[7]: i) is bidirectional: it contains Russian originals translated into Italian and vice versa; ii) it includes several literary works and essays (from 19th, 20th and 21st centuries) as well as some newspaper articles written in the last decade (and this variety distinguishes it from other parallel corpora in NKRJa); iii) like all the other sections of NKRJa, it has three types of annotation: metatextual[8], morphological and semantic.

2. The use of parallel corpora for the analysis of discourse markers

A field in which parallel corpus linguistics seems to have great potential, especially if compared to more traditional research methods, is that of discourse markers (DMs), i.e. multi-functional linguistic elements of various origins (adverbs, verbs, particles, etc.) that can operate at a textual, discursive, interactive, modal, social and contextual level[9]. DMs have come to the attention of researchers especially during the eighties, as a result of a new pragmatic direction in language studies, and since then considerable progress has been made in this area[10]. However, the use of electronic corpora in the description of DMs is still in its initial phase.

The difficulty of producing a fully automatic tool for the analysis of DM is due to the fact that these are procedural and multifunctional elements expressing pragmatic and discursive functions which are clarified only in relation to the context or to the communicative situation, whose automatic annotation is still developing[11]. Moreover, syntactically, DMs are optional (can be removed), relatively mobile in the utterance and come from diverse grammatical classes, on which depends their syntactical integration (Crible, 1917: 106).

Therefore, the discussion on the automatic processing of DMs is currently still focused on the "need for functional paradigmatic studies that include every kind of DMs, possibly in multifunctional approaches for better generalization" in order to "provide a solid basis for comparative or contrastive analysis between languages and frameworks" (Crible, 2017: 100).

Some recent experiments for the identification and annotation of DMs are worth noting, like for example (Bolly et al., 2017), even though the empirical method they present is still matching manual and automatic annotation. For a fully automatic cross-linguistic analysis of DMs, which takes into account not only syntaxis

---

[6] A detailed description of the corpus and its sub-corpora (including all the information about the available annotations) can be found on the corpus website (www.ruscorpora.ru). See also (Aa.Vv 2005) and (Plungjan 2009).

[7] For a description of the Russian-Italian PC and its design criteria, see (Noseda, 2018).

[8] It provides various pieces of information about a text: author, date, genre, number of words, etc.

[9] This list of functional areas summarizes the results of the debate on the classification of DMs – for a review see (Schiffrin 2001; Frediani and Sansò 2017); for the discussion in Italy see (Bazzanella 2001: 41-42) – although we avoid entering into the discussion on labels, whose boundaries are subject to change and have a graduated character (Molinelli, 2018: 277).

[10] For a review of the features of the DMs highlighted by the research, up to the most recent studies see (Frediani and Sansò, 2017).

[11] As far as Russian is concerned, in the NKRJa only the multimodal sub-corpus (4 million words) is pragmatically annotated: the search engine can be interrogated on the basis of specific contexts (at the doctor's, at the restaurant, etc..) and linguistic acts (complaint, prohibition, apology, etc..).
Among Italian corpora, we can name the AVIP corpus (http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/673-corpus-avip-api) and PraTID (http://www.parlaritaliano.it/index.php/it/progetti/35-pratid-un-sistema-di-annotazione-pragmatica-di-dialoghi-task-oriented), which are fully or partially annotated at a pragmatic level.

(Cinque 1999) but also semantics and pragmatics, the annotation of PC, according to Crible (2017: 107), should consider the following levels: ideational (the relation between real-world events), rhetorical (the relation between epistemic and speech-act events), sequential (the shaping of discourse segments) and interpersonal (speaker-hearer relationship). Therefore, the large amount of data that can be consulted today through electronic corpora, as far as DMs are concerned, has yet to find a way to be processed employing a targeted annotation. More specifically, concerning the automatic analysis of DMs in Russian, some supracorpora databases (SCDB), resulting from the processing of some bilingual parallel corpora within NKRJa, have recently been developed (Zatsman, Inkova, Kruzhkov and Popkova, 2016). Their aim is to increase the functionality of parallel corpora for goal-oriented cross-linguistic research on various linguistic elements. For the moment, there is one SCDB for French-Russian contrastive analysis of verbs (Zatsman and Buntman, 2015) and one for textual connectors (Inkova, 2018).

The it-ru PC used for our research does not have an annotation that takes into account pragmatic and discursive parameters; moreover, we still do not have Russian-Italian SCDB for such particular linguistics elements as DMs. Therefore, for the moment, we tested the effectiveness of it-ru PC as a tool for linguistic analysis in the heuristic phase, as it provides a significant number of examples in a short time, allowing researchers to clarify and adjust their intuition regarding a given research question (corpus-based approach) or to formulate new hypotheses (corpus-driven approach) (Mikhailov and Cooper, 2016: 15-16). If this is generally helpful, it is even more useful for DMs, i.e. linguistic elements that both in Italian and Russian have been developing textual, discursive, modal and pragmatic functions that make them multifunctional and often language-specific, but frequently still lacking an adequate description (Proietti, 2000: 227) (Benigni and Nuzzo, 2019: 152–154)[12], especially if we consider current lexicography.

As we will show in this paper, the effectiveness of our PC (in its current form) for the heuristic phase of a contrastive corpus-driven or corpus-based approach – well described in (Crible, 2017) – lies in the fact that:
1) it makes the multi-functionality of DMs easily emerge, clarifying it by contrast with another language, as description through linguistic comparison, "rend le dispositif d'analyse plus puissant: elle peut suggérer, d'une part, de nouvelles hypothèses pour les faits constatés; elle peut, d'autre part, inciter à réexaminer des hypothèses existantes" (Lamiroy, 1984: 224);
2) if a given DM presents recurrent functional equivalences in the language compared, it is possible to determine if in the L2 there are DMs associated with specific functions as well;
3) finally, analyzing quantitative data (even with a relatively small number of examples), we can see the preferential strategies of each language to express certain functions, and in some cases, as illustrated in section 3, it is also possible to make some assumptions about possible structural differences between the two compared languages.

In section 3 we will exemplify the abovementioned points by analyzing Italian *allora* and Russian *ved'*, two of the most frequently used DMs in the respective languages.

3. The DMs *allora* and *ved'*

Concerning the pragmatic-textual multi-functionality of DMs (section 2, point 1), both Russian and Italian lexicographic descriptions are particularly poor and often do not distinguish contextual elements from the functional core meaning of the DM under investigation.

For example, as far as *allora* is concerned, DISC 2008, among the several dictionaries that we have consulted, is the only one providing some clear categories about the discursive use of this word, which can be a temporal adverb, a conjunction or an actual DM. According to DISC, *allora*, as DM, refers to shared knowledge in dialogues (*Allora?*) or in exhortative, imperative and interrogative sentences (*e allora sei pronto?*). This brief description, although correct, is rather uncomplete and it uses contextual categories, such as sentence or text type, without specifying how their role interacts with the functionality of the DM.

As for traditional Russian lexicography, the description of DMs is not better: in both traditional (Ušakov, 1935) and recent dictionaries (Kuznecov, 2000; Efremova, 2001; Ožegov and Švedova 2003) the particle-conjunction *ved'*, whose various meanings are summarized in (Morozov, 2014: 259), is defined as follows: 1) conjunction in those sentences that indicate the cause or the motivation of a previous statement; 2) concessive conjunction; 3) it expresses a hypothetical or possible state; 4) particle that underlines or contradicts what has

[12] In particular, in the article, dedicated to the use of corpora for teaching DMs, the authors underline how even in this field has emerged so far "a lack of contextualization of pragmatic phenomena and a shortage of natural conversational models, exemplifying the real use of language" (Benigni and Nuzzo, 2019: 154).

been said; 5) it emphasizes adversative conjunctions such as *no* [but][13], *a* [but, and], *daže* [even]; 6) in conditional clauses it means *togda* [then], *v takom slučae* [in this case]; 7) it indicates a statement from which a conclusion will be drawn; 8) it gives emotional color to spoken language; 9) in questions and exclamations it means *neuželi ne*?, *razve ne* [really/indeed].

Such a functional heterogeneity, as well as the variety of aspects involved, shows that the core meaning of *ved'* provided by lexicographic descriptions is quite vague and even confused. Moreover, as in the case of *allora*, the problem of distinguishing the function of connector from that of DM remains.
Thanks to our corpus-driven analysis in the it-ru PC, a much more precise and complex description has surfaced.

3.1 *Allora*

Our analysis took into account the first 200 occurrences of *allora* automatically extracted from the corpus (100 in Italian originals and 100 in texts translated from Russian)[14].

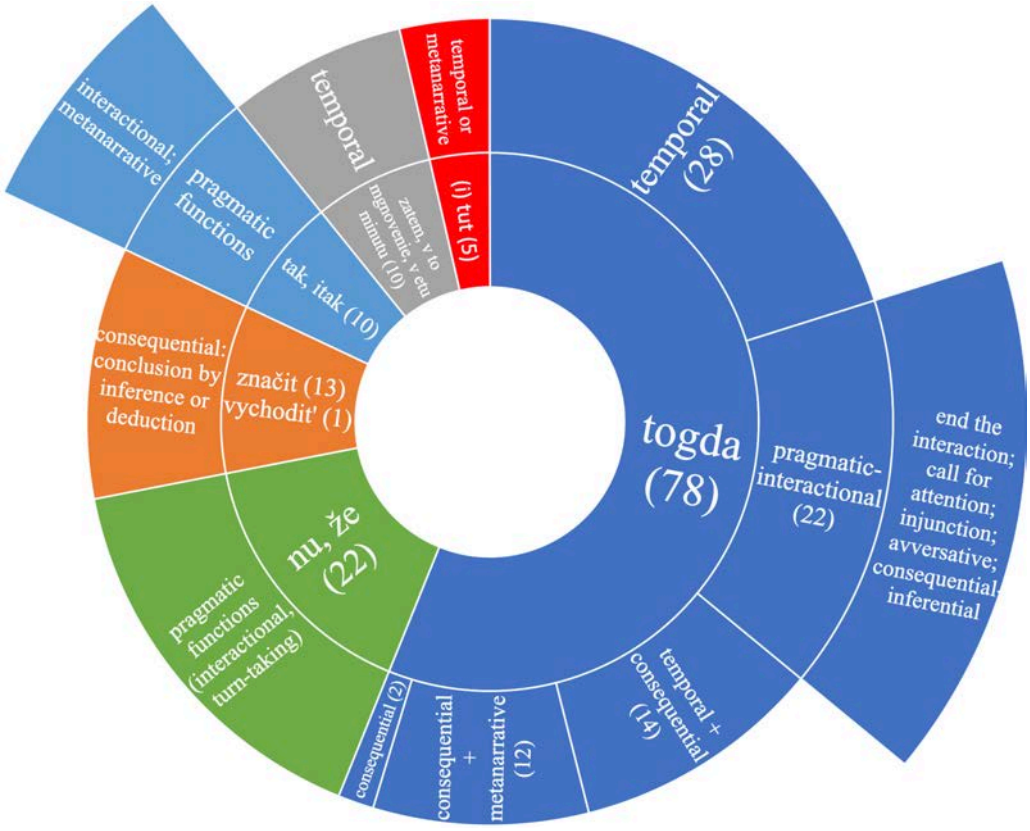
Firstly, we considered Russian DMs corresponding to *allora* both in Russian translations and in Russian original texts; secondly, we examined their different functions. Our goal was, on the one hand, to clarify the multi-functionality of *allora* by contrast with Russian, and on the other to compare our results with the descriptions of this DM provided by traditional lexicography and some linguistic research works. This allowed us to verify if our PC, even in its current form, can be useful to integrate these resources towards a more precise description.

As *Allora* is highly polysemic (it combines temporal, logical and pragmatic values) and multifunctional (it can be an adverb, a connector or a DM), we found out that it does not have full functional equivalents in Russian; in fact, quite frequently (25 occurrences) *allora* does not show any equivalent at all: either it is omitted in the Russian translation or it is inserted in the Italian translation without a corresponding DM in the Russian original; its adverbial and connective values are rather carried out by different and thus highly specialized markers with metatextual/metanarrative, interactive and pragmatic functions (this distinction is provided in Bazzanella, 2001) (see section 2, point 2). More precisely:
– *togda* [then] mostly conveys adverbial and connective meanings;
– *značit* and *vychodit* [so] (connectives) express conclusion by inference or deduction;
– *(i) tut* [and then] often expresses temporal correlation and is combined with the metanarrative function typical of *allora*, that marks the different phases of narration.
– *Tak/itak* [so] add two pragmatic functions to the basic consequential meaning: i) interactional function (beginning or end of the interaction, and turn-taking in a conversation); ii) metanarrative function (restarting the narration or marking the narrative phases);
– *Nu* and *že* [well] never have temporal meaning, but they carry out pragmatic functions, emphasizing the interactional process as well as turn-taking. *Nu* and *že* do not seem to express any consequential component.

These results are summarized in Figure 1, which shows, in addition, quantitative data. In this regard we must point out that we had to leave out some examples due to translation errors, omissions etc.; as a result, the number of examples that we could actually take into account amounts to 164 (including the 25 cases of zero correspondence which are not showed in Figure 1):

[13] The translations that we provide in brackets are approximate since even in English there is never a single equivalent for these words.
[14] This bidirectional approach allows determining if the behavior of a given linguistic unit differs according to text type (i.e. original versions vs translations). In this sense, a bidirectional parallel corpus has proved to be an extremely helpful resource.

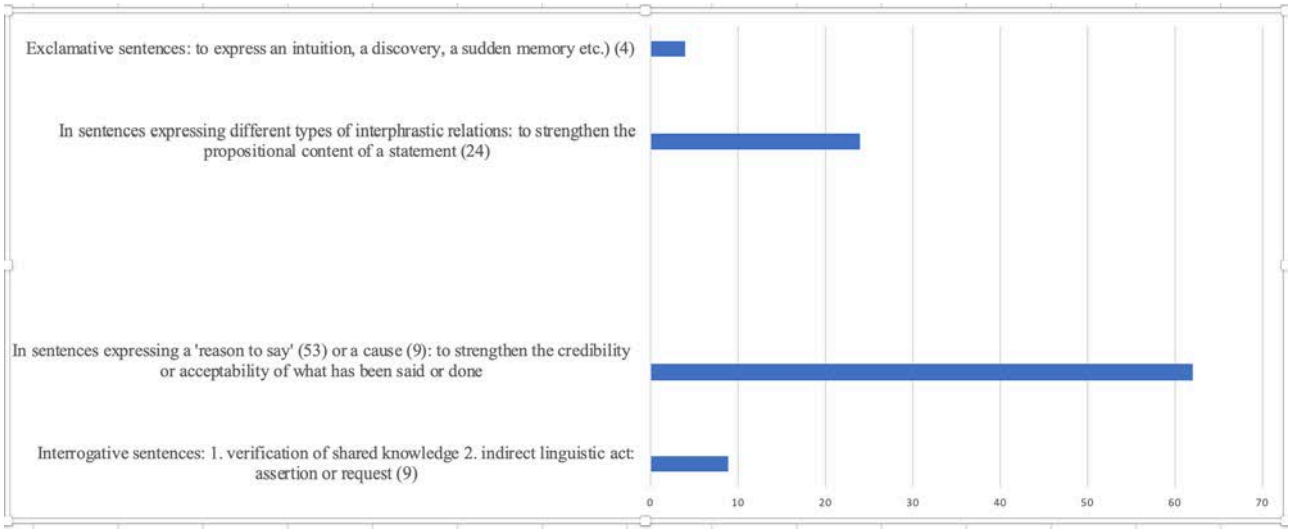
Figure 1: Data resulting from a corpus-driven analysis of DM *allora*



3.2 *Ved'*

Biagini and Bonola (2019, in press) have recently applied to *ved'* a similar heuristic method of investigation using the it-ru PC. They examined the first 100 occurrences automatically extracted from the corpus (both in originals and translated texts). In this case, the analysis was carried out considering first of all the contexts of occurrence. The primary goal was to identify the core meaning of *ved'*, in order to distinguish it from other peripheral values. The results of the analysis are summarized in Figure 2, followed by a brief explanation:

Figure 2: Data resulting from a corpus-driven analysis of DM *ved'*



- unlike what is stated in dictionaries, Biagini and Bonola would not attribute to *ved'* the encoding function of clause linking, even though in our corpus the group of contexts which exhibit an interphrastic relation is the second in terms of entity: *ved'* in fact occurs in sentences that express very different kinds of relations (such as adversative and causal), which, nevertheless, in almost all the examples are codified by conjunctions or are inferable from the propositional content of the statements, instead of directly depending on *ved'*.
- secondly, in more than 50% of the analyzed contexts, *ved'* occurs in the presence of two sentences, the second of which expresses a 'reason to say' (i.e. a reason why something was previously said) instead of a mere causal relation. Strengthening the illocutionary force of the second sentence by referring to a shared background that the speaker wants to recall, *ved'* provides the listener with useful hints to overcome the inferential process. In these contexts, *ved'* realizes the macro-functions of expressing textual cohesion (discourse marker), social cohesion and personal attitude (pragmatic marker).
- thirdly, the semantic core of *ved'* (if used when referring to a shared knowledge) persists in particular when it functions as a pragmatic marker that manages social cohesion and modulates illocutionary force (in questions and 'reasons to say') or as an element which favors the inferential process (in 'reasons to say'). When, on the other hand, it carries out the role of discourse marker favoring textual cohesion, it still refers to shared knowledge, but apart from this, nothing else is presupposed.

If the results described above exemplify points 1 and 2 of section 2, concerning the multi-functionality of *allora* and *ved'* or the specialization of their equivalents in Russian (for *allora*) and in Italian (for *ved'*), for point 3 – i.e. the preferential strategies of Russian and Italian regarding the expression of certain discursive functions – it was very useful to analyze the asymmetries emerged from the it-ru PC, i.e. the cases of omission or addition of *allora* and *ved'* in target texts compared to the originals. This analysis showed that both DMs are sometimes omitted in translation or they are added in the absence of a correspondent marker in the original. In addition to this, neither of the two DMs has a perfect functional equivalent in the target language, but they distribute their many functions on partial equivalents. This demonstrates a certain language-specificity of both DMs (on the relationship between the number of translation variants and language-specificity of DMs see Inkova, 2017).

Moreover, as far as *allora* is concerned, we observed that using this marker we tend to give a logical (consequential) interpretation to the temporal relationship between two circumstances: "in that moment/that circumstance" can, in fact, be interpreted through *allora* also as a consequential relationship. Here we can see the preference of Italian for logical cohesion in the text. Russian, on the contrary, often simplifies this temporal-consequential relation in a strictly temporal sense, translating *allora* with temporal adverbs or adverbial phrases (on this difference between Russian and Italian, a consequence of Latin syntax, see Govorucho, 2007).

4 Conclusions: a hypothesis on the differences between Italian and Russian regarding the use of DMs

Our conclusions regard, firstly, the evaluation of the tool we adopted for our corpus-based contrastive analysis of DMs, i.e. the Russian-Italian bidirectional parallel Corpus of NKRJa. At the moment we can say that this corpus is suitable for the heuristic phase, but it does not yet provide sufficient data to draw general conclusions from a systemic or typological point of view. Any assumption about possible structural differences related to the use of DMs in Russian and Italian should be supported by a larger number of data. Nevertheless, a heuristic analysis allowed us to formulate some preliminary hypotheses.

More precisely, we were able to register the tendency of Russian to express purely pragmatic functions, both cognitive and interactive[15], through an ancient group of primitive particles, such as *ved'*, *nu*, *že*, which are more specialized if compared to DMs of more recent origin, such as *togda*, which maintains an adverbial and connective function as well. On the contrary, Italian tends to form multifunctional DMs of verbal or adverbial origin which combine their pragmatic features with the task of guaranteeing logical cohesion in the text and interphrastic relations.

This is a broad – and according to us new – observation on a structural difference between the two languages, which deserves to be further explored by investigating – both from a diachronic and a synchronic point of view – more Russian and Italian DMs. All this demonstrates how a heuristic corpus-driven study allows, on the one hand, to quickly obtain linguistic descriptions on the functioning of DMs that are more precise than those provided by traditional tools and, on the other, to open up new hypotheses for wide-ranging research.

[15] On this distinction see (Bazzanella, 2001).

References

Aa.Vv. 2005. *Nacional'nyj korpus russkogo jazyka: 2003–2005. Rezul'taty i perspektivy*. Indrik, Moskva.

Mona Baker. 1993. *Corpus linguistics and translation studies – Implications and applications*. Mons Baker, Gill Francis and Elena Tognini-Bonelli (eds). *Text and technology. In honour of John Sinclair*. John Benjamins Publishing, Philadelphia and Amsterdam, pp 233-250.

Carla Bazzanella. 2001. *Segnali discorsivi e contesto.* Wilma Heinrich and Christine Heiss (eds). *Modalità e Substandard.* CLUEB Bologna, pp. 41–64.

Carla Bazzanella, Cristina Bosco, Barbara Gili Fivela, Johanna Miecznikowski and Francesca Tini Brunozzi. 2008. *Segnali discorsivi e tipi di interazione*. Cristina Bosisio, Bona Cambiagli, Emanuela Piemontese and Francesca Santulli (eds). *Aspetti linguistici della comunicazione pubblica e istituzionale*. Guerra, Perugia, pp. 239–265.

Valentina Benigni and Elena Nuzzo. 2018. *L'insegnamento dei segnali funzionali in russo come lingua seconda*. Alberto Manco (ed.) *Le lingue extra-europee e l'italiano: aspetti didattico-acquisizionali e sociolinguistici. Atti del LI Congresso Internazionale di Studi della Società di Linguistica Italiana (Napoli, 28-30 settembre 2017)*. Officinaventuno, Milano, pp. 151-165.

Silvia Bernardini. 2004. *Corpora in the classroom. An overview and some reflections on future developments*. John Sinclair (ed). *How to use corpora in language teaching*. Benjamins, Amsterdam.

Francesca Biagini and Anna Bonola. 2019 (in press). *Descrizione semantico-funzionale delle particelle russe e corpora paralleli. Un'analisi contrastiva (italiano-russo) corpus-based di* ved'. Iliyana Krapova, Svetlana Nistratova and Luisa Ruvoletto (eds). *Studi italiani di linguistica slava: nuove prospettive e metodologie di ricerca*. Edizioni Ca' Foscari, Venezia.

Catherine T. Bolly, Ludivine Crible, Liesbeth Degand and Deniz Uygur-Diestexhe. 2017. *Towards a model for discourse marker annotation. From potential to feature-based discourse markers*. Chiara Frediani and Andrea Sansò (eds). *Pragmatic Markers, Discourse Markers and Modal Particles. New perspectives*. John Benjamins Publishing Company, Amsterdam and Philadelphia, pp. 71-98.

Lars Borin. 2002. *... And never the twain shall meet?* Lars Borin (ed.) *Parallel Corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University*, Sweden, 22-23 April 1999. Rodopi, Amsterdam and New York, pp. 1-43.

Nicoletta Calzolari and Alessandro Lenci. 2004. Linguistica computazionale. Strumenti e risorse per il trattamento automatico della lingua. *Mondo Digitale* 2: 56-69.

Guglielmo Cinque. 1999. *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press on Demand.

Ludivine Crible. 2017. *Towards an operational category of discourse markers: A definition and its model*. Chiara Frediani and Andrea Sansò (eds). *Pragmatic Markers, Discourse Markers and Modal Particles. New perspectives*. John Benjamins Publishing Company, Amsterdam and Philadelphia, pp. 99-124.

DISC. 1997–2008. *Dizionario italiano Sabatini Coletti*. Giunti, Firenze http://dizionari.corriere.it/dizionario_italiano

Tat'jana F. Efremova, (ed). 2001. *Tolkovyj slovar' služebnych častej russkogo jazyka*. Russkij jazyk.

Chiara Frediani and Andrea Sansò (eds). *Pragmatic Markers, Discourse Markers and Modal Particles. New perspectives*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Roman Govorucho. 2007. Složnoe predloženie s vremennym značeniem v ital'janskom i russkom jazykach i problemy rečevogo uzusa. *L'analisi linguistica e letteraria* 15(1): 93–115.

Sylviane Granger. 2003. *The corpus approach: a common way forward for contrastive Linguistics and Translation Studies?* Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson (eds). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Rodopi, Amsterdam and New York, pp. 17–29.

Ol'ga Ju. In'kova. 2017. Principy opredelenija lingvospecifičnosti konnektorov. *Komp'juternaja lingvistika i intellektual'nye technologii* 16(2): 150–60.

Ol'ga Ju. In'kova. (ed.), 2018. *Semantika konnektorov: kontrastivnoe issledovanie*. Torus press, Moskva.

Sergej A. Kuznecov. (ed). 2000. *Bol'šoj tolkovyj slovar' russkogo jazyka*. Norint, Sankt Peterburg.

Stig Johansson. 2003. *Contrastive linguistics and corpora*. Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson (eds). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Rodopi, Amsterdam and New York, pp. 31-44.

Béatrice Lamiroy. 1984. *La valeur heuristique de la comparaison linguistique: un exemple cencernant le français, l'espagnol et l'italien.* Guillet Alain and Nunzio La Fauci (eds). *Lexique-grammaire des langues romanes. Actes du premier colloque Européen sur la grammaire et le lexique comparés des langues romanes, Palerme, 1981*. John Benjamins Publishing Company, Amsterdam and Phladelphia, pp. 223–230.

Irina B. Levontina and Galina V. Denissova. 2017. Ital'janskoe *magari* i ego russkie perevodnye ekvivalenty: raznye diskursivnye strategii. *Komp'juternaja lingvistika i intellektual'nye technologii* 16(2): 261–270.

Mikhail Mikhailov and Robert Cooper (eds). 2016. *Corpus Linguistics for Translation and Contrastive Studies. A guide for research*. Routledge, London and New York.

Evgenij A. Morozov. 2014. Diskursivnye slova *ved'* i *doch*: opyt semantičeskogo analiza (na materiale slovarej sovremennogo russkogo i nemeckogo jayzkov). *Problemy istorii, filologii, kul'tury* 3: 258–60.

Dirk Noël. 2003. Translations as evidence for semantics: an illustration. *Linguistics* 41(4): 757–785.

Valentina Noseda. 2018. La *corpus revolution* russa e il corpus parallelo italiano-russo: storia, criteri di compilazione e usi. *L'Analisi linguistica e letteraria* 24(2): 115–132.

Maeve Olohan. 2004. *Introducing corpora in translation studies*. Routledge, London and New York.

Sergej I. Ožegov and Natal'ja Ju Švedova (eds). 2003. *Tolkovyj slovar' russkogo jazyka*. ITI Technologii, Moskva.

Vladimir A. Plungjan (ed). 2009. *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy*. Nestor-Istorija, Sankt-Peterburg.

Domenico Proietti. 2000. *Comunque* dalla frase al testo. *Studi di grammatica italiana* 19: 175–231.

Deborah Schiffrin. 2001. *Discourse markers: language, Meaning, and Context*. Deborah Schiffrin et al. (eds), *Handbook of Discourse Analysis*. Blackwell Publishers, Oxford, pp. 54–73.

Dmitrij N. Ušakov (ed). 1935-40. *Tolkovyj sovar' russkogo jazyka: v 4 t.* Sovetskaja enciklopedija, Moskva.

Ruprecht von Waldenfels. 2012. Polish tea is Czech coffee: advantages and pitfalls in using a parallel corpus in linguistic research. *Trends in Linguistics* 247: 263–28.

Anna A. Zaliznjak. 2015. Lingvospecifičnye edinicy russkogo jazyka v svete kontrastivnogo korpusnogo analyza. *Komp'juternaja lingvistika i intellektual'nye technologii* 13: 683–695.

Anna A. Zaliznjak, Galina V. Denissova and Irina L. Mikeljan. 2018. Russkoe *kak-nibud'* po dannym parallel'nych korpusov. *Komp'juternaja lingvistika i intellektual'nye technologii* 17: 803–817.

Federico Zanettin. 2012. *Translation-driven corpora*. Routledge, New York.

Igor' M. Zatsman and Nadezhda Buntman. 2015. Outlining goals for discovering new knowledge and computerised tracing of emerging meanings discovery. Andrea Garlatti and Maurizio Massaro (eds), *16th European Conference on Knowledge Management Proceedings. ECKM 2015*. Academic Publishing International Ltd, Reading, pp. 851–860.

Igor' M. Zatsman, Ol'ga Ju. In'kova, Michail G. Kružkov and Natalija A. Popkova. 2016. Representation of cross-lingual knowledge about connectors in supracorpora databases. *Informatika i ee Primeneniya (Informatics and its Apllications)* 10(1): 106-118.

PhiloEditor®: simplified HTML markup for interpretative pathways over literary collections

Claudia Bonsi
University of Milano-Bicocca
claudia.bonsi@unimib.it

Angelo Di Iorio
University of Bologna
angelo.diiorio@unibo.it

Paola Italia
University of Bologna
paola.italia@unibo.it

Francesca Tomasi
University of Bologna
francesca.tomasi@unibo.it

Fabio Vitali
University of Bologna
fabio.vitali@unibo.it

Ersilia Russo
University of Florence
ersilia.russo@unifi.it

Abstract

English. In this paper we introduce PhiloEditor®, a software environment for the representation and coloring of time-based modifications of literary texts meant for scholarly editions. The purpose of the tool has evolved from a simple philological characterization of temporal evolution of literary texts to a critical and hermeneutic approach of interpretative pathways. From a technical point of view, PhiloEditor® is based on an approach to markup that heavily diverges from the traditional XML/TEI and towards a reliance on a custom subset of HTML5, based on a mature theory of XML design patterns, well-behaved and well-formed, readily convertible into XML/TEI or whatever other XML vocabulary needed, and at the same time fully compliant with modern web frameworks and browsers.

Italiano. Nel presente contributo introduciamo PhiloEditor®, un ambiente digitale per la rappresentazione e la marcatura delle modifiche temporali di testi letterari per l'allestimento di edizioni critiche digitali. L'obiettivo della piattaforma è evoluto dalla semplice caratterizzazione filologica dell'evoluzione temporale di testi letterari ad un approccio critico ed ermeneutico di percorsi interpretativi. Da un punto di vista tecnico, PhiloEditor® si basa su un approccio di markup che diverge fortemente dal tradizionale XML/TEI affidandosi ad un sottoinsieme personalizzato di HTML5, basato su una teoria matura di modelli di progettazione XML, ben educati e ben formati, facilmente convertibile in XML/TEI o in qualunque altro vocabolario XML, e allo stesso tempo pienamente conforme ai moderni framework e browser.

1 Introduction

Descriptive markup has been introduced three decades ago as the main mechanism to provide structured annotations over arbitrarily organized text documents. The rigidity of SGML and XML-based languages and the complexity of their editing tools made HTML an attractive markup language for text representations, as well as web pages, academic articles, structured documents and literary collections. However, the syntax of HTML is too flexible and its vocabulary too rich and unspecific. So the idea of channeling and restricting this richness into smaller and simplified subsets of the language was tried: Scholarly HTML, RASH and ADF, for instance, are simple and rigid subsets of HTML5 provided with similar approaches towards a restricted vocabulary and syntax. PhiloEditor®[1] started as a web-based tool that identifies variants and versions of literary texts using such descriptive richness as a mechanism for creating and perusing complex, multiform and coexisting interpretative pathways over literary collections. It has been used for three years now at the Departments of Italian Studies of the University of Rome and at the University of Bologna, in order to study and characterize the differences between earlier and later versions of the same literary texts. PhiloEditor® was originally meant to provide a visualization of the differences between two versions of the same text, as a display of the output of a diff tool over two text documents. To ease the development of this tool, a simplified form of HTML5 immediately displayable in a browser window was used. Over time, in addition to the mere display of differences, a few additional interpretative tools were added to the interface to provide for highlighting of some semantically relevant textual features, rendered with different colors of the text. Finally, a few

[1] A demo is available at http://site1705.web.cs.unibo.it/phed6.2/#.

extractors of data were created to collect and graphically represent some statistics of features in the annotations and in the very text, exploiting the regularities in the controlled use of simplified HTML markup. The purpose of the tool, therefore, has shifted from the merely philological characterization of literary texts (i.e., to represent their temporal evolution) to a critical and hermeneutic approach to them, by providing support for custom, discipline-specific pathways over complex, multidimensional and temporally complex collections of (literary) documents. At the same time, the minimality of the markup language of the tool (relying exclusively on a well-behaved and extremely simplified subset of HTML) allows an extreme homogeneity and control over markup without the necessity of giving up in richness of features. Texts annotated through this simple format are extremely regular and can be exported to a well-known literary XML vocabulary for literary texts such as XML/TEI.

2 Markup patterns in simplified HTML

The pattern theory of documents (Di Iorio *et al.*, 2014) has been created to provide a generative approach to useful patterns in document structures. The advantages in restricting elements to patterns are to achieve *orthogonality*, as each pattern has a specific goal, and fits a specific context with no overlap with other patterns, and *assemblability*, as it is clearly associated to precise nesting rules. By limiting the possible choices, patterns prevent the spread of arbitrary structures and allow authors to create unambiguous, manageable and well-structured markup languages and, consequently, documents which drive reusability and homogeneity (Di Iorio *et al.*, 2012; Di Iorio *et al.*, 2013).

In our theory of patterns applied to PhiloEditor®, elements are divided into four categories according to two axes: text/no-text content and element/no element content. This creates a basic scaffolding of four categories: *marker* is the class of elements that can contain neither text nor other elements (i.e., empty elements), *flats* are the elements that can only contain text, *buckets* are the elements that can only contain other elements, and *mixed* is the class of elements that can contain both text and other elements.

	Cannot contain text	**Can contain text**
Cannot contain elements	Empty element (*Marker*)	Plain text element (*Flat*)
Can contain elements	Plain structural element (*Bucket*)	Both elements and text (*Mixed*)

Figure 1. Main categories of our pattern theory of documents.

We now switch to consider where these elements can be used, that is the context model that can accommodate them. Since only two categories can contain other elements, we have exactly eight patterns of elements.

	Marker	**Flat**	**Bucket**	**Mixed**
Marker	-	-	-	-
Flat	-	-	-	-
Bucket	Meta	Field	Container	Block
Mixed	Milestone	Atom	Popup	Inline

Figure 2. The basic set of patterns.

According to this model:

- A *meta* element is a marker (i.e., an empty element) placed within a bucket, i.e., never close to text fragments. Usually their real position is not relevant, and only its existence has some relevance: it is the

perfect candidate for metadata structures, hence its name. Element <meta> in HTML is clearly and rightly a meta element.

- A *milestone* is a marker (i.e., an empty element) placed within a mixed, i.e., close to the text fragments. Its location in the document is often its most important contribution, i.e. it represents a special position within of the document, hence its name. Elements
 and <img> in HTML are milestones.
- A *field* is a flat element (text allowed, elements not allowed) placed within a bucket, i.e., never close to text fragments. It is just a container of data whose position is not particularly relevant. Its proper use is as a field of a record, hence the name. Element <title> in HTML is a field.
- An *atom* is a flat element (text allowed, elements not allowed) placed within a mixed, i.e., close to text fragments. Its location in the document is often its most important contribution. Since it allows no element content, it is an atomic container of text, hence its name.
- A *container* is a bucket (text not allowed, elements allowed) placed within a bucket. It is a container of elements and it provides the fundamental scaffolding of the document structure. Containers are arguably the most important pattern of the eighth. For instance, elements <html> and <table> in HTML are containers, but there are many others.
- A *popup* is a bucket (text not allowed, elements allowed) placed within a mixed element. It provides an interruption of the usual flow of inline and text elements within a mixed element, creating a separate context for buckets such as containers. It acts as a frontier element between mixed and buckets, and lets the contained elements jump out of the constraints of mixed elements, hence its name. Element <figure> in HTML is a popup.
- A *block* is a mixed element (text and elements are both allowed) placed within a bucket. It is the earliest element in a hierarchy of containment of buckets that can contain text, and therefore acts as a frontier element between buckets and mixed. The most important type of blocks is the paragraph, i.e. a basic, independent container of text and inline elements of a document, a block of text vertically separated from the others of the same type, hence its name. Element <p> in HTML is therefore a block.
- An *inline* is a mixed (text and elements are both allowed) placed within a mixed. An inline element characterizes text fragments according to several criteria, such as style, typography, semantics, etc. Since it does not break the paragraph structure, it stays on the same line as the text, hence its name. Elements <b> or <a> in HTML are inlines, as well as many others.

These eight patterns represent by construction the complete set of possible element types, and no others can exist without loosen the rules. However we have identified some specific regularities within containers, which are very important and complex elements. Rather than creating independent patterns, we identified sub-patterns of containers, inheriting all their characteristics and adding others: *record, table,* hierarchical container (*hcontainer*).

This theory shows that although HTML does not recognize nor use patterns in a systematic way, it is possible to restrict HTML by selecting a reasonable subset of elements expressive enough to capture the typical components of domain-specific documents while being also well-designed, easy to reuse and robust.

3 Simplified HTML for scholarly, technical and literary texts

The discussion to adopt HTML as the markup language of choice even for specialized domain-dependent documents started in the academic publishing domain: through the unification of data formats we can homogenize the process of drafting, submitting and publishing documents. HTML easily supports the embedding of semantic annotations to improve the sharing of communication results, thanks to already existing W3C standards such as RDFa (Sporny, 2015) or JSON-LD (Sporny *et al.*, 2014), so that discoverability, interactivity, openness and usability of the scientific works are increased (Shotton *et al.*, 2009). The HTML language has been widely used as a full-fledged markup language to encode texts, not only plain Web pages but also academic articles, technical documentation, literary artefacts and data reports. The ability to embed semantics within HTML pages, for instance by using RDFa (Herman *et al.*, 2015), is a key factor for this success since it allows designers to express rich

information about any domain and to overcome the limitations of a language developed and used for long time for other purposes. HTML has recently gained importance as a markup language for writing technical documentation as well. ADF (Caponi *et al.*, 2018) is a pattern-based subset of HTML that allows designers to express information about the authoring process such as change-tracking data, templates and reusable fragments and authorship attribution data. The format can be manipulated by a Web editor that relies on the pattern-based structure of the language and its ability to express fine-grained semantic information on each piece of content.

4 PhiloEditor® for the scholarly markup of literary documents

PhiloEditor®, as a web-based environment for reading and annotating variants, is aimed to be a valuable support for scholars in the criticism of variants. Its main characteristics are the display of differences (*deltas*) between two versions of the same document and the classification of semantically relevant fragments of the text (*colouring*), according to a parametric set of features provided by a scholar in a specified domain. PhiloEditor®'s data model is based on a highly simplified version of HTML5 that is fully based on the pattern model described in section 2. PhiloEditor®'s features make it particularly suitable for the study of literary texts in multiple versions, in particular those resulting from the shifts of volition of the author (authorial philology). In fact, it has been tested first on the two printed editions of *I Promessi Sposi* by Alessandro Manzoni (Bonsi *et al.*, 2015), then on the internationally-known *Pinocchio* by Carlo Collodi (Gargano and Italia, 2018), comparing the version published periodically between 1881 and 1883 on "Giornale per i bambini" magazine with the printed edition published by Paggi in 1883.

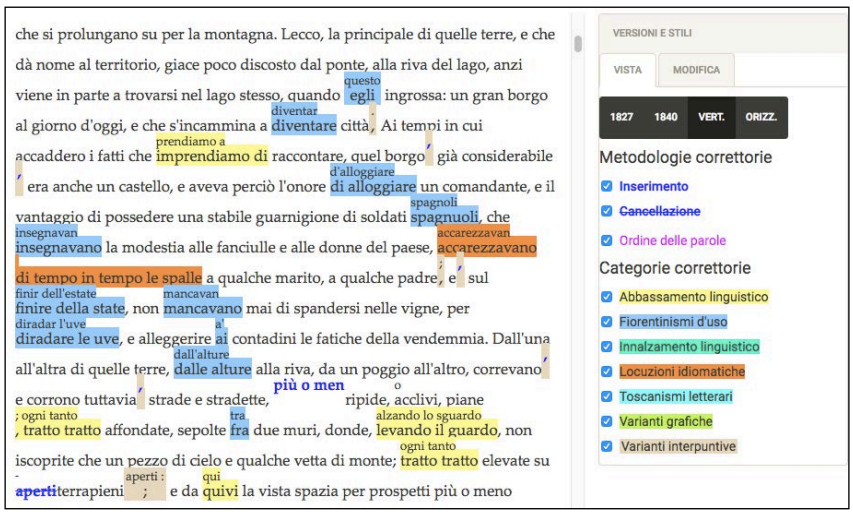


Figure 3. Color coding of the edits in *I Promessi Sposi*.

For versions and variants to be comparable, it's necessary for a delta to be created. Many algorithms exist that perform such tasks, but the kind of delta is also relevant for our purpose. Most algorithms meant for computer code assume that the relevant atomic entity to be differentiated is the individual line, since most programming languages use lines as their atomic semantic unit of production. In other types of documents, atomic entities may be single characters or structural nodes (e.g., a paragraph or a whole chapter). These are not particularly appropriate for literary texts, where long paragraphs with many differences (as in the case of our documents) would make the understanding of the variants difficult. The best choice for literary texts in our opinion is word-based diffing, which identifies whole words as the basic perceivable and understandable difference between variants. We thoroughly tested two Javascript libraries for word-based diffing, Javascript diff algorithm (Resig, 2005) and wikEd diff (Cacycle, 2017). The final choice was the latter, which is more precise and better parameterized. This algorithm reads two text files and outputs a document comparing them by placing the differences as spans of an HTML document. The output needs to be further synthetized and qualified: in particular,

wikEd diff generates a simple list of insertions and deletions. Literary variants do in fact contain relatively few simple deletions and insertions, and mostly replacements (where an old fragment is substituted by a new fragment). A replacement is seen by the diff algorithm as a deletion of the old text followed, in the same position, by an insertion of the new text. Correctly identifying and characterizing replacements is important for a better description of the actual editing operation occurred and this is generated by appropriately grouping the list of insertions and deletions provided by the diff algorithm before the result is displayed to the user. PhiloEditor® provides different ways to display variants: the user can choose an *exclusive*, a *synoptic* or a *layered* representation of the text. In these views, all modifications are characterized as replacements. Complex diff issues can be handled manually: the segment of the variants automatically created can be modified thanks to a simple interface operation that makes possible to divide or combine groups of elements according to the user's needs. Word-based diffing doesn't work with macro-variants though, so for now PhiloEditor® can only manage micro-variants.

The increased familiarity with the temporal evolution of the text brought a growth in curiosity towards the types of modifications. For this reason, we devised a two-layer model of modification types (Italia, 2018): "categories of corrections" to identify the actual methodology of correction (i.e. shifts of textual passages, additions, deletions, inversions of words, etc.) and "linguistic categories" to identify the linguistic categories that drove the correction. In order to describe the text in these terms, a color-based coding schema of the fragments of text affected by each phenomenon was created. Since the two classes could overlap, two separate types of styles were created using the color of the text (categories of correction) and the color of the background (linguistic categories). Displaying these changes visually enables us to reach a more complex perception of linguistic features of literary texts and facilitates not only the representation of important syntactic structures but also their hermeneutic implications.

The variety of textual features that need to be expressed in the markup of documents edited inside PhiloEditor® is limited: one needs to support the hierarchical structure of the document (chapters, paragraphs, text), very little typography (just a few words in italic here and there), plus the application-specific requirements of describing replacements (e.g., the edits generated by the diff engine as insertions and deletions and converted by an internal algorithm) and the colors, i.e., the linguistic categories described by scholars.

```
che si prolungano su per la montagna. Lecco, la principale di quelle terre, e che
dà nome al territorio, giace poco discosto dal ponte, alla riva del lago, anzi viene
in parte a trovarsi nel lago stesso, quando
    <span class="replace florentinism" data-responsible="Teresa">
        <span class="new">questo</span>
        <span class="old">egli</span>
    </span>
 ingrossa: un gran borgo al giorno d'oggi, e che s'incammina a
    <span class="replace florentinism" data-responsible="Teresa">
        <span class="new">diventar</span>
        <span class="old">diventare</span>
    </span>
 città
    <span class="replace punctuation" data-responsible="Teresa">
        <span class="new">.</span>
        <span class="old">,</span>
    </span>
 Ai tempi in cui accaddero i fatti che
    <span class="replace lowering" data-responsible="Teresa">
        <span class="new">prendiamo a</span>
        <span class="old">imprendiamo di</span>
    </span>
 raccontare, quel borgo
    <span class="replace insert punctuation" data-responsible="Teresa">
        <span class="new">,</span>
        <span class="old"> </span>
    </span>
 già considerabile
    <span class="replace insert punctuation" data-responsible="Teresa">
        <span class="new">,</span>
        <span class="old"> </span>
    </span>
 era anche un castello, e aveva perciò l'onore
    <span class="replace florentinism" data-responsible="Teresa">
        <span class="new">d'alloggiare</span>
        <span class="old">di alloggiare</span>
    </span>
 un comandante, e il vantaggio di possedere una stabile guarnigione di soldati
```

Figure 4. The HTML source of *I Promessi Sposi*.

Correspondingly, we have created an extreme simplification of the HTML5 used by the application: <section>, <p> and <i> are used for the document, and is used for both modifications and color coding (since most of the colors are applied to modifications anyway). HTML classes are created and

used both to provide semantics and rendering characteristics, and an additional data-* attribute is used to provide basic provenance attribution. Given that the same span can be associated to multiple classes, clashes in categorization are impossible. On the other hand, the HTML that is being created is rigorously well-formed and homogeneous, and generating a correct and valid XML/TEI source is immediate and straightforward.

Using HTML class names for semantic characterization has both advantages and issues. Of course, classes make the association of special rendering to fragments easy. On the other hand, it is extremely difficult to impose constraints on the list of classes that can be used, e.g., to specify within a schema that elements of the class "replace" must first contain an element with the class attribute containing "new" and then an element with the class attribute containing "old".

5 PhiloEditor® and Scholarly Editions: markup tools for literary texts

Most of the available online editions of literary texts are based on full-text transcriptions of original texts into electronic form (Franzini, 2012), typically using the XML/TEI model, where the sources (*witnesses* in philology jargon) are traditional primary sources. The infrastructure is generally based on standard web programming languages, both client- and server-side. Users have limited access to the digital sources without any awareness of the backend software employed. Most of the editions, as said, are XML/TEI documents, written by hand with a stand-alone application such as Oxygen, and converted as needed into HTML/CSS documents using XSLT stylesheets. All projects in this domain are built on the same basic structural components: they consist of a set of files (*assets*) stored inside an information architecture such as a database or file system (*structure*) where they can be accessed (*services*) and shown on a browser (*use/display*) (Druker *at al.*, 2014). Different phases (assets, structure, services, use/display) mean different tools. Thus, the DiRT Directory is a registry of digital research tools for scholarly use, while the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) "breaks down the research lifecycle into high-level 'goals', each with a set of 'methods' ".

In the field of literary texts, editing is the first important step of the process: editors need to use simple applications in order to describe their documents and their features. TextGrid, for instance, is a real research environment with all the necessary tools and services to support the entire research process, especially in digital scholarly editing. The downloadable "TextGridLaboratory" is an editor for XML/TEI markup with a view on the source code and a traditional visualization mechanism. Catma (Computer Assisted Text Markup and Analysis) is an online environment that manages analysis, manual and automatic annotations, and visualizations of documents. Juxta is an open-source tool for comparing and collating multiple witnesses of a text work, useful for an analytic visualization of textual variants. Tapas is meant for visualizing, storing and sharing XML/TEI documents, while EVT (Edition Visualization Technology) is a downloadable open source tool that creates digital editions from XML-encoded texts. The final styling of documents is entrusted to CSS style-sheets and is easily customizable.

As shown, XML/TEI is the fundamental data model for all documents, and it is used as a work format rather than as an output format, requiring scholars to learn it in depth. Simplified HTML could be an alternative that, while maintaining complete exportability to XML/TEI, allows application designers to rely on web technologies much easier to work with, and avoiding exposing literary scholars to angle brackets altogether.

6 Conclusions

PhiloEditor® is but one of the activities in which we are pushing for the use of a simple HTML instead of a full-fledged custom XML vocabulary, although striving to identify a well-formed, well-behaved, totally descriptive and specialized subset of HTML. In PhiloEditor® two specific markup needs, the description of modifications and the coloring of semantic characterization of the texts, have been expressed within a standard and very simple subset of HTML. Similar activities, such as RASH or ADF, are used in other completely different, but still very specialized, domains. Overall the response to the features and the flexibility of PhiloEditor® has been overwhelmingly positive. The objective of the next

future is to slowly convert PhiloEditor® into a full-scale environment for all the needs of the collection, digitization and publication of scholarly editions of literary texts.

References

Bonsi C, Di Iorio A, Italia P, Vitali F. 2015. *Manzoni's electronic interpretations*, in The Mechanic Reader. Digital methods for literary criticism, special issue of "Semicerchio", LIII (2015/2), pp. 91-99.

Cacycle. 2017. wikEdDiff, https://en.wikipedia.org/wiki/User:Cacycle/wikEdDiff

Caponi A, Di Iorio A, Vitali F, Alberti P and Scatá M. 2018. *Exploiting patterns and templates for technical documentation.* In DocEng '18: ACM Symposium on Document Engineering 2018, August 28–31, 2018, Halifax, NS, Canada. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3209280.3209537

Di Iorio A, Italia P, Vitali F. 2015. *Variants and Versioning between Textual Bibliography and Computer Science*, in F. Tomasi, R. Rosselli Del Turco, and A.M. Tammaro (Eds.), Humanities and Their Methods in the Digital Ecosystem. Proceedings of Third AIUCD Annual Conference (AIUCD2014). Selected papers. ACM, New York.

Di Iorio A, Peroni S, Poggi F, Vitali F, Shotton D. 2013. *Recognising document components in XML-based academic articles*. In: Proceedings of the 2013 ACM symposium on document engineering. New York. ACM. 181-184.

Di Iorio A, Peroni S, Poggi F, Vitali F. 2012. *A first approach to the automatic recognition of structural patterns in XML documents*. In: Proceedings of the 2012 ACM symposium on document engineering. New York. ACM. 85-94.

Di Iorio A, Peroni S, Poggi F, Vitali F. 2014. *Dealing with structural patterns of XML documents*. Journal of the American Society for Information Science and Technology 65(9):1884-1900.

Dirt Directory, https://dirtdirectory.org/

Drucker J., Kim D., Salehian I., Bushong A. 2014. *Introduction to Digital Humanities. Concepts, Methods, and Tutorials for Students and Instructors*. http://dh101.humanities.ucla.edu/?page_id=15

Edition Visualization Technology (EVT), http://evt.labcd.unipi.it/

Franzini, G. 2012-, *Catalogue of Digital Editions*. DOI: 10.5281/zenodo.1161425, https://dig-ed-cat.acdh.oeaw.ac.at/

Gargano T, Italia P. 2018. *Philoeditor 3.0: Pinocchio*, in P. Ponti e M. Marazzi (a cura di), *«Senza giudizio... e senza cuore»,* Atti del convegno di studi su Pinocchio, (18-19 maggio 2017 – Università Cattolica del Sacro Cuore e Università degli Studi di Milano), numero monografico della "Rivista di letteratura italiana", a. XXXVI, n. 2 (2018), pp. 133-144.

Herman I, Adida B, Sporny M, Birbeck M. 2015. RDFa 1.1 Primer - Third Edition Rich Structured Data Markup for Web Documents, W3C Working Group Note 17 March 2015.

Italia P. 2018. *Filologia d'autore digitale e multidiscipinare. Dall'Authorship alla Fotonica*, in G. Sampino, F. Scaglione (a cura di), Saperi Umanistici nella Contemporaneità, Atti del Convegno Internazionale dei dottorandi (17-18 settembre 2015 – Università degli Studi di Palermo), La Biblioteca di Classico Contemporaneo 6, Palumbo Editore, Palermo 2018, pp. 322-334.

Juxta, http://www.juxtasoftware.org/

Meister J.C., Petris M., Gius E., Jacke J. CATMA 5.0 (2016) [software for text annotation and analysis]: http://www.catma.de

Resig J. 2005. Javascript Diff Algorithm, https://johnresig.com/projects/javascript-diff-algorithm/

Shotton D, Portwin K, Klyne G, Miles A. 2009. Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLOS Computational Biology 5(4):e1000361

Sporny M, Kellogg G, Lanthaler M. 2014. JSON-LD 1.0—a JSON-based serialization for linked data. W3C Recommendation 16 January 2014. World Wide Web Consortium. https://www.w3.org/TR/json-ld/

Sporny M. 2015. HTML+RDFa 1.1: support for RDFa in HTML4 and HTML5. W3C recommendation 17 March 2015. World Wide Web Consortium. http://www.w3.org/TR/rdfa-in-html/

TaDiRAH, https://dirtdirectory.org/tadirah

Tapas project, http://tapasproject.org/

TextGrid Consortium. 2006–2014. TextGrid: A Virtual Research Environment for the Humanities. Göttingen: TextGrid Consortium. textgrid.de

An Empirical Study of Versioning in Digital Scholarly Editions

Martina Bürgermeister
University of Graz
martina.buergermeister@uni-graz.at

Abstract

English. This paper contributes to increasing the reliability in Digital Scholarly Editions (DSE) by emphasizing the use of versioning. These mechanisms have the potential to render DSE more trustworthy. Within an analysis of existing DSE projects it will be shown, what kind of versioning strategies have been implemented and why. Versioning strategies are assessed according to a definition of versioning, which has three criteria: creation of versions for each change, documentation of changes, and availability of previous versions.

Italiano. Questo articolo contribuisce a sviluppare una migliore valutazione dell'attendibilità delle edizioni scientifiche digitali (DSE), enfatizzando l'uso delle versioni. Questi meccanismi hanno il potenziale per rendere la DSE più affidabile. Nell'ambito di un'indagine sui progetti DSE esistenti verrà mostrato che tipo di strategie di versioning sono state implementate e perché. Le strategie di versioning sono analizzate secondo una definizione di versioning, che ha tre criteri: creazione di versioni ad ogni modifica, documentazione delle modifiche, e disponibilità delle versioni precedenti.

Digital publishing and a textuality that is dynamic, collaborative, distributed and interdependent lead to the digital scholarly edition (DSE) facing additional technological challenges in contrast to print editions. The *MLA Committee von Scholarly Editions* takes care of this and recommends the use of technologies appropriate to the goals of the edition, "in recognition of the fact that technologies and methods are interrelated in that no technical decisions are innocent of methodological implications and vice versa (MLA 2016, 1)." The committee also suggests a design of the DSE that will be as durable and sustainable as possible. This applies not only to DSE, but also to the entire scientific practice. In the "Guidelines of good scientific practice", like the one from the University of Parma, it says: "I ricercatori dell'Ateneo devono aver cura che tutti i dati, primari e secondari, generati dalle loro attività di ricerca siano archiviati e conservati in modo corretto ed appropriato, garantendone la sicurezza e l'accessibilità […]" Besides being responsible for data safety and access, researchers should support a prompt data exchange and reuse: "Con l'intento di rendere la ricerca più aperta, globale e collaborativa e garantirne un controllo di qualità, i dati dovrebbero essere messi a disposizione dei colleghi che vogliano replicare lo studio o elaborare nuove ricerche a partire da essi […] (Università di Parma, 2018)".

The reuse of data is a matter of trust in data itself and the IT infrastructure as maintainer and provider. In this case, transparency is a key factor. In the following it will be argued, that versioning is more than a measure to guarantee data authenticity, integrity and accessibility. Versioning as an integrated part of the DSE creates transparency in an ongoing scholarly discourse. This objective is achieved, firstly, when new versions of the DSE or components of it are created if changes are made to it. Second, if what has been changed is also communicated; and third, if previous versions are made available. These three criteria form the basis of my analysis of versioning DSEs. Using an empirical approach, I would like to find out why versioning plays a role in these projects and how it is implemented.

The implementation of versioning in a DSE is not yet state-of-the-art. In order to find out which DSEs have a versioning strategy and what it looks like, the "Catalogue of digital editions" (ÖAW) was used as a finding aid. Since the list is very extensive, the analysis includes edition projects mainly from the past 15 years, which have developed over a relatively longer period of time. Since there are far more DSEs without versioning than with it, the empirical basis was mainly provided by Bleier's preliminary work (2019). He has evaluated 100 DSEs on criteria such as citation recommendation, permalinks and versioning. Finally, those cases that assign version numbers, but neither convey what has changed, nor make past versions available, were excluded from further consideration. Their exclusion from this analysis is justified by the fact that they fulfil only one of three conditions for versioning DSEs discussed here. The review has shown that so far three different versioning strategies have been applied, which meet at least two of the three conditions formulated above. An overview of the discussed strategies is provided in section 5 (Table 1).

1 Versioning as documentation

DSE projects of this type implement the assignment of sequential version numbers for changes and document what has been changed. A list of past version numbers with descriptions of the revisions and where they were made is provided. In these projects, the DSE is seen as an ongoing process, which is documented.

The strategy is pursued in the DSE "Auchinleck Manuscript", edited by David Burnley and Allison Wiggins. The project started in 2000 and was launched in July 2003 with version 1. The edition has a citation recommendation with version number indication. The editors recommend citing version 1.1 because it corresponds to the current version (March 15, 2004). The version number applies to the entire project. There is a version documentation with the title “Archive of site updates”. It consists of three lists: One with version number and date, another with "Corrections to texts and to textual notes" and an extra column "Record of other changes", which lists changes to the bibliography and the side menu. Allison Wiggins explains the version documentation theme as follows:

> All changes made to the content of this site are recorded here in the archive of site updates. Each time a batch of updates is added, the site is designated a new version number. This system ensure that users can keep track of any changes made and can reference the site materials accurately (Burnley and Wiggins, 2019).

At about the same time (March 2003), the DSE "The Old Bailey Proceedings Online" was published. It is an editorial long-term project that presents London processes from the period 1675-1913 and Tim Hitchcock, Robert Shoemaker, Clive Emsley, Sharon Howard and Jamie McLaughlin are responsible for the edition. A version documentation can be found in the "What's New Archive". As in the previous example, an introductory text refers to the non-final, process-like state of the DSE. It is followed by a listing and description of what has changed in previous versions: the current status is March 2018 with version 8.

A version documentation is also available in the edition project "The Online Froissart". The editors Peter Ainsworth and Godfried Croenen cite another reason for documenting versions:

> The Online Froissart is a collaborative, interdisciplinary and incremental project. Given the sheer size of Froissart’s Chroniques, the number of surviving manuscripts and their dispersal across many libraries in several different countries, it has been decided not to delay publication of available transcriptions until all of these materials have been transcribed and annotated. The Online Froissart, therefore, publishes all currently available transcriptions and other materials produced by the project team, plus updates, and augments the website and its underlying datasets on a regular basis (annually from 2012 onwards). (Ainsworth and Croenen 2018)

Due to the large amount of material and its scattered provenance, an intentional decision was made that the publication of the DSE should be as early as possible, but the edition will be regularly updated.

2 Versioning as version management

Versioning is implemented in the same way in the DSE as it can be found in software development. Version management systems save all changes to text documents as versions. All versions can be restored, and these systems are also designed to handle collaborative writing processes. Examples of DSEs that implement this versioning strategy are "Papyri.info" and "The Devonshire Manuscript".

“The Devonshire Manuscript” is a Wikibook edition, whose main editor is the Devonshire Manuscript Group.

The aim is to discuss the edited sources as widely as possible and to change the role of the scientific editor from the sole authority for the text to that of a moderator: “The social edition is a work that brings communities together to engage in conversation around a text formed and reformed through an ongoing, iterative, public editorial process.” (Wikibooks, 2014) One of the main contributors to this project is Ray Siemens (2012, 453), whose motivation to support the DSE with the help of social media is as follows:

> Such tools facilitate a model of textual interaction and intervention that encourage us to see the scholarly text as a process rather than a product, and the initial, primary editor as a facilitator, rather than a progenitor, of textual knowledge creation. (ibid.)

It is therefore about the editorial process as an iterative and collaborative activity. In this respect, it is essential for the progress of the project to keep all iterations available in the form of commented versions.

Like all pages of Wikipedia, all pages of a Wikibook are based on the same software MediaWiki (since version 1.5) and have a revision history (under the tab View history). The revision history in the form of a table contains all edits of a page in the wiki. Each change to a page creates a change line that contains information about the person who made the edit, the time when the edit was made, and a reference to the new wiki text in the text table. Elements of the revision table are preserved permanently, unless the page is deleted.

The versions of Papyri.info are accessible in a similar way, but via a different interface. Papyri.info aggregates papyrological resources from different databases and makes them available for editing. This DSE has been in existence since 2006 and interested editors can still add or change data today. A peer review of the revisions ensures the quality of the content. The implementation of the strongly social and collaborative project approach is made possible by a version management software that manages different users and their contributions. A method that was developed to facilitate the software development process is used here to manage the editorial processes. The DSE is stored in a Git repository. All editorial processes are recorded, versioned and recoverable. A look at the repository in Github shows 100 contributors[1] and more than 100,000 commits[2].

Git is a version control system that is used for collaborative software development. As already mentioned, the changes to the files are tracked. These programs provide access to any version of the file so that any changes can be undone. Each version has a timestamp and an author. It is always possible to see who changed what and when.

In the case of Papyri.info, the edited texts are saved as XML files in Git. Via the platform Github the repository can be viewed and the different versions of the texts can be displayed. The version comparison is done line by line and any changes to the file will be recognized by the software and automatically saved as a new version.

It makes no difference which version management software is used, the understanding of what a version is remains the same in the software development domain. In the examples mentioned so far, which have version documentation, the same version number stood for a whole series of revisions. If versioning takes place via version control systems, every saved change becomes a new version, which has no further semantic meaning. It makes no difference for the system if you make many changes or just fix a typo, it is always a new version. This strategy certainly has its advantages especially in a collaborative editing process.

3. Versioning as retrievable milestone versions

In contrast to the open DSEs mentioned above, which are geared towards a high frequency of changes, DSEs of this type are updated at intervals under new version names. These versions can also be called milestone versions, because the question of when to publish the next update is a project specific decision. Former versions are findable via a permalink and can be retrieved. A version name or number applies to the entire edition. This versioning strategy is evident in teams of editors who deliberately publish changes, revisions, and enhancements collectively. Every single editing step is not shown. What is desired as a research process in the case of Wikibooks is not part of the intention to publish an edition in this case.

This group includes the DSE "Der Sturm. Digital Source Edition on the History of the International Avant-Garde", developed and edited by Marjam Trautmann and Torsten Schrade since 2018. This edition project is still a "work in progress": the team of editors will gradually open up new sources and publish them promptly. The motive for this approach is explained as follows:

> Dies kommt den interessierten Forschungscommunities zugute, da somit ein schneller Zugriff auf eine beständig wachsende Gesamtedition gewährleistet ist. Ein weiterer Vorteil dieses iterativen Vorgehens ist, dass sich somit auch das Forschungsdatenmodell und die benötigten Softwarekomponenten kontinuierlich und in Einklang mit den hinzukommenden Quellen und ihren jeweiligen Spezifika weiterentwickeln können. (https://sturm-edition.de/projekt/methodik.html)

All developed sources have several permalinks that represent the versions. The permalinks are constructed in such a way that the identifier ends with the version information:

> Version 1: https://sturm-edition.de/id/Q.01.19140115.FMA.01/**1**
> Version 2: https://sturm-edition.de/id/Q.01.19140115.FMA.01/**2**

"Humboldt Digital" also integrates this type of versioning. The edition "Humboldt Digital" is a publication of the Academy Project "Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung" at the Berlin-Brandenburg Academy of Sciences and Humanities. In this project, each entry offers the possibility to view past milestone versions. The specific version number is inserted after the domain name. E.g. "https://edition-humboldt.de/**v4**/H0002656".

[1] Github Glossary: "A collaborator is a person with read and write access to a repository, who has been invited to contribute by the repository owner."

[2] Github Glossary: "A commit, or "revision", is an individual change to a file (or set of files). It's like when you save a file, except with Git, every time you save it creates a unique ID (a.k.a. the "SHA" or "hash") that allows you to keep record of what changes were made when and by who. Commits usually contain a commit message which is a brief description of what changes were made."

4. Mix type: Versioning as retrievable, documented milestone versions

If milestone versions are also documented, then all three of the above criteria for versioning DSEs are met: changes will be published at intervals as new versions. Old versions remain viewable and what the version stands for and what has been changed are documented.

An example of this is the "August Wilhelm Schlegel Edition". This DSE is about bringing together August Wilhelm Schlegel's correspondence. The project runs until 2020 under the direction of Jochen Strobel and Claudia Bamberg. The first beta version was published on 2 June 2014; version-07-19 was published on July 1, 2019. Specific versions can be addressed by entering the name after the domain. For example, a letter in the up-to-date version has a Permalink like this:

"https://august-wilhelm-schlegel.de/**version-07-19**/briefid/1599".

Under the menu item "Versions" is stated that every three months a new version with numerous resources will be published, while all previous versions remain fully accessible in the version archive.

The DSE "Johann Wolfgang Goethe: Faust", edited by Anne Bohnenkamp, Silke Henke and Fotis Jannidis, also has a version archive. The entire project is versioned at regular intervals and the current version is 1.2. In this case, the version is specified in the subdomain:

"http://v1-2.faustedition.net/document?sigil=B.a&page=59&view=print&section=5#l813"

5. Overview of applied versioning strategies

	version name	work in progress	collaborative	documented changes	former versions available
Versioning as documentation	X	X		X	
Versioning as version management	X	X	X	X	X
Versioning as retrievable milestone versions	X	X			X
Mix type: Versioning as retrievable, documented milestone versions	X	X	X	X	X

Table 1. Overview of applied versioning strategies. X means: is provided.

6. Conclusion

The empirical analysis of DSE has shown that integrating versioning is still relatively rare in DSE projects. This is the case despite the fact that there are manifold practical motives for implementing a versioning strategy. Such as the idea that the editorial process should be shared with the public; the project has a long duration and editors would like to publish interim results; the material to be edited is too extensive and will be made accessible in publishable stages; or the editorial team and the collections are so scattered that it would make sense and be beneficial for the overall project to publish partial results. For whatever reason, all projects intend to communicate changes and to be transparent in such a way that makes the DSE more reliable and trustworthy.

All projects presented here clearly pursue the strategy of making the DSE available to the public, although the editions are work in progress. The most technologically undemanding practice to communicate the changes is to maintain a version documentation. Many more ways to make changes transparent are given when the edition is linked to a version control system. In these cases, it is also easy to make a version comparison and

to understand the contributions of individual editors. Each resource has its own version history and can be retrieved. In the presented projects of the type 'versioning as retrievable milestone version' this is instead not the aim of the editors. The version number refers to the complete edition. Individual contributions can be searched under the milestone version number. A version control system is not absolutely necessary for the technical implementation, but it is an advantage. It is part of the workflow to have a published version and at the same time to have a new version in processing. The encapsulated data storage in the system is important for the addressability and availability of past versions. In addition, the information resources gain in quality if the traceability of the changes is also collectively available as documentation. By this means, versioning works like an apparatus in a broader sense, and one which verifies editorial decisions by making the work in progress transparent to the users.

References

Ainsworth, Peter, and Croenen, Godfried, eds. 2018. *The Online Froissart.* Version 8. Accessed 1 September 2019. https://www.dhi.ac.uk/onlinefroissart/apparatus.jsp

Burnley, David and Wiggins, Allison, eds. 2004. *Auchinleck Manuscript.* Version 1.1. Accessed 1 September 2019. https://auchinleck.nls.uk/editorial/refer_site.html

Bleier, Roman 2019: "How to cite this digital edition?" Forthcoming.

Bohnenkamp, Anne, Henke, Silke, and Jannidis, Fotis, eds. 2019. *Johann Wolfgang Goethe: Faust.* Version 1.2. Accessed 1 September, 2019. http://v1-2.faustedition.net/document

Driscoll, Matthew James, and Pierazzo, Elena eds. 2016. Digital Scholarly Editing: Theories and Practices, edited by, Cambridge, UK: Open Book Publishers. DOI: 10.11647/OBP.0095.02.

Ette, Ottmar, ed. 2019. *edition humboldt digital.* Berlin-Brandenburgische Akademie der Wissenschaften, Berlin. Version 5. Accessed 1 September, 2019. https://edition-humboldt.de/index.xql?l=de

Github, Glossary: *Contributors* https://help.github.com/en/articles/github-glossary Accessed 14 August 2019.

--- *Commit* https://help.github.com/en/articles/github-glossary Accessed 14 August 2019.

Hitchcock, Tim, Shoemaker, Robert, Emsley, Clive, Howard, Sharon, and McLaughlin, Jamie, eds. 2018. *The Old Bailey Proceedings Online* Version 8. Accessed 1 September, 2019. https://www.oldbaileyonline.org/

MLA Committee on Scholarly Editions. MLA Statement on the Scholarly Edition in the Digital Age. Web publication, May 2016, https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/MLA-Statement-on-the-Scholarly-Edition-in-the-Digital-Age

Österreichische Akademie der Wissenschaften (ÖAW), eds.2019. *Catalogue of digital editions.* Accessed 1 September 2019. https://dig-ed-cat.acdh.oeaw.ac.at/

Siemens, Ray et al. 2012. *Toward Modeling the Social Edition: An Approach to Understanding the Electronic Scholarly Edition in the Context of New and Emerging Social Media.* In: Humanities Commons: 445 - 461 https://hcommons.org/deposits/item/mla:83/, http://dx.doi.org/10.17613/M6KS3G

Strobel, Jochen, and Bamberg, Claudia, eds. 2019. *August Wilhelm Schlegel Edition.* Version-07-19. Accessed 14 August, 2019. https://august-wilhelm-schlegel.de/version-07-19

Trautmann, Marjam, and Schrade, Torsten, eds. 2019. *Der Sturm. Digital Source Edition on the History of the International Avant-Garde.* Accessed 1 September 2019. https://sturm-edition.de/

The Duke Collaboratory for Classics Computing and the Institute for the Study of the Ancient World, eds. 2019. Papyri.info Accessed 14 August 2019. http://papyri.info/

Università di Parma: Linee Guida per la Buona Pratica Scientifica e Disseminazione della Ricerca. 2018, https://www.unipr.it/sites/default/files/albo_pretorio/allegati/19-06-2018/linee_guida_dr_emanazione.pdf

Wikibooks,The Free Textbook Project. 20 Sep 2018. *The Devonshire Manuscrip*t. Accessed 15 September 2019. https://en.wikibooks.org/w/index.php?title=The_Devonshire_Manuscript&oldid=3469218

Wikibooks, The Free Textbook Project. 23 May 2014. *The Devonshire Manuscript/A Note on this Edition.* Accessed 15 September 2019. https://en.wikibooks.org/w/index.php?title=The_Devonshire_Manuscript/A_Note_on_this_Edition&oldid=265930_6

ELA: fasi del progetto, bilanci e prospettive

Emmanuela Carbé
Università di Siena - QuestIT
emmanuela.carbe@unisi.it

Nicola Giannelli
Università di Siena - QuestIT
giannelli@quest-it.com

Abstract

English. The paper considers the phases of a start-up project (March 2019-February 2020) developed with the aim of building ELA-Eurasian Latin Archive, a digital platform containing Latin and multilingual texts from 12th to 18th century concerning East Asia. The balance includes a general evaluation of the work, beginning with the initial stage and the project planning up to the "lesson learned", as well as some reflections on sustainability and expected future implementations.

Italiano. Il contributo ripercorre le fasi di un progetto di start-up (marzo 2019-febbraio 2020) nato per la realizzazione di ELA-Eurasian Latin Archive, una piattaforma digitale di testi latini e multilingua dei secoli XII-XVIII riguardanti l'Estremo Oriente. Il bilancio qui proposto comprende una valutazione generale del lavoro a partire dall'avvio e dalla pianificazione del progetto fino alle "lezioni apprese", con alcune riflessioni sugli sviluppi attesi in termini di sostenibilità e implementazioni future.

1 Introduzione

Tra le cinque fasi che dovrebbero caratterizzare ogni progetto, ovvero avvio, pianificazione, esecuzione, controllo e chiusura, l'ultima viene talvolta sottovalutata o non messa in atto, dimenticando che un progetto per definirsi tale deve avere le caratteristiche di unicità e durata limitata, con la produzione di un risultato il più possibile in linea con il piano di costi, tempo e qualità stabiliti. Ne consegue il rischio di trasformare un lavoro in un pericoloso work-in-progress, che può subire improvvisi arresti per mancanza di copertura finanziaria, per lo scioglimento del gruppo di lavoro, o per altre cause interne ed esterne. La qualità di un progetto andrebbe dunque valutata anche nella sua capacità di arrivare a una conclusione con la formalizzazione delle lezioni apprese (Mastrofini, 2017), incluse quelle meno positive (Dombrowski, 2019), al fine di costruire un patrimonio comune di conoscenze non meno importante del progetto stesso. Questo patrimonio, se conservato e condiviso, può essere utile per realizzazioni future e per la costruzione di nuove reti di collaborazione.

Il presente contributo si propone di fare il punto su un progetto di start-up, DAS-MeMo[1], avviato a marzo 2018 dall'Università di Siena sotto la direzione di Francesco Stella e in collaborazione con l'azienda QuestIT e la casa editrice Pacini. Questa fase del progetto prevede la realizzazione della piattaforma ELA-Eurasian Latin Archive, che raccoglie documenti in lingua latina e multilingua dal XII al XVIII secolo riguardanti l'Estremo Oriente. Il progetto è inserito all'interno di un contesto più ampio, caratterizzato da metodologie e strumenti di lavoro già collaudati grazie a numerose esperienze del PI e del suo gruppo di lavoro nell'ambito dei progetti digitali. Viene dunque tracciata la storia del progetto in tutte le sue fasi, dall'avvio del lavoro alla sua chiusura, prevista per febbraio 2020, e presentato un bilancio di ciò che è stato realizzato, delle lezioni apprese, e degli sviluppi futuri con l'avvio della fase di consolidamento e implementazione.

[1] DAS-MeMo (Data Mining e analisi statistica su fonti testuali storiche del periodo medievale e moderno, www.dasmemo.unisi.it) ha ricevuto il contributo di Regione Toscana per un assegno di ricerca cofinanziato con le risorse POR FSE 2014-2020 nell'ambito del progetto Giovanisì.

2 Metodologia di progetto

Nella fase di avvio è stata elaborata la documentazione per il piano del progetto, basato su obiettivi, tempi e risorse. Nel corso della pianificazione sono stati individuati i passi da realizzare per il raggiungimento degli obiettivi, sono stati assegnati i ruoli e decise le tempistiche delle consegne. È stata condotta un'analisi SWOT per identificare fin da subito le possibili problematiche e i fattori di rischio (si veda, a titolo di esempio, l'analisi applicata al caso BEIC di Consonni e Weston, 2015). La fase di start-up, della durata di 24 mesi, aveva i seguenti obiettivi: 1. Creazione di un modello e di un workflow di lavoro, includendo momenti di revisione e di controllo della qualità. 2. Definizione e creazione del corpus, con relativa codifica. 3. Progettazione, analisi dei requisiti e realizzazione, in collaborazione con l'azienda partner del progetto, di un prototipo della piattaforma, con: a. pagine informative gestibili tramite un comune CMS; b. digital library; c. tool di analisi linguistiche e semantiche; d. backend per la gestione della Digital Library. 4. Indagine preliminare di una parte del corpus con primi risultati, pubblicati in e-book grazie all'editore partner del progetto. 5. Disseminazione e comunicazione dei risultati su più livelli; 6. Realizzazione di un piano di sostenibilità. Dalla pianificazione del progetto è nata una lista di specifiche basate su metodo MoSCoW (vd. par. 3). Il progetto è stato monitorato sia internamente, con un controllo costante dello stato di avanzamento dei lavori anche per eventuali modifiche del cronoprogramma, sia esternamente, con relazioni intermedie inviate ai soggetti cofinanziatori del progetto.

In questa ultima fase, ormai prossima alla chiusura dei lavori, rimane dunque da compiere una revisione generale e una valutazione di ciò che è stato fatto. In un intervento sul sito del King's Digital Lab, Arianna Ciula (2019a) spiega le motivazioni che hanno portato alla realizzazione della *Checklist for Digital Outputs Assessment* (Ciula, 2019b), una guida pubblicata per facilitare la revisione dei progetti in vista della prossima valutazione REF (Research Excellence Framework), utilizzata nel Regno Unito per monitorare la qualità della ricerca. Ciula rileva che se in altre discipline una valutazione e autovalutazione del lavoro parte da basi piuttosto definite, quando si opera nel contesto di progetti digitali stabilire dei criteri condivisi comporta criticità e incertezze. La checklist propone dodici punti tematici (con relativi esempi) per la valutazione di prodotti digitali, che abbiamo deciso di adottare in questa fase finale. I punti servono per verificare vari aspetti del progetto: i credits, con la corretta attribuzione dei lavori svolti (incluse le realizzazioni dei data model e dell'architettura) e degli enti che hanno partecipato e/o finanziato il progetto, che devono essere correttamente menzionati con i loro loghi; il controllo delle licenze e copyright, che devono essere esplicite e il più possibile conformi ai principi FAIR; la messa a disposizione della documentazione relativa al progetto, che includa esplicitamente una riflessione sul valore aggiunto dei risultati conseguiti; l'attenzione all'accessibilità, alla *user experience* e alla funzionalità dell'interfaccia; controllo delle versioni del prodotto, con la conservazione delle eventuali versioni precedenti; la presenza di indicazioni su come citare, di identificatori persistenti e DOI; il piano di sostenibilità e accessibilità e delle informazioni sull'utilizzo del prodotto.

3 Analisi del corpus e creazione della piattaforma

Dopo una valutazione dei casi di studio, e in particolare di ALIM – Archivio della Latinità Italiana del Medioevo (Russo, 2005; Ferrarini, 2017; Manos, 2018), si è proceduto alla definizione del corpus attraverso un censimento di trecento testi, effettuato sulla base di alcuni repertori cartacei e online (tra questi: Bibliotheca Sinica 2.0 e CCT-Christian Texts Database). Il censimento è stato allestito e arricchito con l'utilizzo di OpenRefine (Hooland, Verborgh and De Vilde, 2013; Williamson, 2017), includendo i riferimenti bibliografici di ogni record, l'eventuale presenza di digitalizzazioni online e i diritti di utilizzo della risorsa. In corso d'opera sono stati aggregati metadati relativi allo stato di avanzamento del progetto: se un testo è preso in carico dal gruppo di lavoro, sono aggiunte informazioni circa il software utilizzato per l'eventuale digitalizzazione o trattamento degli OCR, il responsabile della trascrizione e codifica (con relativo ORCID), i tempi di realizzazione, la licenza di pubblicazione, il grado di attendibilità della risorsa.

È stato poi creato un modello di lavoro per la realizzazione della piattaforma, basato sulla MoSCoW analysis, che ha permesso di focalizzare meglio gli obiettivi prioritari (Must have), i desiderabili ma non essenziali (Should have), i desiderabili ma non strettamente necessari (Could Have), e quelli che possono

essere pianificati per il futuro (Would have). Si dà qui conto dei contenuti essenziali: 1. Must have: una piattaforma che raccoglie testi latini contenenti inserti multilingua (cinese, giapponese, coreano, ma anche pinyin); un solido motore di ricerca, in grado di realizzare ricerche mirate anche con filtri e indicizzazioni in base a una lista predefinita di metadati; un modello di codifica in XML TEI; possibilità di visualizzare il testo e scaricarlo in più formati (TXT; PDF; XML); un backend per il caricamento e la gestione dei documenti; un identificativo persistente per ciascun documento; un set base di tool per analizzare i documenti (indici di parole, lemmi, type, stopwords; frequenze assolute e relative; concordanze; Type/Token Ratio, N-grams; numero totale di parole per documento e altre analisi quantitative di semplice acquisizione); definizione della *user policy*; messa a punto di metodi di lavoro collaborativi per future implementazioni del progetto con più unità di ricerca (es. Wiki). 2. Should have: tecniche di georeferenziazione; codifica di luoghi, date e persone menzionati nei testi; strumenti più raffinati per l'analisi linguistica dei testi (PoS, Burrows Delta, frequenze in base a sillabe) anche in considerazione del problema specifico che pongono i documenti multilingua; analisi semantica dei testi; topic modeling; un progetto più complesso di backend per i collaboratori, con la possibilità di modificare i documenti attraverso un editor di testo e di gestire il flusso di lavoro. 3. Could have: un progetto più specifico per alcune tipologie di documenti, come ad esempio le lettere, numerose all'interno del corpus; la possibilità di creazione, da parte dell'utente, di un corpus personalizzato con analisi testuali comparate attivando processi in tempo reale; integrazione della codifica TEI con modelli semantici (Ciotti et al., 2016; Ciotti, 2018); Named Entity Recognition per luoghi, persone e date (Erdmann et al., 2016; Simon et al., 2017); 4. Would have: inclusione delle digitalizzazioni dei documenti (Rosselli Del Turco, 2015; 2019), che talvolta contengono disegni, mappe e altre rappresentazione grafiche di particolare rilevanza.

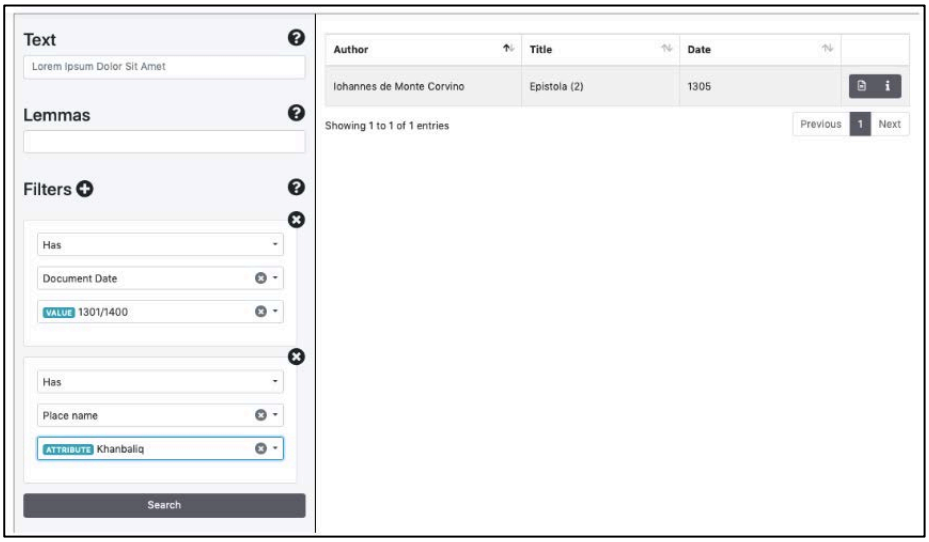


Figura 1. Il prototipo della piattaforma.

Il prototipo della piattaforma è stato realizzato con più componenti: per l'interfaccia è stato utilizzato il CMS Wordpress, scelto soprattutto per una facile gestione delle pagine informative e per la pubblicazione della documentazione del progetto; la Digital Library è sviluppata in Java EE, con una UI basata su tecnologie web e realizzata utilizzando Javascript; si basa sul motore di ricerca Elastic Search con possibilità di effettuare ricerche full text anche con l'utilizzo della sintassi Lucene. Al suo interno è stato aggiunto un tool dedicato alle analisi linguistiche, attualmente alla sua versione 1.0: si tratta di un strumento realizzato in Python e basato su CLTK (Burns, 2019) e NLTK (Bird et al., 2015), chiamato ELA-tool, che restituisce in formato JSON i risultati delle analisi linguistiche previste dai requisiti primari del progetto (Must Have). Nel momento in cui avviene l'upload del documento, ELA-tool processa il testo. I risultati vengono poi memorizzati e utilizzati sia per la visualizzazione delle analisi linguistiche, sia da Elastic Search per raffinare le possibilità di ricerca. Nell'ottica di poter effettuare continue migliorie del tool procedendo con il rilascio

di nuove versioni perfezionate, è stato previsto che nel backend per l'upload e la gestione dei documenti sia presente una funzione di refresh che esegue i processi di ELA-tool sui testi già caricati in precedenza su esplicita richiesta di un utente amministratore della piattaforma.

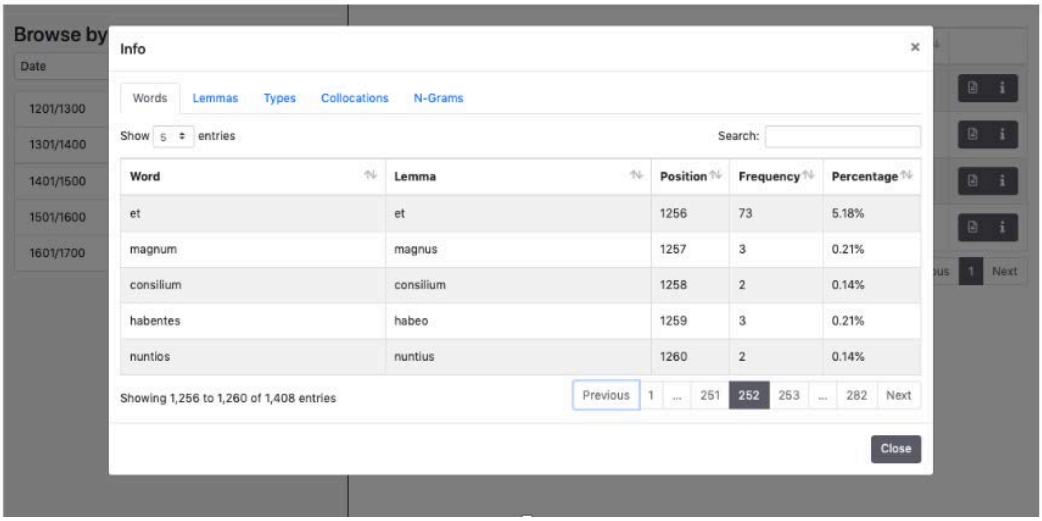


Figura 2. Visualizzazione dei risultati processati dal tool di analisi linguistica.

I testi sono codificati in XML seguendo lo schema TEI P5 (Tei Consortium 2015), con un TEI header modellato grazie al censimento realizzato con OpenRefine; si pone una particolare attenzione alla codifica di luoghi, date, persone (Wikidata, VIAF e, per i luoghi, primariamente Pleiades Gazetteer) e l'eventuale presenza di inserti in lingue diverse dal latino (è il caso, ad esempio, di *Sapientia Sinica* di Costa e Intorcetta).

4 Bilanci, prospettive

Alla conclusione della fase di start-up, ELA ospiterà i primi cento testi tratti dal censimento, scelti per importanza e per varietà nelle caratteristiche, per permettere una verifica sul campo del modello di codifica scelto e intervenire con correzioni in corso di implementazione. La piattaforma sarà ospitata presso il centro di calcolo dell'Università di Siena, con un progetto di business continuity e disaster recovery, con un piano programmato di snapshot e backup.
Rispetto agli obiettivi del progetto, i "must have" risultano oggi completamente raggiunti, così come quasi tutti i "should have". In questo contesto, pare utile una revisione e valutazione del lavoro sulla base di indagini qualitative e quantitative. Tra queste, appare prioritaria un'analisi dei risultati ottenuti dal sistema di lemmatizzazione di CLTK, attraverso una comparazione con altri lemmatizzatori, sulla scorta di Mambrini e Passarotti (2019) e Eger et al. (2015, 2016). La revisione del lavoro potrà includere anche una valutazione sulla *user experience* della piattaforma, con un questionario per gli utenti che utilizzano il prototipo.
Se la fase di start-up è prossima alla sua conclusione, quella di consolidamento e implementazione di Eurasian Latin Archive va ora pianificata nell'ottica della sostenibilità sul medio e lungo periodo per il raggiungimento di risultati più ambiziosi: l'avvio a regime della piattaforma e la programmazione del suo incremento in termini di numero di documenti trattati, con un piano redazionale che comprenda anche la pubblicazione e il mantenimento di tutte le sezioni della piattaforma; l'ampliamento dell'utenza prevista: allo stato attuale ELA è pensato soprattutto per un pubblico specializzato, ma non si esclude, in futuro, la possibilità di rivolgersi a una comunità più ampia di utenti, anche attraverso nuovi percorsi di disseminazione e riutilizzo dei materiali; infine l'integrazione con altri progetti: ELA è stato primariamente pensato per essere interoperabile con ALIM, tuttavia si ritengono necessari, anche per la sostenibilità del

progetto stesso, la condivisione dei dati e degli strumenti realizzati (che saranno messi a disposizione su GitHub) e un dialogo costante per operare a fianco di altri progetti in corso (Passarotti et al., 2019).

Bibliografia

Patrick J. Burns. 2019. Building a Text Analysis Pipeline for Classical Languages. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, ed. by M. Berti, De Gruyter, Berlin, Boston:159-176. DOI: 10.1515/9783110599572-010*

Steven Bird, Erwan Klein, and Edward Loper. 2015. *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit* [version updated for Python 3 and NLTK 3]. URL: `https://www.nltk.org/book/`

Fabio Ciotti. 2018. A Formal Ontology for the Text Encoding Initiative. *Umanistica Digitale,* 3:137-153. DOI: 10.6092/issn.2532-8816/8174.

Fabio Ciotti, Marilena Daquino, Francesca Tomasi. 2016. *Text Encoding Initiative Semantic Modeling. A Conceptual Workflow Proposal*. Digital Libraries on the Move, ed. by D. Calvanese, D. De Nart, C. Tasso C., vol. 612, Springer, Cham. DOI: 10.1007/978-3-319-41938-1_5.

Arianna Ciula. 2019a. What Makes Good Honey? KDL Checklist for Digital Outputs Assessment in the REF. *Thoughts and reflections from the Lab.* Aug. 7, 2019.

Arianna Ciula. 2019b. *KDL Checklist for Digital Outputs Assessment*. Aug. 6, 2019. DOI: 10.5281/zenodo.3361580

Chiara Consonni and Paul G. Weston. 2015. Finding a Needle in a Haystack. *Digital Libraries on the Move*, ed. by D. Calvanese, D. De Nart, C. Tasso, IRCDL 2015. Communications in Computer and Information Science, vol. 612. Springer, Cham. DOI: 10.1007/978-3-319-41938-1_18.

Quinn Dombrowski. 2019. *Towards a Taxonomy of Failure*, Jan. 30, 2019.

Steffen Eger, Rüdiger Gleim and Alexander Mehler. 2016. Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. Proceedings of the 10th International Conference on Language Resources and Evaluation. European Language Resources Association:1507-1513.

Steffen Eger, Tim Vor der Brück and Alexander Mehler. 2015. Lexicon-assisted Tagging and Lemmatization in Latin: A Comparison of Six Taggers and Two Lemmatization Methods. Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). Association for Computational Linguistics:105–113.

Alexander Erdmann et al. 2016. Challenges and Solutions for Latin Named Entity Recognition. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*:85-93.

Edoardo Ferrarini. 2017. ALIM ieri e oggi. *Umanistica digitale*, 1(2017):7-17. DOI: 10.6092/issn.2532-8816/7193.

Seth van Hooland, Ruben Verborgh and Max De Vilde. 2013. Cleaning Data with Open Refine. *The Programming Historian*, 2.

Francesco Mambrini and Marco Passarotti. 2019. *Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin*. Proceedings of the 13th Linguistic Annotation Workshop (LAW XII), Association for Computational Linguistics, Florence, 2019, ed. by A. Friedrich and D. Zeyrek:71-80.

Traianos Manos. 2018. ALIM: Archivio della Latinità Italiana del Medioevo. Accessed October 20, 2017. DM Reviews - June 2018. *Digital Medievalist*, 11(1): 4. DOI: 10.16995/dm.79.

Enrico Mastrofini. 2017. *Guida ai temi ed ai processi di project management. Conoscenze avanzate e abilità per la gestione dei progetti*. ISPM – Istituto italiano di Project Management, Franco Angeli, Milano.

Passarotti Marco et al. 2019. Lila: Linking Latin – A Knowledge Base of Linguistic Resources at NLP Tools. Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LKD-PS 2019), Leipzig, May 2019, ed. by T. Declerck and J.P. McCrae:20-23.

Roberto Rosselli del Turco et al. 2015. Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions. *Journal of the Text Encoding Initiative*, 8:1-21. DOI: 10.400/jtei.1077.

Roberto Rosselli del Turco et al. 2019. Visualisation with EVT: Simplicity is Complex. *Poster session of the Digital Humanities Conference 2019, Utrecht*, July 2019.

Luigi Russo. 2005. ALIM, Archivio della latinità italiana del Medioevo. *Reti Medievali Rivista* 6, 1(2005):149-151. DOI: 10.6092/1593-2214/181.

Rainer Simon et al. 2017. Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2. *Journal of Map & Geography Libraries*, 13(1):111-132.

Tei Consortium. 2015. *P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.6.0. Last updated on 16th July 2019.

Evan Peter Williamson. 2017. Fetching and Parsing Data from the Web with Open Refine. *The Programming Historian, 6.*

Digitized and Digitalized Humanities: Words and Identity

Claire Clivaz
Swiss Institute of Bioinformatics
claire.clivaz@sib.swiss

Abstract

English. This paper analyses two closely related but different concepts, *digitization* and *digitalization*, first discussed in an encyclopedia article by Brennen and Kreiss in 2016. Digital Humanities mainly uses the first term, whereas business and economics tend to use the second to praise the process of the digitalization of society. But *digitalization* was coined as a critical concept in 1971 by Wachal and is sometimes used in post-colonial studies. Consequently, humanist scholars are invited to avoid the "path of least resistance" when using *digitalization*, and to explore its critical potential. The paper concludes by considering the effect of the digitalization perspective and by expressing author's point of view on the issue.

Italiano. Questo articolo analizza due concetti correlati ma differenti fra loro, "digitization" e "digitalization", discussi la prima volta in una voce di enciclopedia da Brennen e Kreiss nel 2016. Nelle scienze umane digitali si utilizza sostanzialmente il primo termine, mentre in economia si tende a utilizzare il secondo per sottolineare il processo di digitalizzazione della società. Ma il termine "digitalization" era stato creato nel 1971 da Wachal come un concetto critico, ed era stato utilizzato in alcuni studi sul post-colonialismo. Di conseguenza, gli studiosi nelle scienze umane sono invitati a evitare di utilizzare "digitalization" in modo triviale, e ad esplorare il suo potenziale critico. L'articolo termina con alcune considerazioni sugli effetti della prospettiva della digitalizzazione, presentando il punto di vista dell'autore.

1 Introduction: Words and Identity in Digital Humanities

As the 2020 AIUCD conference topic underlines, the identity and definition of the Humanities that has met the computing world, is in constant reshaping (Ciotti, 2019)[1]. The English language has acknowledged the important turn from *humanities computing* to *digital humanities* at the beginning of the 21st century (Kirschenbaum, 2010), whereas French-speaking scholarship is wrestling between *humanités numériques* (Berra, 2012; Doueihi, 2014) and *humanités digitales* (LeDeuff, 2016; Cormerais–Gilbert, 2016; Clivaz, 2019). Moreover, new words are often tested to express the intensity of what is at stake: if Jones has chosen the term "eversion" for describing the present state of the digital turn (Jones, 2016), the French thinker Bernard Stiegler focuses on "disruption" (Stiegler, 2016). German and Hebrew link digital humanities naming with the vocabulary of spirit/mind, whereas the outmoded word *humanités* has come back in French through the naming of the *humanités numériques*, recalling the presence of the body (Clivaz, 2017).

Inscribed in this linguistic effervescence, a phenomenon has so far not drawn the attention of the humanist scholarship: the difference between *digitization* and *digitalization*, or between *digitized* and *digitalized* Humanities. The present paper will explore, as far as possible, the emergence of this dualistic vocabulary, inside and outside of digital humanities scholarship, looking for its meanings and implications. It represents only a first overview about the scare definitions and occasional uses of "digitalization", even if the debate between digitization and digitalization can sometimes inform implicitly the discourse, as we will see in Section 4 (Smithies, 2017). Section 2 will first comment similarity and difference between both words, looking for "digitalization" definitions, and its uses. Section 3 discusses in detail the only definition article we have so far debating these two concepts. Section 4 considers more broadly the digitalization perspective and presents the author's point of view on the issue, including its articulation to the AIUCD 2020 topic.

2 Looking for "digitalization" definition and uses

English native speakers would surely ask first if there is really a difference between "digitization" and "digitalization." "Digitalization" does not benefit from its own entry in Wikipedia or in the *Collins Dictionary*

[1] Many thanks are due to the reviewers for their remarks, to Andrea Stevens for her English proof-reading, and to Elena Giglia for her translation of the Italian abstract.

online.[2] However, the *Oxford English Dictionary* (OED) dates the first use of digitalization as equivalent to digitization in 1959,[3] whereas the medical sense appeared in 1876.[4] OED presents also *digitalization* as meaning "the adoption or increase in use of digital or computer technology by an organization, industry, country, etc."[5] In the Wikipedia entry "digital transformation", a similar definition is given for "digitalization": "unlike digitization, *digitalization* is the 'organizational process' or 'business process' of the technologically-induced change within industries, organizations, markets and branches."[6] A most decisive shift in the sense of a difference between the two words can be seen in the *International Encyclopedia of Communication Theory and Philosophy*, which published an entry on "Digitalization" by J. Scott Brennen and Daniel Kreiss in 2016. They argue in favour of a distinction from "digitization" (Brennen–Kreiss, 2016). This publication is in itself a quite clear signal, according to our cultural and scholarly habits, that "digitalization" exists with its own meanings, since it has been defined in an encyclopedia. As far as I have been able to determine, it is the only article trying to define both concepts and is discussed in detail in Section 3.

As we see, references to digitalization's definition are quite scare. So far, there it is not even possible to do a systematic overview of its theoretical background based in the scholarly literature because it is not discussed, with the exception of the Brennen–Kreiss article. But if we look at its uses, some aspects clearly emerge. "Digitalization" is mainly used in the business and economical world, and very infrequently in digital humanities. For example, according to Jari Collin in a 2015 Finnish volume of collected essays, digitalization refers to the understanding of "the dualistic role of IT in order to make right strategic decisions on IT priorities and on the budget for the coming years. IT should not be seen only as a cost center function anymore!" (Collin, 2015, 30). Digitalization seems to be "one of the major trends changing society and business. Digitalization causes changes for companies due to the adoption of digital technologies in the organization or in the operation environment" (Parvianien et al., 2017, 63).

According to Mäenpää and Korhonen, "from the retail business point of view, the 'digitalization of the consumer' is of essence. People are increasingly able to use digital services and are even beginning to expect them. To a certain extent, this is a generational issue. The younger generations, such as Millennials, are growing up with digitalization and are eagerly in the forefront of adopting new technology and its affordances" (Mäenpää–Korhonen, 2015, 90). In 2018, Toni Ryynäen and Torsti Hyyryläinen, members of the *Helsinki Institute of Sustainability Science* at the *Faculty of Agriculture and Forestry*, published an article seeking to fill the gap between the digitalization process and digital humanities, by focusing on the concern for "new forms of e-commerce, changing consumer roles and the digital virtual consumption" (Ryynäen – Hyyryläinen, 2018, 1). In this process, the role of digital humanities is described in a way that is quite hard to recognize for DHers, at least for those not involved in digital social sciences: "A challenge for digital humanities research is how to outline the most interesting phenomena from the endless pool of consumption activities and practices. Another challenge is how to define a combination of accessible datasets needed for solving the chosen research tasks" (Ryynäen – Hyyryläinen, 2018, 1).

In light of such clear descriptions of what "digitalization" means for business and economy, digital humanities scholarship demonstrates a deafening silence about this notion. The 2004 and 2016 editions of the reference work *Companion to Digital Humanities* do not mention the word. In the established series *Debates in the Digital Humanities*, one finds one occurrence in the five volumes, under the pen of Domenico Fiormonte (2016). As a third example, the collected essays *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, edited by Willard McCarty (2010), can only surprise the reader: indeed, "digitalization" stands in the title, but the word is then totally absent from the volume. When questioned about this discrepancy, McCarty answered that the publisher had requested to have this word in the title. This request has led to a damaging side effect in terms of Google searches: if one searches for "digitalization" and "digital humanities", one gets several book titles that do not contain no mention of this word other than a reference to *Text and Genre*'s title. It is also the case in my 2019 book *Ecritures digitales.*

[2] Entry "digitization" in Wikipedia: https://en.wikipedia.org/wiki/Digitization; entry "digitalize" in the *Collins Dictionary* online: https://www.collinsdictionary.com/dictionary/english/digitalize. All hyperlinks have been last checked on 30/11/19.

[3] Entry "digitalization n.2", *OED*, https://www.oed.com/view/Entry/242061

[4] Entry "digitalization n.1", *OED*, https://www.oed.com/view/Entry/52616: "the administration of digitalis or any of its constituent cardiac glycosides to a person or animal, esp. in such a way as to achieve and maintain optimum blood levels of the drug. Also: the physiological condition resulting from this".

[5] Entry "digitalization n.2" in the *Oxford English Dictionary* online: https://www.oed.com/view/Entry/24206

[6] Entry "digital transformation" in Wikipedia: https://en.wikipedia.org/wiki/Digital_transformation#Digitization_(of_information)

Digital writing, digital Scriptures: the unique occurrence of "digitalization" occurs in my reference to McCarty's collected essays (Clivaz, 2019).

One can sometimes meet infrequent uses of digitalization in digital humanities, such as a 2013 article by Amelia Sanz. She uses the word to describe Google Books and the Hathi Trust's effect on Spanish literature: "Digital Libraries as Google Books or Hathi Trust include numerous works belonging to our study period among its *digitalized* collections in US universities, because most of these forgotten authors make part of the Spanish diaspora after the Civil War (1936-39) and during the subsequent dictatorship (1939-1975). In fact, European copyright legislation has made Google *digitalize* only works prior to 1870 in Spain, and, unfortunately for Spanish researchers, those works appear to be in 'limited access' due to the existing diffusion/circulation rights, but available in 'full text' mode for researchers located in the US" (Sanz, 2013, n.p.). The two italicized words are the unique occurrences of *digitalization* vocabulary in an article focused on the effects of digitization. When asked about her use of these two words, Sanz answered that it was probably a misuse of language, since she is not a native English speaker.

Usually in digital humanities scholarship, one speaks about "Humanities digitized" (Shaw, 2012)[7], and the mutation to the digital sphere is seen as a pre-step before the processes of interpretation.[8] Uses of digitalization and cognate terms remain rare, like Domenico Fiormonte, who is also a non-native English speaker and the only one to use *digitalization* in the series *Debates in Digital Humanities*: "In the last ten years, the extended colonization, both material and symbolical, of digital technologies has completely overwhelmed the research and educational world. Digitalization has become not only a vogue or an imperative, but a normality. In this sort of 'gold rush', the digital humanities perhaps have been losing their original openness and revolutionary potential" (Fiormonte, 2016, n.p.). Fiormonte compares digitalization to a colonization process: if there is some consciousness of the digitalization vocabulary in humanities, it can be indeed found in research about cultural diversity and colonialism, such as in a 2007 article by Maja van der Velden, "Invisibility and the Ethics of the Digitalization: Designing so as not to Hurt Others."

Van der Velden studies "the designs of Indymedia, an Internet-based alternative media network, and TAMI, an Aboriginal database, [...] informed by the confrontations over different ways of knowing" (2007, 81). She points to the fact that, "if we understand knowledge not as a commodity but as a process of knowing, something produced socially, we must ask about the nature of digitalization itself. As the Aboriginal elders say, 'Things are not real without their story'" (2007, 82). She documents in this way two examples of non-Western digital projects, in which the diversity of the source codes and standards has led to recurrent negotiations: "the confrontations over issues of privacy and control resulted in different ways of organizing access and information management" (2007, 89). Van der Velden's article allows one to understand, from a humanist point of view, what is at stake in the concept of *digitalization*, a perspective that the next section develops. But it should be underlined that, even in this article pointing to cultural and digital control issues, *digitalization* is not discussed as such. The apparent lack of awareness about this binomial vocabulary and its implication for DH scholarly literature appears to be a real blind spot that section 4 considers.

3 Claiming a Critical Use of *Digitalization* in Humanities

In their overview article, Brennen and Kreiss give a general definition of "digitalization" similar to the one presented in Section 2: "We [...] define digitization as the material process of converting analog streams of information into digital bits. In contrast, we refer to digitalization as the way many domains of social life are restructured around digital communication and media infrastructures" (Brennen–Kreiss, 2016, 1). They usefully remind us that "digitization is a process that has both symbolic and material dimensions" (2016, 2), and that "analog and digital media, [...] all forms of mediation necessarily interpret the world" (2016, 3). The authors also consider that "the first contemporary use of the term 'digitalization' in conjunction with computerization appeared in a 1971 essay first published in the North American Review. In it, Robert Wachal discusses the social implications of the 'digitalization of society' in the context of considering objections to, and the potential for, computer-assisted humanities research. From this beginning, writing about digitalization has grown into a massive literature" (2016, 5). The reference to Wachal's article is a very interesting one, and it deserves more attention than the co-authors devote to it. Moreover, they omit any reference to Maja van der Velden's article or to similar approaches in Brennen and Kreiss's article. The "winners" of their digitalization

[7] One can also see uses of *digitalization* in the humanities in archaeology, notably in conjunction with 3D discussion (Ercek –Viviers –Warzée, 2009).

[8] See Earhart – Taylor (2016): "Our *White Violence, Black Resistance* project merges foundational digital humanities approaches with issues of social justice by engaging students and the community in digitizing and interpreting historical moments of racial conflicts."

definition are scholars from the vein of Manuel Castells, who argues that "technology is society, and society cannot be understood or represented without its technological tools" (Brennen–Kreiss, 2016, 5).

To get a deeper understanding of the critical potential of *digitalization*, it is worth reading Wachal's 1971 article. He uses *digitalization* in just one sentence: "The humanist's fears are not entirely without foundation, and in any case, as a humane man he naturally fears the digitalization of the society. He doesn't like to be computed. He doesn't want to be randomly fingered by a credit card company computer" (1971, 30). The entire article is an ironic confrontation between the habits of a humanist scholar and what a programmer and a computer could do for humanities. As a computer programmer teacher himself, Wachal remembers the term coined by Theodor Nelson, "cybercrud": "putting things over on people [by] saying using computers. When you consider that this includes everything from intimidation ('Because we are using automatic computers, it is necessary to assign common expiration dates to all subscriptions') to mal implementation ('You're going to have to shorten your name - it doesn't fit in to the computer'), it may be that cybercrud is one of the most important activities of the computer field" (1971, 30). In other words, computer scholars have a clear awareness about their world, as Nelson and Wachal after him demonstrate. After this *captatio benevolentiae*, Wachal raises what is for him the main issue with the humanist point of view on computing: "Dare we hope that the day has come when humanists will begin asking some new questions?" (1971, 33), referring also to artificial intelligence (1971, 31). His "personal view", as announced in the article title, is an open call that is still worth humanist scholars' attention.

The complex elements of the discussion of the digitization/digitized *vs* digitalization/digitalized divide indicates that it is surely time for DHers to pay attention to this binomial expression, so successfully deployed in business or economy that a publisher can get it in a title of collected essays that does not contain the word *digitalization* at all. It is time to form an understanding of *digitalization* that still denounces "cybercrud" when needed, or helps us to pay attention to "the confrontations over issues of privacy and control resulted in different ways of organizing access and information management" (van der Velden, 2007, 89). To express it in an electronic vocabulary, Brennen and Kreiss present a "path of least resistance" to the definition of digitalization, according to the path describing the third potential state of an electronic circuit (open, closed, or not working), because electricity follows the "path of least resistance."[9] But it is a core skill of the humanities to renounce the paths of least resistance and to wrestle with words, concepts, and realities. In that perspective, the last Section will develop some tracks to further the debate.

4 The effect of the "digitalization" perspective

The binomial expression "digitization" *versus* "digitalization" enters in the international debate through the English language. Such a distinction does not exist in French, Italian, or German, for example. But the inquiry of this article demonstrates that it this concept is worthy of exploration in an effort to grasp what is at stake in an explicit way in the English language. It represents surely one further argument in favor of a multilingual approach to digital epistemology, like the one developed in *Digital writing, digital Scriptures* (Clivaz, 2019).

I firstly underline how striking it is that even in the few occurrences where humanist scholars consciously use the term "digitalization" (van der Velden, Fiormonte), it is not discussed *per se*: a blind point exists in the scholarly discussion apart of Brennen and Kreiss's article. After all, the first use of "digitalization" in relation to the computer sphere was by a programmer (Wachal, 1971), but nowadays its use in critical discussion is mainly found under the pen of scholars outside of humanities who make claims about the "essence" of "the 'digitalization of the consumer'" (Mäenpää–Korhonen, 2015, 90; quoted in Section 2). In light of this consumerist perspective, DH scholars are generally confident in the traditional critical impact of their methodologies and knowledge. Alan Liu, for example, writes that "the digital humanities serve as a shadow play for a future form of the humanities that wishes to include what contemporary society values about the digital without losing its soul to other domains of knowledge work that have gone digital to stake their claim to that society" (2013, 410). In the same line, the HERA 2017 call hopes that the humanities, when digitized, will be able "to deepen the theoretical and empirical cultural understanding of public spaces in a European context."[10]

But it could secondly be argued that the blind point of the absent discussion about digitization/digitalization demonstrates an overconfidence of the digital humanities in its capacity to not lose the soul of the humanities in digital networks. Other voices are indeed more sensitive to the limitations imposed on humanities research

[9] See "Path of Leaf Resistance", *Wikipedia*, https://en.m.wikipedia.org/wiki/Path_of_least_resistance

[10] See "HERA Public Spaces", 31.08.17, http://heranet.info/2017/08/31/hera-launches-its-fourth-joint-research-programme-public-spaces/

by digital constraints, as we have seen with Maja van der Velden: even if she uses the word "digitalization" without discussing it, her article clearly points to digital control issues in the practice of building a database or a virtual research environment. From a more general and theoretical point of view, James Smithies strongly underlines in his book *The Digital Humanities and the Digital Modern* the same issues, even if the word *digitalization* is totally absent in it. He suggests that "our digital infrastructure […] has grown opaque and has extended into areas well outside scholarly or even governmental control" (2017, 11). His discourse becomes overtly political when he affirms the existence of a "point of entanglement between the humanities and neoliberalism, implicating digital humanists and their critics in equal measure" (2017, 218).

We are probably reaching here the main root of the silence about the digitization/digitalization challenge in DH debates: this binomial expression points to the political dimension of the digital revolution in humanities, to its economic and institutional implications, something that we prefer to let aside, consciously or unconsciously. This fear is also described by Wachal: "The humanist's fears are not entirely without foundation, and in any case, as a humane man he naturally fears the digitalization of the society" (1971, 30; quoted in Section 3). Listening to Wachal, and almost fifty years later to Smithies, can begin to lead us beyond the "path of leaf resistance" of Brennen and Kreiss. We should consider digitalization rather as the top of a mountain: it can be reached only through the *via ferrata* of the debates about cultural and multilingual diversity, about multiple source codes and standards, a multiplicity that preserves, at the end, diversity in human-computing knowledge productions.

Moreover, we are probably reaching right now the start of the DH awareness of this linguistic debate. As I end this article, I have opened the debate in the list Humanist Discussion Group and Simon Tanner has signaled his interest in the point, referring to Brennen and Kreiss' definition: "I have found the difference to be significant enough to seek to define it for my current book and in the past it has been a source of confusion or conflation that has not been helpful. I make it very clear to our students in the Masters of Digital Humanities or the MA Digital Asset and Media Management that they should not use the interchangeably" (Tanner, 2019).

Third, since the binomial expression digitization/digitalization is a vehicle for its own impact and meaning within the DH epistemology, is it possible to tie these concepts to the general challenge raised by the AIUCD 2020 call for papers? Notably, this discussion raises the following questions: "is it still necessary to talk about (and make) a distinction between 'traditional' humanists and 'digital' humanists? Is the term 'Digital Humanities' still appropriate or should it be replaced with 'Computational Humanities' or 'Humanities Computing'? Is the computational dimension of the research projects typically presented at AIUCD conferences that methodologically distinctive?"[11] At the root of these problems stands of course an important debate in Italian speaking DH, present in the name itself of the national DH organization, the AIUCD. This name mentions "Humanities Computing" (*informatica umanistica*) and "digital culture" (*cultura digitale*): *AIUCD - Associazione per l'Informatica Umanistica e la Cultura Digitale*.[12] But beyond this specific Italian perspective, the importance of collaboration between DHers and other humanist scholars concerns all of us.

The dialectic between *Humanities Computing* and *Digital Humanities* will in all cases remain in the historical memory of the DH development. But I am personally not convinced that a "step back" in the form of a return to *Humanities Computing*, motivated by a desire to keep all the humanists together under the banner of the *informatica umanistica*, is viable. Why? When the *Harvard Magazine* published in 2012 one of its first articles about the digital humanities, it was entitled "Humanities Digitized" (Shaw, 2012). It has always been meaningful for me to think in that direction. As I have argued elsewhere in detail, we could "begin to speak about the *digitized humanities*, or simply about *humanities* again, instead of *digital humanities*. Such an evolution might occur, if one looks at the evolution of the expression 'digital computer' which was in common usage during the fifties, but it has been now replaced by the single latter word 'computer' (Williams, 1984, 310; Dennhardt, 2016). When humanities finally become almost entirely digitized, perhaps it is safe to bet that we will once again speak simply about *humanities* in English or about *humanités* in French, thus making this outmoded word again meaningful through the process of cultural digitization" (Clivaz, 2019, 85–86).

According to this perspective, the debate between "humanities digitized" or "humanities digitalized", with all its cultural, economic, material, institutional and political dimensions, could signal a third step after *Humanities Computing* and *Digital Humanities*. This third step would stand at the crossroads where all humanists could meet up again, in an academic world definitively digitized, but hopefully not totally digitalized. It is up to all of us to decide if, in the third millennium, Humanities will be digitized or digitalized.

[11]See "Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale. Call for papers", https://aiucd2020.unicatt.it/aiucd-call-for-papers-1683.

[12] See AIUCD, www.aiucd.it.

References

Aurélien Berra. 2012. Faire des Humanités Numériques. *Read/Write Book* 2. Pierre Mounier (ed.). Paris, OpenEdition Press, 25–43. http://books.openedition.org/oep/238

J. Scott Brennen and Daniel Kreiss. 2016. *Digitalization. International Encyclopedia of Communication Theory and Philosophy* 23 October: 1–11. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118766804.wbiect111

Fabio Ciotti. 2019. Oltre la galassia delle Digital Humanities : per la costituzione di una disciplina di Informatica Umanistica. AIUCD2019. Book of Abstracts. Teaching and Research in Digital Humanities' Era. Stefano Allegrezza (ed). Udine, AIUCD 2019, 52–56. http://aiucd2019.uniud.it/wp-content/uploads/2019/01/BoA-2019_PROVV.pdf

Claire Clivaz. 2017. Lost in Translation? The Odyssey of "digital humanities" in French. *Studia UBB Digitalia* 62/1:26-41. https://digihubb.centre.ubbcluj.ro/journal/index.php/digitalia/article/view/4

Claire Clivaz. 2019. *Ecritures digitales. Digital writing, digital Scriptures*. DBS 4. Leiden, Brill.

Jari Collin. 2015. Digitalization and Dualistic IT. *IT Leadership in Transition. The Impact of Digitalization on Finnish Organizations.* Science + Technology 7. Jari Collin, Kari Hiekkanen, Janne J. Korhonen, Marco Halén, Timo Itälä, and Mika Helenius (eds.). Helsinki, Aalto University Publication Series, 29–34.

Franc Cormerais and Jacques Gilbert. 2016. Introduction. *Le texte à venir. Etudes Digitales* 1:11–16.

Robert Dennhardt. 2016. *The Term Digital Computer (Stibitz 1942) and the Flip-Flop (Turner 1920)*. München, Grin Verlag.

Milad Doueihi. 2014. Préface: quête et enquête. *Le temps des humanités digitales*. Olivier Le Deuff (dir.). Limoges, FyP editions, 7–10.

Amy A. Eahart and Toniesha L. Taylor. 2016. Pedagogies of Race: Digital Humanities in the Age of Ferguson. *Debates in the Digital Humanities*, volume 2. Matthew K. Gold and Lauren F. Klein (eds.). https://dhdebates.gc.cuny.edu/read/untitled/section/58ca5d2e-da4b-41cf-abd2-d8f2a68d2914-ch21

Rudy Ercek, Didier Viviers and Nadine Warzée. 2010. 3D Reconstruction and Digitalization of an Archaeological Site, Itanos, Crete. *Virtual Archaeology Revie* volume 1/1:81–85. DOI: 10.4995/var.2010.4794

Domenico Fiormonte. 2016. Toward a Cultural Critique of Digital Humanities. *Debates in the Digital Humanities*, volume 2. Matthew K. Gold and Lauren F. Klein (eds.). https://dhdebates.gc.cuny.edu/read/untitled/section/5cac8409-e521-4349-ab03-f341a5359a34-ch35

Steven E. Jones. 2016. The Emergence of the Digital Humanities (the Network Is Everting). *Debates in the Digital Humanities,* volume 2. Matthew K. Gold and Lauren Klein (ed.). Minneapolis, University Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/52

Alan Liu. 2013. The Meaning of the Digital Humanities. PMLA 128/2:409-423. https://doi.org/10.1632/pmla.2013.128.2.409

Willard McCarty (ed.). 2010. *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behavious, Products and Institutions*. Cambridge, OpenBook Publishers.

Raimo Mäenpää and Janne J. Korhonen. 2015. Digitalization in Retail: The Impact on Competition. *IT Leadership in Transition. The Impact of Digitalization on Finnish Organizations.* Science + Technology 7. Jari Collin, Kari Hiekkanen, Janne J. Korhonen, Marco Halén, Timo Itälä, and Mika Helenius (eds.). Helsinki, Aalto University Publication Series, 89–102.

Matthew G. Kirschenbaum. 2010. What Is Digital Humanities and What's It Doing in English Departments? *ADE Bulletin* 150:55–61. http://mkirschenbaum.files.wordpress.com/2011/03/ade-final.pdf

Olivier Le Deuff. 2016. Humanités digitales *versus* humanités numériques, les raisons d'un choix. *Le texte à venir. Etudes Digitales* 1: 263–264.

Päivi Parvianien, Jukka Kääriäinen, Maarit Tihinen, and Susanna Teppola. 2017. Tackling the digitalization challenge: how to benefit from digitalization in practice. *International Journal of Information Systems and Project Management* 5/1: 63–76. DOI: 10.12821/ijispm050104.

Toni Ryynänen and Torsti Hyyryläinen. 2018. Digitalisation of Consumption and Digital Humanities - Development Trajectories and Challenges for the Future. *CEUR Workshop Proceedings* Vol-2084: 1–8. http://ceur-ws.org/Vol-2084/short11.pdf

Amelia Sanz. 2013. Digital Humanities or Hypercolonial Studies? *E-Prints Complutense*. Madrid, University of Madrid. https://eprints.ucm.es/50610/

Jonathan Shaw. 2012. Humanities Digitized. Reconceiving the study of culture. *Harvard Magazine* May–June: 40–44 and 73–75. http://harvardmag.com/pdf/2012/05-pdfs/0512-40.pdf

James Smithies. 2017. *The Digital Humanities and the Digital Modern*. Basingstoke, Palgrave Macmillan.

Bernard Stiegler. 2016. *Dans la disruption. Comment ne pas devenir fou?* Paris, Les liens qui libèrent.

Simon Tanner. 2019. *Delivering impact with digital resources: Planning strategy in the attention economy*. London, Facet Publishing. Forthcoming. Quotation in Simon Tanner. 2019. Digitization vs digitalization. *Humanist Discussion Group* 33.390/4. https://dhhumanist.org/volume/33/392/

Maja van der Velden. 2007. Invisibility and the Ethics of the Digitalization: Designing so as not to Hurt Others. *Information Technology Ethics: Cultural Perspectives.* Sonja Hongladarom and Charles Ess (eds.). Hershey et al., Idea Group Reference Ed., 81–93.

Robert Wachal. 1971. Humanities and Computers: A Personal View. *The North American Review* 256/1:30–33. https://www.jstor.org/stable/25117163

Bernard O. Williams. 1984. *Computing with Electricity, 1935–1945*. PhD Dissertation. Lawrence, University of Kansas.

La geolinguistica digitale e le sfide lessicografiche nell'era delle *digital humanities*: l'esempio di VerbaAlpina

Beatrice Colcuc
Ludwig-Maximilians-Universität München
beatrice.colcuc@romanistik.uni-muenchen.de

Abstract

English. The increasing use of new technologies and the possibilities they offer have led to a change in the way in which research and data processing are conceived. Collaboration between projects and data exchange are modern practices, the rules of which have been summarized in an acronym (FAIR) containing the four basic principles of digital research. The project VerbaAlpina of Munich University investigates the idioms of the Alpine area and, since the beginning, has been carried out according to these principles. However, in order to act according to the recommendations and pursue its research objective, VerbaAlpina, as a multilingual project, had to match the lexical-semantic variation with the universality of concepts. To establish concepts in a universal way is a difficult task because, to represent a concept, it is nevertheless always necessary to use a certain language. For this reason, VerbaAlpina has started to apply the procedures provided by external projects such as Wikidata or the resources of the German National Library for language-independent labelling of concepts on the one hand and linguistic forms on the other. Thus, the paper reflects the challenges of modern lexicography and the possibilities of overcoming problems in the digital era exemplified by the project VerbaAlpina.

Italiano. L'utilizzo delle nuove tecnologie e le possibilità da esse offerte hanno condotto a un mutamento nel modo di concepire la ricerca e il trattamento dei dati. La collaborazione tra i progetti e lo scambio di dati sono pratiche, le cui regole sono state riassunte nell'acronimo FAIR indicante i quattro princìpi fondamentali della ricerca digitale. Il progetto VerbaAlpina dell'università di Monaco di Baviera si occupa dello studio degli idiomi dell'area alpina e, fin dalla sua concezione, è stato portato avanti secondo tali princìpi. Tuttavia, per poter agire secondo le raccomandazioni e perseguire il proprio obiettivo di ricerca, VerbaAlpina si è dovuto confrontare con la problematica lessicale-semantica relativa alla mancanza di universalità dei concetti. Fissare dei concetti in maniera universale è un arduo compito perché la riproduzione di una data idea si effettua sempre attraverso una determinata lingua. Per questo motivo, VerbaAlpina ha iniziato ad applicare procedure fornite da progetti esterni quali Wikidata oppure le risorse della Biblioteca Nazionale Tedesca per l'etichettatura dei dati indipendente dalla lingua. L'articolo vuole presentare una riflessione sulle sfide della lessicografia moderna e sulle possibilità di superamento delle problematiche nell'era delle digital humanities presentando l'esempio concreto del progetto VerbaAlpina.

1 Trattare i dati (lessicografici) nell'era delle *digital humanities*

La ricerca dell'era pre-digitale è stata contraddistinta da modalità di concezione progettuale fortemente individuali: la raccolta, l'analisi e l'illustrazione dei dati venivano effettuate da persone (o gruppi di persone) che operavano singolarmente. La comunicazione scientifica si compiva attraverso le pubblicazioni dei libri cartacei, i quali venivano conservati in luoghi circoscritti come ad esempio le biblioteche, e ciascuno studio rappresentava un progetto concluso in sé (cfr. Krefeld, 2016). Inoltre, molti dati rimanevano nelle mani degli stessi ricercatori che li avevano raccolti e, in generale, la comunità scientifica non compiva molti sforzi per mettere a disposizione del grande pubblico i dati provenienti dai diversi progetti scientifici.

Ormai, l'avvento delle nuove tecnologie non è più un fenomeno di recente datazione. Il passaggio dal cartaceo al digitale è avvenuto in maniera sempre più repentina, contribuendo al mutamento dell'approccio scientifico nei confronti delle modalità di ricerca. La rivoluzione digitale ha rinsaldato l'idea della condivisione e, allo stesso tempo, provveduto a fornire gli strumenti necessari per favorire e facilitare l'interscambio di dati come pure, più in generale, l'interazione tra diversi progetti. Ciononostante, la creazione di una rete di progetti e la condivisione dei relativi dati non sono fornite da una mera digitalizzazione degli strumenti di ricerca. La collaborazione rappresenta l'essenza primaria del fare scienza, poiché è sulla base delle conoscenze già presenti che si costruisce il progresso. Per operare in maniera collettiva nell'era digitale è però necessario che i dati di ricerca soddisfino alcuni requisiti fondamentali. In primo luogo i dati devono essere strutturati, descritti

ed eventualmente etichettati in maniera tale da prestarsi a essere maneggiati in sedi esterne al loro progetto originario e devono poter continuare a essere accessibili anche in un momento successivo all'eventuale chiusura del progetto. In questo contesto si inserisce l'idea di Web Semantico (e successivamente dei Linked Open Data), nata con l'obiettivo di rimodellare l'ambiente virtuale di internet. Numerosi sono i progetti che si collocano all'interno di questo pensiero: essi formano la cosiddetta *Linked Open Data Cloud* (cfr. Cyganiak e Jentzsch, 2007 -) e costituiscono, al contempo, una comunità interconnessa attraverso relazioni basate sui loro insiemi di dati (cfr. Bizer et al. 2009, 154).

Le esigenze di condivisione e connessione dei dati sul web sono inoltre state formulate in maniera esplicita nel 2016, quando un grande numero di ricercatori provenienti da diversi Paesi ha pubblicato le linee guida per la gestione moderna dei dati di ricerca (cfr. Wilkinson, Dumontier et al., 2016). Tali raccomandazioni sono state racchiuse nell'acronimo FAIR, una sigla che raccoglie i quattro princìpi fondamentali sui quali dovrebbero essere basate la comunicazione e la cooperazione scientifiche nell'era digitale. Secondo tali princìpi, i dati della ricerca dovrebbero essere rintracciabili (*findable*), accessibili (*accessible*), interoperabili (*interoperable*) e riutilizzabili (*reusable*). Per essere rintracciabili, i progetti ai quali i dati appartengono, devono essere reperibili attraverso portali centrali quali, ad esempio, i cataloghi delle biblioteche. I dati di ricerca che non sono soggetti ad alcuna restrizione giuridica (come potrebbero essere ad esempio i dati strettamente personali) devono essere messi a disposizione del grande pubblico e rinunciare, di conseguenza, al diritto d'autore. Al fine di poter essere interoperabili, inoltre, i dati devono essere innanzitutto scissi, successivamente strutturati ed essere descritti in maniera precisa. Il riutilizzo dei dati si rende infine possibile attraverso una corretta applicazione dei tre princìpi precedenti: si tratta di princìpi estensibili che non si lasciano mettere in contrapposizione l'un l'altro (cfr. Force11, 2011-2017; GoFair, 2011-2017; Lücke, 2018).

Le possibilità offerte dalla digitalizzazione in termini tecnici hanno altresì consentito di valutare da un nuovo punto di vista una delle questioni storiche relative al trattamento dei dati lessicografici. In linea di massima, le opere lessicografiche possono essere strutturate in maniera semasiologica (le parole di un dato idioma sono elencate seguite dal loro significato) oppure onomasiologica (si descrivono i significati e vi si collegano le diverse parole che ad essi conducono). Nell'era analogica tali opere erano strutturate secondo uno o secondo l'altro modo: in ambito romanzo, si ricordino, tra gli altri, l'atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale (AIS), di stampo onomasiologico, oppure, il *Dicziunari Rumantsch Grischun* (DRG), strutturato, invece, in maniera semasiologica. Una concezione in entrambi i sensi non era possibile a causa di limiti pratici imposti dalle modalità di pubblicazione del passato, mentre oggigiorno, le possibilità fornite dalla digitalizzazione offrono nuovi strumenti e la concezione di un'opera lessicografica che vada in due direzioni è realizzabile. Ciononostante, benché l'aspetto tecnico non rappresenti più alcun problema alla messa in pratica di tale approccio bidirezionale, le sfide si aprono soprattutto dal punto di vista contenutistico, come si evincerà dai capitoli che seguiranno.

La compilazione digitale di opere lessicografiche quali i dizionari sembra essere oggi un'attività relativamente consolidata. Non sono rari i dizionari che offrono la possibilità di essere consultati in rete, anche se, alcuni di essi, non si sono mai realmente distanziati da una concezione cartacea. Tali opere, presentano ancora un grande margine di sviluppo e ampliamento per potersi definire digitalizzate in modo interoperabile secondo i gradi definiti in Lücke (2016). A titolo illustrativo, ma non esaustivo, si pensi alla versione online del *Romanisches Etymologisches Wörterbuch* (Meyer-Lübke, 1935): il contenuto dell'opera è messo a disposizione online in formato PDF, ma, ai sensi della digitalizzazione, in essa potrebbero essere implementate diverse funzioni, tra cui, ad esempio, la ricerca di singoli lemmi. Lo scopo di una lessicografia basata sul web non si limita alla mera presentazione virtuale di una determinata opera. Lo sviluppo digitale è bensì rappresentato da un più ampio tentativo di costituzione di reti lessicali e semantiche messe in relazione tra di loro. Progetti volti a fornire tali interconnessioni sono stati iniziati già verso la metà degli anni Ottanta. A titolo esemplificativo si pensi a *WordNet*, il database lessicale per la lingua inglese, ma anche a *EuroWordNet* nato negli anni Novanta come rete semantica per le lingue europee (cfr. Fellbaum 2006, 665, 669). Tali progetti costituiscono un primo tentativo di strutturare il materiale in maniera semantica e non solo lessicale come si è soliti fare nei dizionari cartacei e, soprattutto, si inseriscono nel panorama dei lavori relativi all'elaborazione del linguaggio naturale multilingue. In tempi più recenti si colloca la concezione di *BabelNet* (https://babelnet.org/), una rete semantica multilingue, automatizzata e di ampia copertura, ovvero un dizionario enciclopedico costituito unendo il contenuto lessicale di *WordNet* al sapere enciclopedico di *Wikipedia* attraverso processi automatizzati di integrazione dei contenuti di ambedue i database (cfr. Navigli e Ponzetto, 2010, 216).

2 *VerbaAlpina*: geolinguistica e lessicografia digitali

Mentre la digitalizzazione dei primi dizionari e corpora lessicali si colloca tra gli anni Ottanta e gli anni Novanta (cfr. Chiari 2012, 98), più recente risulta invece essere il passaggio dal cartaceo al digitale per quanto riguarda gli atlanti linguistici. Relativamente all'area alpino-romanza, diversi atlanti sono oggi disponibili in rete in formati PDF o JPG ma non hanno percorso tutte le tappe del passaggio dal cartaceo al digitale (cfr. Lücke, 2016; Knapp, 2017). È, a titolo esemplificativo, il caso di *NavigAIS* (Tisato, 2009-2018) all'interno del quale, nonostante la sua presenza su internet sia lodevole, potrebbero essere implementate diverse funzionalità, tra le quali ad esempio la rintracciabilità di forme attestate oppure una visualizzazione quantificata dei dati, la possibilità di consultare singoli gruppi di dati linguistici in prospettiva onomasiologica o semasiologica, come pure l'esportazione dei dati.
Alle esigenze della ricerca lessicografica e atlantistica in chiave moderna, cerca di dare una risposta il progetto *VerbaAlpina* dell'Università Ludwig-Maximilian di Monaco di Baviera, nato nel 2014 con l'intento di indagare lo spazio linguistico delle Alpi nella sua storica unità linguistico-culturale (cfr. Krefeld e Lücke, 2014 -). Fin dalla sua concezione, completamente digitale e pensata non solo sul web, ma per il web, il progetto VerbaAlpina ha operato nel pieno rispetto dei princìpi FAIR (che sarebbero stati formulati solamente due anni dopo) e promosso un'idea innovativa di lessicografia e atlantistica linguistica (cfr. Krefeld, 2018). Oltre all'aspetto linguistico, una parte consistente del progetto è specificatamente dedicata alla creazione di strumenti per la gestione dei dati di ricerca nei progetti digitali e pensati per il web.

2.1 Concezione e presentazione del progetto

Nucleo centrale dell'attività di VerbaAlpina è la raccolta strutturata di una precisa cornice semasiologica e onomasiologica, costituita dagli ambiti terminologici alpini, alla quale è possibile accedere attraverso una cartina interattiva. La ricerca prende in esame il lessico dialettale degli idiomi alpini, in modo particolare le parole relative agli ambiti della natura (flora, fauna, formazioni paesaggistiche), della cultura alpina storica (lavorazione del latte) e di quella corrente (turismo). I dati raccolti e analizzati da VerbaAlpina sono puramente dialettali, mentre i termini relativi alle lingue standard non sono presi in considerazione. Diversi sono gli scopi perseguiti da VerbaAlpina: in primo luogo il progetto intende documentare e analizzare in prospettiva linguistica e storico-etimologica la regione alpina, uno spazio fortemente frammentato per quanto riguarda le lingue e i dialetti ivi parlati. I confini dell'area di ricerca sono definiti dalla Convenzione delle Alpi (http://www.alpconv.org/), un trattato tra i Paesi del territorio alpino atto a promuovere e sviluppare questa area montana in diversi ambiti (cfr. Lücke 2018a).

I dati sono forniti dagli atlanti linguistici e dai dizionari relativi all'area di ricerca, analogici o digitali, pubblicati nel corso del tempo. In un primo momento, il materiale linguistico georeferenziato proveniente dalle fonti affronta un percorso di digitalizzazione attraverso un sistema di trascrizione basato esclusivamente sui caratteri ASCII (cfr. Krefeld e Lücke, 2016). In un secondo momento, il materiale trascritto viene sottoposto a una *tokenizzazione*, un processo che separa in singoli *token* (parole) il materiale trascritto in un momento precedente. L'interesse principale del progetto si esplica nella presentazione dei punti di coesione tra i diversi idiomi e le diverse famiglie linguistiche presenti sul territorio alpino soprattutto in prospettiva lessicologica. Per l'adempimento di tale scopo, il materiale linguistico viene raggruppato in tipi di base, ossia secondo la radice lessicale comune a diverse attestazioni che possono appartenere anche a diverse famiglie linguistiche[1] e in tipi morfolessicali, vale a dire in forme di un solo tipo di base, appartenenti a un'unica famiglia linguistica che presentano caratteristiche grammaticali comuni quali la parte del discorso, il genere e gli elementi di formazione delle parole (cfr. Krefeld e Lücke, 2016a). Ad esempio, il tipo di base latino *CASEU(M) 'formaggio' è presente sia in area linguistica germanica, sia in area romanza nelle forme deu. *Käse* e ita. *cacio*, i quali, a loro volta, rappresentano due tipi morfolessicali differenti.

I dati linguistici storici rilevati dagli atlanti e dai dizionari sono completati attraverso una piattaforma di crowdsourcing sviluppata all'interno del progetto (https://www.verba-alpina.gwi.uni-muenchen.de/en/?page_id=1741&db=191). La piattaforma si rivolge direttamente ai parlanti dei dialetti delle Alpi al fine di raccogliere materiale linguistico attuale e poter osservare lo spazio alpino anche in prospettiva diacronica. Una volta aperta la pagina, viene chiesto agli utenti di scegliere una lingua di navigazione tra quelle proposte (francese, italiano, sloveno, tedesco). Successivamente, vengono mostrate le istruzioni per l'utilizzo

[1]In molti casi non è dato sapere se la radice lessicale comune a diverse parole sia da ricollegare allo stesso sostrato linguistico oppure a un contatto linguistico più recente. Per questo motivo VerbaAlpina utilizza il termine "tipo di base", in quanto "etimo" si riferisce generalmente allo strato linguistico immediatamente precedente (cfr. Krefeld e Lücke 2016a).

della piattaforma e gli utenti sono invitati a inserire l'idioma alpino di cui essi sono i parlanti. Nel caso in cui un idioma non sia presente nella lista, gli utenti hanno la possibilità di segnalarlo direttamente alla redazione che provvederà a inserirlo. Innanzitutto, gli utenti sono chiamati a inserire il nome del comune di cui padroneggiano l'idioma. Cliccando sull'apposito campo "concetto", appare una lista con tutti i concetti esistenti nella banca dati di VerbaAlpina. Da qui, gli utenti possono scegliere per quali concetti inviare parole. I dati raccolti attraverso il crowdsourcing vengono trattati alla pari dei dati provenienti dai dizionari e dagli atlanti, con l'unica differenza che non sono sottoposti al processo di trascrizione. Per questi dati, la tokenizzazione avviene solamente qualora si tratti di un sintagma costituito da più elementi. La tipizzazione di queste parole avviene alla stregua dei dati raccolti dai dizionari e dagli atlanti linguistici. A livello di database, le singole attestazioni provenienti dal crowdsourcing ricevono un identificatore e sono collegate ai concetti di cui rappresentano le diverse forme dialettali. Successivamente al trattamento strutturato, i dati analizzati da VerbaAlpina possono essere visualizzati sulla cartina interattiva (https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=27&db=191). Tramite l'utilizzo di filtri appropriati, i dati sono accessibili in prospettiva onomasiologica (si rappresentano tutte le attestazioni linguistiche collegate a un determinato concetto) oppure semasiologica (si rappresentano i concetti legati a un preciso tipo morfolessicale). Inoltre, la visualizzazione può essere impostata in modalità qualitativa, attraverso la quale è possibile evincere la distribuzione generale delle attestazioni linguistiche, oppure quantitativa, ossia indicante il numero di dati all'interno di una certa area. I dati possono essere visualizzati in prospettiva geografico-fisica oppure astratta: la prima mostra i dati distribuiti su una cartina fisica, attraverso la seconda, invece, i dati sono presentati su una mappa a nido d'ape.

Parallelamente all'attività linguistica, il progetto ha profuso un grande impegno nella gestione dei dati digitali con l'obiettivo ultimo di promuovere la sostenibilità e la durabilità del progetto anche dopo la sua chiusura definitiva. La descrizione di una parte delle attività che sono state intraprese in questo senso avviene nel corso del presente contributo. VerbaAlpina si impegna a utilizzare strumenti tecnologici adatti al web e applicabili al pensiero open source, come ad esempio *Wordpress* per la piattaforma centrale, *Leaflet* per la cartina interattiva e le banche dati relazionali *MySQL*. Dato che VerbaAlpina intende altresì fungere da creatore di nessi tra istituzioni e progetti già esistenti al fine di interscambiare, integrare e completare i dati linguistici riguardanti l'area alpina, per i diversi partner sono messe a disposizione banche dati all'interno delle quali i progetti cooperanti possono inserire i loro dati e collegarli così a quelli di VerbaAlpina.

2.2 Status quo

L'area di ricerca di VerbaAlpina prende in considerazione tutti gli idiomi parlati nell'arco alpino. Si tratta di una superficie di 190.000 km^2 comprendente alcune regioni di sei Paesi diversi (Austria, Francia, Germania, Italia, Slovenia e Svizzera) e due interi stati (Liechtenstein e Montecarlo) per un totale di quasi 6000 comuni, i quali rappresentano per VerbaAlpina le unità di georeferenziazione. Considerato che in linguistica l'unanimità di opinioni su una definizione unitaria di dialetto rappresenta ancora una visione remota (e, a dire il vero, di scarso interesse per la disciplina stessa), non è possibile indicare un numero, nemmeno approssimativo, di varietà alpine locali parlate dalla Francia alla Slovenia. Pur concedendo grande importanza all'aspetto locale delle varietà, l'analisi linguistica di VerbaAlpina si eleva al livello delle tre famiglie linguistiche che occupano il territorio alpino (romanza, germanica e slava). Per questo motivo, attraverso il processo di tipizzazione di cui sopra, il variegato materiale linguistico locale è raggruppato in tipi morfolessicali etichettati rispettivamente con le sigle ISO 639-5 relative alle famiglie linguistiche: *roa*. per romanzo, *gem*. per germanico e *sla*. per slavo (cfr. ISO 639-5). La base di conoscenza di VerbaAlpina racchiude ad oggi (novembre 2019) 55.407 stimoli (si tratta solitamente dei titoli delle carte degli atlanti di riferimento) ai quali sono collegate 165.521 attestazioni linguistiche, distribuite tra 3.989 concetti e riassunte in 9.556 tipi morfolessicali. Per quanto riguarda i dati provenienti dal crowdsourcing, si contano 1.065 informanti diversi e 15.249 parole totali inviate dagli utenti.

3 La lessicografia tradizionale e le sfide per il futuro nell'era delle *digital humanities*

Come è stato già accennato, solitamente, i dizionari classici sono strutturati in maniera semasiologica, ovvero il lessico ivi contenuto viene elencato partendo dall'unità lessicale (parola) alla quale sono collegati i diversi significati. La fortuna di questo modello di dizionario è da ricondurre essenzialmente a due ragioni: da un lato tali opere lessicografiche hanno il compito di raccogliere e illustrare il lessico appartenente a un dato idioma (e in questo senso fungono da ausili per la documentazione di una lingua); dall'altro lato, concepire un'opera

lessicografica in prospettiva semasiologica risulta essere un'operazione di più facile realizzazione. Per riprodurre una serie ordinata di segni che compongono un'unità lessicale si può contare su un sistema codificato e standardizzato di caratteri: la scrittura stessa. La difficoltà di creare un'opera lessicografica partendo dal contenuto concettuale (prospettiva onomasiologica) è invece molto più estesa. Il contenuto semantico dell'unità lessicale non può essere delineato così facilmente, né tantomeno può essere standardizzato. L'utilizzo di una determinata espressione, non predice nulla sulle caratteristiche intrinseche del concetto al quale è collegata la parola che si è cercato di riprodurre. Fondamentalmente, sia per quanto riguarda la riproduzione di una singola parola, sia per quanto riguarda la descrizione di un significato, si fa appello alla scrittura, ovvero, alla lingua. Tuttavia, tale ricorso alla lingua è problematico giacché è possibile utilizzare solamente un determinato idioma alla volta, mentre invece, alla luce di quanto detto poc'anzi e nell'ottica di una scienza interconnessa, sarebbe opportuno potersi riferire ai concetti indipendentemente dalla singola lingua. Il mero ricorso ai codici linguistici ostacola inoltre la condivisione dei dati e la loro connessione ad altri database, limitando in parte una più ampia collaborazione tra progetti scientifici.

4 L'approccio al problema sull'esempio di VerbaAlpina

Per l'accorpamento dei contenuti provenienti dai vari atlanti linguistici e dai dizionari, anche il progetto VerbaAlpina si è dovuto misurare con la suddetta questione. All'interno della banca dati relazionale che funge da base del progetto, è stata creata una tabella che racchiude i concetti (si tratta prevalentemente dei contenuti tematici delle mappe linguistiche). Le singole mappe degli atlanti sono quindi collegate al concetto appropriato corrispondente. Per quanto riguarda il contenuto semantico di un concetto, il minimo comune denominatore è rappresentato dall'insieme delle informazioni che compongono il dato concetto. Dal momento che VerbaAlpina tratta dati provenienti da diverse fonti esterne al progetto, la gestione e l'uniformizzazione degli stessi può essere concepita solamente attraverso una descrizione accurata dei dati che, nell'insieme, formano un singolo concetto. Tuttavia, il metodo di gestione appena descritto consente solo la comparabilità all'interno del progetto, mentre per collegare tra di loro diversi gruppi di dati, sarebbe auspicabile e necessaria una soluzione globale e indipendente dalla lingua.

La sfida della standardizzazione è stata intrapresa da lungo tempo all'interno delle biblioteche, dove l'esigenza dell'uniformità mira a creare uno standard ad esempio per la gestione dei dati relativi agli autori delle diverse pubblicazioni o per la realizzazione di differenti indicizzazioni tematiche. È da questa esigenza dell'ambito bibliotecario che nasce l'idea del cosiddetto *authority control*, ossia un sistema normato per la costituzione di un archivio (*authority file*) che possa contenere dati organizzati secondo uno stesso modello. In Germania, a partire dagli anni Ottanta del secolo scorso sono stati creati diversi sistemi per la standardizzazione dei dati relativi a persone (PND: *Personennamendatei*), enti (GKD: *Gemeinsame Körperschaftsdatei*) e voci (SWD: *Schlagwortdatei*). Inoltre, tra il 2009 e il 2013, la Biblioteca Nazionale Tedesca (*Deutsche Nationalbibliothek*) e altre associazioni bibliotecarie di lingua tedesca hanno intrapreso un'iniziativa volta a creare il cosiddetto GND (*Gemeinsame Normdatei*), un sistema di controllo di autorità che riassume tutti gli elenchi sopraccitati in un unico file. Oltre alle tradizionali entità quali organismi o persone, il GND raccoglie anche concetti.

Anche il database enciclopedico *Wikidata*, creato allo scopo di supportare *Wikipedia*, funziona attraverso il controllo di autorità (cfr. Wikidata a). In Wikidata, ogni concetto è registrato tramite un numero identificatore (ID) e descritto nel dettaglio attraverso relazioni gerarchiche. La concezione partecipativa di Wikidata ha permesso l'inserimento nella piattaforma di un numero considerevole di entità referenziabili. A partire dal 2018, tutti i concetti di VerbaAlpina sono stati connessi ai Q-ID (o *Q-item*) di Wikidata. Tale connessione attraverso gli identificatori permette di collegare con altre banche dati esterne le informazioni altrimenti gestite solo all'interno del progetto. In questo senso, le risorse interne, come ad esempio il database multimediale, possono essere collegate in maniera decentralizzata ed essere messe a disposizione di diversi progetti. Questa concezione permette ai dati di essere costantemente arricchiti di informazioni aggiuntive. Allo stesso modo, è possibile pensare anche alla presentazione dei contenuti in diverse lingue, in quanto Wikidata mette a disposizione le traduzioni delle denominazioni relative ai diversi concetti (o ai relativi Q-ID). La collaborazione tra Wikidata e VerbaAlpina prevede non solo una connessione attraverso l'applicazione dell'identificatore, ma anche una partecipazione attiva di quest'ultimo al database enciclopedico. VerbaAlpina dispone infatti di un account proprio sulla pagina di Wikidata al fine di mappare eventuali concetti ivi mancanti. Al momento (novembre 2019), 1000 concetti di VerbaAlpina sono muniti di un corrispettivo Q-ID. Una parte consistente di concetti contenuti nella banca dati di VerbaAlpina deve essere ancora elaborata e ogni entrata

corredata del rispettivo Q-ID, un'attività che è in costante aggiornamento. Al momento, è stata data priorità all'applicazione degli identificatori ai concetti di VerbaAlpina, in un secondo momento avverrà anche la mappatura di concetti mancanti su Wikidata. L'inserimento dei Q-ID di Wikidata nella banca dati di VerbaAlpina avviene in maniera del tutto manuale.

Le proposte sopraccitate si riferiscono a un tentativo di standardizzazione del contenuto semantico di un'ampia quantità di concetti. Dal momento che VerbaAlpina non si occupa solamente del trattamento di materiale semantico, ma anche di tipi morfolessicali, la creazione di un controllo di autorità applicabile anche a tali forme linguistiche sarebbe auspicabile per l'identificazione univoca del contenuto lessicale. Per un'etichettatura di questo genere, la situazione si presenta in maniera diversa: il GND non fornisce ancora un sistema mirato per la gestione dei dati in questo senso, mentre Wikidata offre la possibilità di creare attestazioni lessicali contenenti le informazioni legate al lemma stesso, alla lingua e alla categoria lessicale. Ogni attestazione lessicale è correlata a un identificatore L (L-ID) che viene generato automaticamente (cfr. Wikidata b). Le singole voci possono anche essere completate con informazioni quali genere e significato. Per varietà linguistiche di estensione meno ampia quali ad esempio il ladino dolomitico oppure il friulano, Wikidata richiede non solo l'inserimento del lemma vero e proprio, ma anche di indicare la cosiddetta *spelling variant*, ossia l'ortografia utilizzata per rappresentare un determinato lemma indicata mediante un codice linguistico. Ad esempio, se si desidera inserire il lemma *paurùns* (forma ladina per il concetto 'siero di latte dopo la prima separazione della materia solida'), viene richiesto di indicare secondo quale ortografia è stato inserito il lemma stesso (cfr. Wikidata c). Inoltre, è possibile anche aggiungere le informazioni relative al numero e al caso linguistico e collegarle al rispettivo concetto. Quest'ultimo sistema di referenziazione consente di collegare ai singoli lemmi anche le informazioni riguardanti le derivazioni o l'etimologia, una pratica che faciliterebbe, di conseguenza, il lavoro lessicografico. Tale modalità di etichettatura del materiale linguistico è stata implementata da VerbaAlpina solo recentemente, ma sarà portata avanti in maniera progressiva con le stesse modalità applicate per i concetti.

La connessione tra i concetti di VerbaAlpina e quelli di Wikidata attraverso l'applicazione di un identificatore non è fine a se stessa, ma si inserisce in un'ottica di condivisione più ampia in grado trovare punti di aggancio anche con il progetto *GeRDI* (*Generic Research Data Infrastructure*). Quest'ultimo nasce nel 2016 come aggregatore di dati allo scopo di offrire a tutti i progetti di ricerca in Germania la possibilità di archiviare, condividere e riutilizzare dati. GeRDI impiega Wikidata come base di conoscenza, realizzando così un sistema di consultazione di dati interdisciplinare e multilingue (cfr. Mutter, 2018).

Wikidata rappresenta agli occhi di VerbaAlpina una piattaforma centrale attraverso la quale quest'ultimo può mettersi in relazione con altri progetti collegati analogamente alla stessa base di conoscenza. Un esempio potrebbe essere il dizionario enciclopedico BabelNet, anch'esso connesso a Wikidata. La connessione diretta tra VerbaAlpina e Babelnet sarebbe interessante per quanto riguarda l'identificazione dei lemmi, ma, quest'ultimo non dispone ancora di numeri univoci per questo tipo di materiale lessicale, né raccoglie dati dialettali, centrali invece per l'attività di VerbaAlpina. Ad ogni modo, benché non in maniera diretta, VerbaAlpina e Babelnet dispongono entrambe di Wikidata come progetto di collaborazione comune.

5 Prospettive e attività future

Facendo di nuovo riferimento ai princìpi FAIR menzionati all'inizio, organizzare e gestire i dati linguistici nel quadro dei sistemi di identificazione descritti poc'anzi, rappresenta un passo importante verso il rispetto di questi princìpi. I dati strutturati si presentano non solo più accessibili da parte di altri progetti e più facilmente reperibili grazie al collegamento in rete, ma la standardizzazione dei riferimenti esatti ne favorisce anche il trattamento dal punto di vista dell'interoperabilità. VerbaAlpina non è leale a questi princìpi solamente per quanto riguarda l'interoperabilità dei dati, ma anche relativamente alla loro rintracciabilità (*f*) attraverso l'inserimento del progetto nei cataloghi della biblioteca universitaria dell'università di Monaco di Baviera. VerbaAlpina sposa in toto l'idea di libero accesso alla conoscenza come bene comune e utilizza solamente licenze Creative-Commons (CC) rinunciando, di conseguenza, al diritto d'autore e permettendo l'utilizzo dei dati con la sola restrizione dell'obbligo di citazione (cfr. Lücke, 2016a); il riutilizzo dei dati è reso possibile attraverso la loro esportazione tramite un'interfaccia di programmazione di un'applicazione (eng. *Application Programming Interface*; API), la quale permette l'accesso all'intero dataset di VerbaAlpina. Una documentazione e spiegazione dettagliata è consultabile al seguente indirizzo: https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844&db=191. Attraverso tale API i dati di VerbaAlpina sono accessibili meccanicamente (*machine readable*) e possono essere scaricati, modificati ed elaborati ulteriormente. La connessione dei dati con altri dataset è garantita attraverso lo schema di metadati di DataCite

(https://datacite.org), la quale si trova ancora in fase di sviluppo. Nonostante l'accesso meccanico ai dati dall'esterno sia già possibile mediante l'API e la connessione dei dati con altri dataset attraverso i metadati, le procedure per inserire i dati nella nuvola dei Linguistic Linked Open Data (https://linguistic-lod.org/) sono in fase di avvio allo scopo di creare un'ulteriore connessione tra progetti e contribuire in questo senso all'idea di web strutturato. Operare nell'era delle *digital humanities* significa creare conoscenza interconnessa, condivisibile, accessibile, una conoscenza più ampia e coesa. Equivale a creare strumenti e a metterli a disposizione non solo della comunità scientifica, ma anche del grande pubblico. Si tratta di un'amplificazione dell'originale pensiero umanista: creare sapere, renderlo accessibile e diffonderlo affinché l'umanità possa accrescere le proprie conoscenze.

Bibliografia

AIS = Karl Jaberg e Jakob Jud. 1928-1940. *Sprach- und Sachatlas Italiens und der Südschweiz*. 8. vol. Riniger&Co, Zofingen.

API = VerbaAlpina. 2014 -. API Dokumentation. https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844&&db=191 [accesso 29/11/2019].

BabelNet = BabelNet | The largest multilingual encyclopedic dictionary and semantic network. https://babelnet.org/ [accesso 25/11/2019].

Cartina interattiva = VerbaAlpina. 2014 -. Cartina interattiva. https://www.verba-alpina.gwi.uni-muenchen.de/it/?page_id=27&db=191 [accesso 25/11/2019]

Christian Bizer, Tom Heath, e Tim Berners-Lee. 2009. Linked data - the story so far. *International Journal of Semantic Web and Information System* 5(3):1-22. doi:10.4018/jswis.2009081901 [accesso 12/11/2019].

Christiane Fellbaum. 2006. Wordnet and wordnets. In Keith Brown, ed. by., *Encyclopedia of Language and Linguistics*, Elsevier, Oxford: 665-670.

Christina Mutter. 2018. Wikidata. *VerbaAlpina-it 19/1* (creazione: 18/1), *Metodologia*. https://doi.org/10.5282/verbaalpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DW%23105

Convenzione delle Alpi. 1995-. https://www.alpconv.org/it/home/ [accesso 29/11/2019].

Crowdsourcing = VerbaAlpina. 2014 -. Crowdsourcing-Tool. https://www.verba-alpina.gwi.unimuenchen.de/en/?page_id=1741&db=191 [accesso 26/11/2019].

DataCite: https://datacite.org/index.html [accesso 28/11/2019].

DRG = Dicziunari Rumantsch Grischun. 1939-2013. Institut dal Dicziunari Rumantsch Grischun, Coira.

Force11, ed. by. 2011-2017. *The Fair Data Principles*. https://www.force11.org/group/fairgroup/fairprinciples [accesso 03/09/2019].

GeRDI = Generic Research Data Infrastructure. 2016 -. https://www.gerdi-project.eu/ [accesso 14/11/2019].

GoFair, ed. by. 2011-2017. *FAIR Principles*. https://www.go-fair.org/fair-principles. [accesso 03/09/2019].

Graziano Tisato, ed. by. 2009-2018. NavigAIS. AIS Digital Atlas and Navigation Software, Padova, Istituto di Scienze e Tecnologie della Cognizione (ISTC) - Consiglio Nazionale delle ricerche (CNR). Versione 1.47. http://www3.pd.istc.cnr.it/navigais/ [accesso 20/11/2019].

Isabella Chiari. 2012. Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto. *Bollettino di Italianistica*. Per Tullio De Mauro II, pp. 94-125.

ISO 639-5 = Library of the Congress Registration Authority. 2009. Codes for the representation of Names of Languages – Part 5: Alpha-3 code for language families and groups. https://www.loc.gov/standards/iso639-5/index.html [accesso 17/11/2019].

Katharina Knapp. 2017. Elenco dei siti atlantistici e lessici dialettali online. https://www.kit.gwi.uni-muenchen.de/?p=12110 [accesso 17/11/2019].

Linguistic Linked Open Data: https://linguistic-lod.org/ [accesso 28/11/2019].

Mark Wilkinson, Michel Dumontier et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3. https://www.nature.com/articles/sdata201618 [accesso 05/09/2019].

Richard Cyganiak und Anja Jentzsch. 2007 - . Linked Open Data Cloud. https://lod-cloud.net/ [accesso 31/10/2019].

Roberto Navigli e Simone Paolo Ponzetto. 2010. Babelnet: building a very large multilingual semantic network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 216-225. Association for Computational Linguistics. https://www.aclweb.org/anthology/P10-1023/ [accesso 28/11/2019].

Stephan Lücke. 2016. Digitalizzazione. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DD%2315

Stephan Lücke. 2018. FAIR-Prinzipien. *VerbaAlpina-it* 19/1 (creazione 18/2), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D191%26letter%3DP%23128

Thomas Krefeld e Stephan Lücke, ed. by. 2014 –. *VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit*. München, online. http://dx.doi.org/10.5282/verba-alpina

Thomas Krefeld e Stephan Lücke. 2016. Codice Beta. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D191%26letter%3DB%237

Thomas Krefeld e Stephan Lücke. 2016a. Tipizzazione. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D191%26letter%3DT%2358

Thomas Krefeld. 2016. Comunicazione scientifica nel web. *VerbaAlpina-it* 19/1 (creazione 16/1), *Metodologia*. https://doi.org/10.5282/verba-alpina?urlappend=%2Fit%3Fpage_id%3D21%26db%3D191%26letter%3DC%2362

Thomas Krefeld. 2018. I principi FAIR nel progetto VerbaAlpina, ossia il trasferimento della geolinguistica alle Digital Humanities. *VerbaAlpina-de 19/1* (creazione 18/2). https://www.verba-alpina.gwi.uni-muenchen.de/?p=8212&db=191-rf1-8212

Wikidata a. Introduction. https://www.wikidata.org/wiki/Wikidata:Introduction [accesso 05/09/2019].

Wikidata b. Lexikographische Daten/Dokumentation. https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation/de [accesso 05/09/2019].

Wikidata c. https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Glossary [accesso 17/11/2019].

Wilhelm Meyer-Lübke. 1935. Romanisches Etymologisches Wörterbuch, versione online. urn:nbn:de:bvb:355-ubr07799-0 [accesso 17/11/2019].

Una proposta di ontologia basata su RDA per il patrimonio culturale di Vincenzo Bellini

Salvatore Cristofaro
Istituto di Scienze e Tecnologie della Cognizione
CNR, Italia
salvatore.cristofaro@istc.cnr.it

Daria Spampinato
Istituto di Scienze e Tecnologie della Cognizione
CNR, Italia
daria.spampinato@cnr.it

Abstract

English. The rich cultural heritage preserved in the Belliniano Civic Museum of Catania has been studied and promoted in the last years mainly thanks to the BellinInRete project. It includes collections of objects (or *resources*) of a very different nature: paintings, photos, pianos, autograph scores, manuscript sheets, books preserved in the Museum's library, etc. In order to make the Belliniano Museum's heritage interoperable and reusable by scholars and cultural operators, we propose to semantically organize it in a unique homogeneous container, the OntoBellini ontology, designed and developed according to the Linked Open Data and Semantic Web paradigms. The wide variety of the involved museum resources, not even fully digitalised and catalogued, led us to the idea of experimenting with the RDA (*Resource Description and Access*) standard for creating library, archive and cultural heritage resource metadata. In this paper we describe the ongoing work towards the realization of the OntoBellini ontology.

Italiano. Il ricco patrimonio culturale conservato al Museo Civico Belliniano di Catania è stato studiato e promosso negli ultimi anni principalmente grazie al progetto BellinInRete. Tale patrimonio culturale comprende collezioni di oggetti (o *risorse*) di natura molto variegata: dipinti, foto, pianoforti, partiture autografe, fogli manoscritti, libri conservati nella biblioteca del Museo, ecc. Al fine di rendere il patrimonio del Museo Belliniano interoperabile e riutilizzabile da studiosi, operatori culturali ed utenti generici, se ne propone l'organizzazione semantica in un unico contenitore omogeneo, l'ontologia OntoBellini, progettata e sviluppata secondo i paradigmi del Linked Open Data e del Semantic Web. La grande varietà delle risorse museali coinvolte, non ancora completamente digitalizzate e catalogate, ha condotto all'idea di sperimentare lo standard RDA (*Resource Description and Access*) per la creazione di metadati di risorse di biblioteche, archivi e beni culturali. In questo articolo viene descritto il lavoro in corso per la realizzazione dell'ontologia OntoBellini.

1 Introduzione

Il patrimonio culturale conservato al Museo Civico Belliniano di Catania comprende collezioni di oggetti (o *risorse*) di natura molto variegata riconducibile ai settori museale, bibliografico e archivistico con la specificità del domino musicale. Allo stato attuale, le risorse identificate consistono di (circa):

- 250 oggetti tra dipinti, foto, pianoforti, spille, orologi, mobili, poster, medaglie, tessuti, ecc.;
- 4.500 fogli manoscritti di documenti e lettere;
- 9.300 fogli di spartiti manoscritti;
- 1.900 partiture a stampa;
- 50 opuscoli musicali a stampa;
- 280 libri della biblioteca del museo;

– 60 dischi in vinile di varie composizioni musicali.

Negli ultimi anni, questo ricco patrimonio culturale è stato promosso in particolare dal progetto BellinInRete (Del Grosso et al., 2018). Il progetto BellinInRete nasce dalla collaborazione tra il Comune di Catania, l'Istituto di Scienze e Tecnologie della Cognizione del CNR e il Dipartimento di Scienze Umanistiche dell'Università degli Studi di Catania. Esso mira a rinnovare e creare un cambiamento duraturo nella valorizzazione del Museo Civico Belliniano di Catania.

Al fine di rendere il patrimonio culturale del Museo Belliniano interoperabile e riutilizzabile da studiosi, operatori culturali ed utenti generici si propone l'organizzazione semantica di questo patrimonio in un unico contenitore omogeneo, l'ontologia OntoBellini, progettata e sviluppata secondo i paradigmi del Linked Open Data e del Semantic Web. La grande varietà delle risorse museali coinvolte, non ancora completamente digitalizzate e catalogate, ha condotto all'idea di sperimentare lo standard di metadatazione RDA (*Resource Description and Access*).[1] RDA è un *package* di concetti e istruzioni per la creazione di metadati di risorse eterogenee di biblioteche, archivi e beni culturali (Bianchini and Guerrini, 2016).

In questo articolo viene descritto il lavoro in fase di sviluppo per la realizzazione dell'ontologia OntoBellini.[2]

L'articolo è organizzato come segue. Nella Sezione 2 vengono esaminati brevemente alcuni lavori correlati e nella Sezione 3 viene descritto il lavoro svolto e in corso di realizzazione relativo all'analisi e alla rappresentazione delle risorse del Museo Belliniano, motivando l'esplorazione e lo sfruttamento di RDA per la costruzione dell'ontologia OntoBellini. La Sezione 4 presenta, a titolo di esempio, una descrizione tassonomica ad alto livello della parte dell'ontologia OntoBellini che si intende sviluppare relativa ad un corpus di lettere di Vincenzo Bellini e, infine, nella Sezione 5 si traggono le conclusioni e si discutono suggerimenti per lavori futuri.

2 Lavori correlati

Negli corso degli anni sono state presentate varie proposte riguardanti l'organizzazione semantica del patrimonio culturale dei musei. Molte di esse si basano sul *CIDOC Conceptual Reference Model* (CIDOC-CRM),[3] che rappresenta lo standard internazionale per lo scambio controllato di informazioni riguardanti i beni culturali dal 2006. CIDOC-CRM fornisce un'ontologia di base generale che può essere adottata in contesti concernenti il patrimonio culturale per sviluppare sistemi informativi semantici basati sul web e per migliorare la condivisione delle informazioni. Basandosi su CIDOC-CRM, sono stati sviluppati vari modelli di organizzazione della conoscenza volti a migliorare l'espressività semantica nel dominio del patrimonio culturale e per affrontare questioni specifiche non completamente contemplate da altri modelli esistenti. Questo è il caso, ad esempio, delle ontologie *entry OA* e *entry F* presentate in (Daquino et al., 2017), che arricchiscono le capacità descrittive di CIDOC-CRM attraverso la definizione di svariate possibili relazioni tra opere d'arte (entry OA) e fotografia (entry F), seguendo gli standard italiani promossi dall'ICCD[4] *Scheda OA* e *Scheda F*, rispettivamente. In (Moraitou et al., 2019) è possibile trovare un ampio elenco di altri progetti e proposte basate su CIDOC-CRM nel settore dei beni culturali.

Nel contesto della promozione del patrimonio culturale è emerso di recente lo standard RDA. Gli obiettivi principali di RDA sono l'identificazione e la messa in relazione di *entità* a livello astratto. Inizialmente, RDA implementava il modello di dati *Functional Requirements for Bibliographic Records* (FRBR), classificando le risorse informative in termini di una gerarchia di entità a quattro livelli chiamata WEMI (*Work*, *Expression*, *Manifestation*, *Item*).[5] Successivamente, dal novembre 2016, il comitato direttivo di RDA ha concordato l'adozione dell'*IFLA Library Reference Model* (LRM)[6] come modello

[1] http://www.rda-rsc.org/. Tutti gli url citati in questo contributo sono stati visitati il 27 novembre 2019.

[2] Il presente articolo costituisce una versione aggiornata del contributo ad opera degli stessi autori dal titolo *OntoBellini: towards an RDA based ontology for Vincenzo Bellini's cultural heritage* presentato al convegno JOWO 2019, 23 settembre 2019.

[3] http://www.cidoc-crm.org/

[4] ICCD (*Istituto Centrale per il Catalogo e la Documentazione*) - http://www.iccd.beniculturali.it/

[5] https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8915

[6] https://www.ifla.org/publications/node/11412

concettuale per lo sviluppo di RDA,[7] sostituendo FRBR.

RDA aspira a fornire uno standard universale per il *data-recording*, un codice univoco per rappresentare risorse eterogenee che si possono trovare in:

(A) biblioteche (manoscritti, libri, musica e film);

(B) archivi (documenti istituzionali, documenti personali e familiari e documentazione commerciale);

(C) musei (opere d'arte, costumi, oggetti e foto).

Si evidenzia che, nel contesto italiano, le risorse relative a biblioteche, archivi e musei sono gestite, attraverso l'utilizzo di norme solide e riconosciute, dalle rispettive istituzioni ICCU[8], ICAR[9] e ICCD. Mentre l'Associazione Italiana MAB[10] esplora le prospettive di convergenza tra i professionisti e le competenze in materia di musei, archivi e biblioteche.

Negli ultimi anni RDA ha attratto l'interesse di diverse istituzioni culturali pubbliche sia europee che d'oltre oceano che lo hanno adottato e implementato, sperimentandone applicazioni alla catalogazione e condivisione di risorse bibliotecarie (si vedano, ad esempio, (Ducheva and Pennington, 2017) e (Panchyshyn et al., 2019)).

3 (Ri)organizzazione dei dati museali

Nell'ambito del progetto BellinInRete, il patrimonio del Museo Civico Belliniano è stato parzialmente studiato e analizzato da esperti musicologi e da specialisti con competenze museali, archivistiche e bibliotecarie, con l'obiettivo di recuperare informazioni sulle risorse del museo. Queste sono state quindi rappresentate formalmente come record di dati che comprendono diversi campi di informazione (si veda sotto). La collezione di questi record costituisce la base su cui si fonda la presente proposta di organizzazione semantica del patrimonio belliniano.[11]

I record delle risorse museali sono stati creati seguendo gli standard italiani ICCD e ICCU per la catalogazione e la documentazione (*Scheda OA*, *Scheda F* e *schede SBN*). Il numero dei campi di ciascuno record e il loro significato dipende dal tipo di risorsa rappresentata dal record stesso. Sono stati identificati 14 diversi *tipi base* di risorse museali, ossia:

Manoscritti	***Testi a stampa***	***Musica manoscritta***	***Musica a stampa***
Materiale grafico	***Arredi***	***Dipinti***	***Documenti***
Foto	***Medaglie***	***Statue***	***Strumenti musicali***
	Tessuti	***Oggetti generici***	

All'interno di ciascun tipo base, le risorse sono suddivise, a loro volta, in sottotipi più specializzati. Ad esempio, i ***Manoscritti*** comprendono: *lettere (originali)*, *bozze e minute di lettere*, *copie di lettere*, *certificati di battesimo*, *certificati di morte*, *certificati di matrimonio*, *note di spesa*, *bollettini medici*, ecc. Il ***Materiale grafico*** comprende i *poster*, mentre gli *spartiti* rientrano nella ***Musica manoscritta***. Gli ***Oggetti generici*** comprendono oggetti personali di Vincenzo Bellini, come *orologi* e *spille*, e altri oggetti di vita quotidiana come *cucchiai*, *coltelli*, *tazze*, ecc. In Figura 1 si riporta una selezione di campi di record in forma tabellare. Si noti che il blocco di informazioni memorizzato in alcuni campi di record presenta un basso livello di granularità che potrebbe essere ulteriormente raffinato suddividendo il blocco tra campi aggiuntivi di dati atomici. Questo è il caso, ad esempio, del campo **formato** (si veda nella Figura 1 la penultima colonna della tabella più in alto) che viene utilizzato per descrivere alcune caratteristiche fisiche di un manoscritto, come dimensioni, numero di pagine, foliazione, direzione della scrittura, ecc. Si osservi anche che alcuni campi di record sono specifici per il particolare tipo di risorsa

[7] http://www.rda-rsc.org/ImplementationLRMinRDA

[8] ICCU (*Istituto Centrale per il Catalogo Unico*) - https://www.iccu.sbn.it/it/

[9] ICAR (*Istituto Centrale per gli Archivi*) - www.icar.beniculturali.it/

[10] MAB (*Musei Archivi Biblioteche*) - http://www.mab-italia.org/

[11] Si noti che, attualmente, il numero dei record creati corrisponde a circa il 70% del numero totale delle risorse stimate del museo. Il coinvolgimento delle rimanenti risorse è programmato per il prossimo futuro.

autore	editore	soggetto	descrizione	data	formato	lingua
Manoscritti						
Dall'Ongaro, Francesco <1808-1873>	Trieste : autografo	Perucchini, Giovanni Battista - Lettere e carteggi	Contiene copia autografa di Dall'Ongaro di quattro lettere, l'ultima delle quali, sebbene anch'essa autografa di Francesco Dall'Ongaro, riproduce lo scritto di una donna non identificata e risulta incompleta	1842-04-21	1 lettera, cc. 4rv ; mm 270 x 210	ITA
Materiale grafico						
	Milano : Tipografia Pirola			1834-05-03	1 manifesto ; 41 x 30 cm	ITA
Musica a stampa						
Bellini, Vincenzo <1801-183j>	Milano : Gio. Ricordi ; Firenze : Gio. Ricord e C.°, [1829]		Spartito del duetto Io troverò nell'Asia nell'atto II dell'opera Zaira. Nel margine superiore sinistro: «Al museo Belliniano - | 12/7.1934.XII - Bazan Ascanio»	1829	1 spartito (15 p.) ; 24 x 33 cm	ITA
Musica manoscritta						
Bellini, Vincenzo <1801-1835>	[S.l.] : copia, [1819-1826]		Spartito dell'aria "Quando incise sul quel marmo". Sul frontespizio nell'angolo superiore destro: «Al museo Belliniano | M° Bazan Ascanio»	1819	1 spartito manoscritto (8 c.) ; 220 x 270 mm	ITA

tipologia	descrizione	dimensioni	stato di conservazione	indicazioni specifiche	provenienza	cronologia
Arredi						
MOBILE	LIBRERIA CON CASSETTINI E	CM. 100 X 163	DISCRETO	RIPORTA INTARSI FLOREALI SUL FRONTALE	DONO SALVATORE POLLINA	XIX SECOLO
TAVOLO	IN LEGNO OVALE	CM. 71 X 47	DISCRETO	STILE IMPERO SOSTENUTO DA UNA CHIMERA	DONO	XIX SECOLO
Dipinti						
QUADRO	RITRATTO GIOVANILE DI BELLINI	CM. 63 X 74	OTTIMO	DIPINTO AD OLIO SU TELA	DONO	XIX SECOLO
Medaglie						
MEDAGLIA	REAL ORDINE DI FRANCESCO I DI BORBONE	CM. 4 CIRCONFERENZA	BUONO	EFFIGE DEL RE – IN ARGENTO IN CUSTODIA ROSSA	-	28 SETTEMBRE 1829
Statue						
STATUA	BELLINI ALL'ETA' DI 5 ANNI	H. CM. 183	DISCRETO	IN GESSO SU PIEDISTALLO IN LEGNO	DONO DEI PARENTI DELLO SCULTORE	XIX SECOLO
Strumenti musciali						
PIANOFORTE	PIANOFORTE VERTICALE INGLESE	CM. 104 X 240	DISCRETO	NON FUNZIONANTE	DONO CATERINA NICOLOSI CIRELLO	XIX SECOLO
Tessuti						
BACHECA MURALE	TRE ANGOLI DI TAPPETO CON TRE OPERE BELLINIANE	CM. 51 X 143	DISCRETO	RAPPRESENTANO "STRANIERA, PIRATA, SONNAMBULA" – RIPORTA N° 62	DONO DEL NIPOTE ASCANIO BAZAN	1831
Oggetti generici						
OROLOGIO	A CILINDRO IN ORO		PESSIMO	CASSA ARABESCA	PROPRIETA' DI BELLINI	XIX SECOLO

Figura 1: Alcuni record corrispondenti alle risorse del Museo Belliniano: le righe verdi contengono i nomi dei campi dei record; le righe blu indicano i tipi base delle risorse. Si noti che la tabella più in alto coinvolge solo risorse cartacee.

museale descritta da questi campi. Ad esempio, il campo **lingua** (cfr. Figura 1) è stato specificamente utilizzato per rappresentare la(e) lingua(e) delle risorse scritte e non può certo essere applicato agli oggetti; così come non ha senso parlare (ad esempio) della foliazione di un tavolo o di una sedia (infatti la foliazione è una informazione specifica del campo **formato** relativo ai manoscritti).

Il patrimonio del Museo Belliniano coinvolge anche alcuni *oggetti fisici composti* (come contenitori per medaglie e cornici fotografiche) che richiedono una struttura gerarchica di record per essere ragionevolmente descritti.[12] Inoltre, il Belliniano conserva anche alcuni libretti musicali che non sono stati ancora catalogati. Si sottolinea ulteriormente che diversi documenti d'archivio (come i vari certificati), hanno ricevuto ad oggi un'analisi solo approssimativa: all'interno del progetto BellinInRete si prevede di creare metadati dettagliati per essi seguendo gli standard adottati dal *Sistema Archivistico Nazionale Italiano*[13] gestito dall'ICAR.

Come emerge dalle considerazioni precedenti, le rappresentazioni delle risorse del Museo Belliniano create presentano, allo stato attuale, un carattere eterogeneo con un basso livello di granularità che rende difficile tradurle in una base di conoscenza ontologica espressiva ed efficace.[14] (Si noti che ciò deriva in

[12]Allo stato attuale, tali oggetti composti non sono ancora stati disassemblati per motivi di conservazione, e quindi, al momento, è stato possibile recuperare poche informazioni descrittive per essi.

[13]http://san.beniculturali.it/SAN

[14]Si osservi comunque che recentemente una parte (ristretta) delle risorse museali del Belliniano, consistente in un corpus di lettere di Vincenzo Bellini, è stata oggetto di studi sistematici approfonditi che hanno evidenziato diversi aspetti semanticamente interessanti facilmente formalizzabili in un'ontologia. (Ciò verrà discusso nella Sezione 4.)

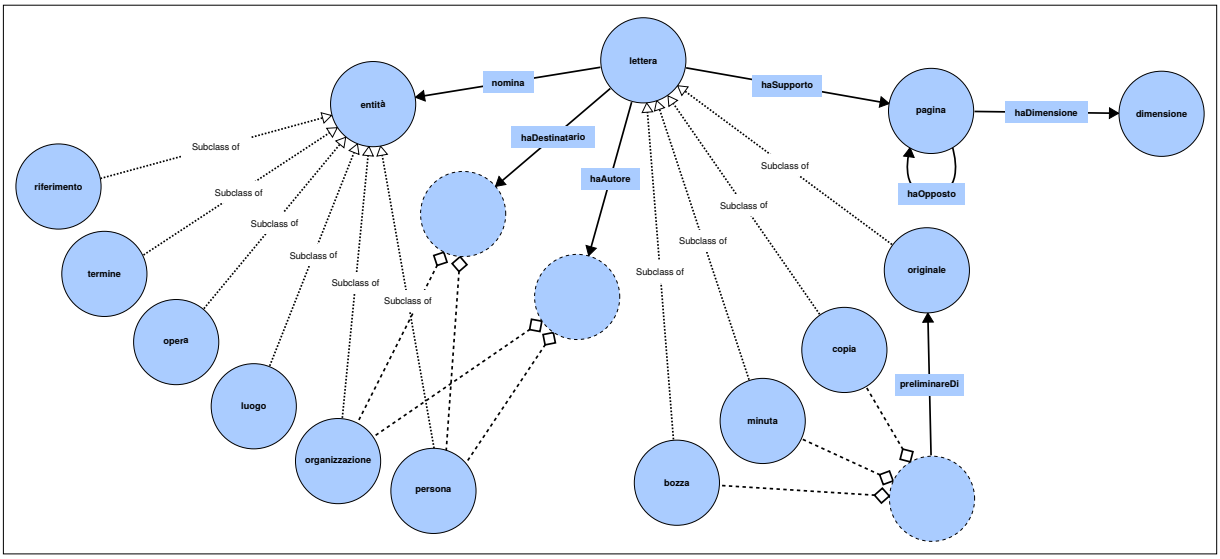


Figura 2: Schema ontologico del corpus epistolare belliniano.

parte dai particolari criteri di rappresentazione adottati per la creazione dei record di dati corrispondenti alle risorse del museo.)

Al fine di migliorare tali rappresentazioni sarebbe innanzitutto utile pulire e raffinare i record dei dati acquisiti, in modo da ottenere una collezione più uniforme. Quindi, le istruzioni RDA potrebbero poi essere proficuamente sfruttate per ottenere una (ri)organizzazione più efficace dei dati. Difatti se si dovesse rappresentare soltanto la collezione museale (composta da oggetti unici) si potrebbe utilizzare CIDOC-CRM che è stato progettato principalmente per questa tipologia di risorse. Ma volendo utilizzare un unico modello di rappresentazione dei dati per l'intero patrimonio belliniano, RDA, attraverso il meccanismo di classificazione WEMI (ereditato da FRBR) si presta meglio alla descrizione delle risorse.[15]

In termini molto generali, le principali attività coinvolte nello sviluppo dell'ontologia OntoBellini, possono quindi essere schematizzate come segue. Dopo una prima fase di ristrutturazione dei dati, con l'obiettivo di creare collezioni di record più omogenee e più dettagliate (come descritto sopra), si prevede di identificare i concetti e le proprietà alla base dell'ontologia OntoBellini in conformità con il framework entità-relazioni di RDA, e quindi sviluppare l'ontologia stessa rendendola accessibile via web.

4 Il caso delle lettere Belliniane

Una parte peculiare del progetto BellinInRete riguarda la rappresentazione, l'organizzazione e la codifica secondo lo standard TEI-XML (*Text Encoding Initiative*),[16] di un corpus di lettere della corrispondenza di Vincenzo Bellini (*corpus epistolare*) che forniscono informazioni interessanti per diversi aspetti legati alla vita sociale e all'attività artistica in ambito musicale del compositore catanese (si veda (Del Grosso et al., 2018)). In questa sezione viene fornita, a titolo esemplificativo, una descrizione ad alto livello della tassonomia di base della parte dell'ontologia OntoBellini che si intende sviluppare relativa al solo corpus epistolare, presentandone il corrispondente schema ontologico (cfr. Figura 2). Questo corpus epistolare, infatti, è stato recentemente studiato ed analizzato in maniera più dettagliata ed approfondita rispetto alle altre risorse museali e, a differenza di queste ultime, le informazioni disponibili ad esso relative risultano, allo stato attuale, più complete e strutturate e permettono di delineare un quadro più concreto e definito degli *item di conoscenza* da rappresentare e dedurre attraverso l'ontologia. Più specificatamente, l'analisi del corpus epistolare belliniano ha condotto alle seguenti considerazioni. Il corpus epistolare è

[15] Si menziona che, basandosi sui modelli IFLA FR, è stata sviluppata un'estensione di CIDOC-CRM, ossia FRBRoo (http://www.cidoc-crm.org/frbroo-0), che intende rappresentare la semantica delle informazioni bibliografiche e facilitare l'integrazione e lo scambio di risorse bibliografiche e museali. Tuttavia IFLA LRM, come modello concettuale sottostante di RDA, consente un livello di generalità maggiore rispetto a FRBRoo, poiché include meno dettagli di quest'ultimo.

[16] https://www.tei-c.org/

costituito essenzialmente da 4 tipologie di *documenti epistolari* (o *lettere*) ossia: *bozze* (di lettere), *minute* (di lettere), *copie* (di lettere) e *originali* (di lettere).[17] Ogni documento epistolare ha uno o più *autori*, è indirizzato ad uno o più *destinatari* ed è fisicamente contenuto (scritto) su un *supporto* costituito da una o più facciate di fogli di carta (*pagine*) aventi differenti *dimensioni*: uno stesso documento può essere infatti frammentato su diverse pagine e (parti di) documenti diversi possono trovarsi su una stessa pagina o su *pagine opposte* (fronte-retro) di uno stesso foglio di carta. Inoltre, in ogni documento epistolare vengono *nominate* (in maniera esplicita o implicita) diverse *entità* interessanti quali *persone*, *organizzazioni*, *luoghi*, *opere* (musicali), *termini* (musicali e non) e *riferimenti* (bibliografici); in particolare, il complesso delle *persone* e delle *organizzazioni* include gli *autori* e i *destinatari* delle lettere menzionati sopra.

Le precedenti considerazioni si traducono nello schema ontologico riportato in Figura 2 che rappresenta la tassonomia di base della parte dell'ontologia OntoBellini relativa al corpus epistolare del Belliniano.[18]

Si osservi che l'organizzazione semantica proposta e rappresentata in Figura 2 per il corpus epistolare è molto generale e di alto livello. Ai fini dell'espressività essa può (e deve) essere specializzata imponendo degli opportuni *vincoli semantici* attraverso l'introduzione di appositi *assiomi di classi e proprietà*. Ad esempio, in riferimento alla rappresentazione in Figura 2, sarebbe ragionevole assumere che le classi **bozza**, **minuta**, **copia** e **originale** siano *disgiunte* (si veda la nota n. 17) e che inoltre le proprietà **haDimensione** e **haOpposto** siano entrambe *funzionali* e con la seconda ulteriormente *simmetrica* e con *inversa funzionale*.[19] In aggiunta si potrebbe postulare anche la validità della proprietà che gli autori e i destinatari delle bozze, delle copie e delle minute coincidano con quelli delle rispettive versioni originali, e così via.

5 Conclusioni e lavori futuri

In questo lavoro è stata proposta l'organizzazione semantica del patrimonio culturale conservato nel Museo Civico Belliniano di Catania attraverso un'ontologia condivisa –l'ontologia OntoBellini–, basandosi sulla grande quantità di dati attualmente acquisiti per le risorse del museo. Il basso livello di granularità e il carattere eterogeneo di questi dati richiede tuttavia una riorganizzazione preliminare degli stessi al fine di renderli più omogenei e facilmente codificabili nell'ontologia. A tal fine si prevede di sfruttare le indicazioni RDA per la creazione di metadati di risorse di biblioteche e beni culturali. A titolo di esempio è stata brevemente descritta una proposta di organizzazione semantica ad alto livello relativa ad un corpus di lettere di Vincenzo Bellini custodite nel museo Belliniano, presentandone il corrispondente schema ontologico.

Bibliografia

Dean Allemang and James Hendler. 2001. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Elsevier Science, second edition.

Carlo Bianchini and Mauro Guerrini. 2016. RDA, Resource Description and Access: The metamorphosis of cataloguing. *JLIS.it* 7(2).

Marilena Daquino, Francesca Mambelli, Silvio Peroni, Francesca Tomasi, and Fabio Vitali. 2017. Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data. *Journal on Computing and Cultural Heritage* 10(4).

[17] Si osservi che, concettualmente ogni bozza, minuta o copia è considerata una *versione preliminare* di una qualche lettera originale (anche se l'originale non è sempre presente tra le risorse del museo). Si noti inoltre che, nel presente contesto, per *copia* si intende sostanzialmente la trascrizione fedele di un documento originale e che la minuta rappresenta una versione più aggiornata, completa e possibilmente corretta rispetto ad una bozza. I concetti bozza, minuta, copia e originale sono da intendersi come entità distinte.

[18] La rappresentazione grafica dello schema ontologico riportata in figura è stata ottenuta attraverso il tool di visualizzazione interattiva di ontologie WebVOWL accessibile all'url http://vowl.visualdataweb.org/webvowl.html. Si osservino le classi anonime rappresentate dai nodi con contorno tratteggiato che corrispondono all'unione delle classi ad esse collegate (si veda (Allemang and Hendler, 2001)).

[19] Ricordiamo che due classi A e B sono disgiunte ($A \cap B = \emptyset$) se non hanno istanze in comune e che una proprietà R è funzionale se per ogni individuo A esiste al più un unico individuo A in relazione con A attraverso A. Si veda (Allemang and Hendler, 2001).

Angelo M. Del Grosso, Salvatore Cristofaro, Maria R. De Luca, Emiliano Giovannetti, Simone Marchi, Graziella Seminara, and Daria Spampinato. 2018. Le lettere di Bellini: dalla carta al web. In *Quaderni di Umanistica Digitale: AIUCD 2018 - Book of abstracts*. pages 60–64.

Dilyana Ducheva and Diane Pennington. 2017. Resource Description and Access in Europe: Implementations and perceptions. *Journal of Librarianship and Information Science* 51(2):387–402.

Efthymia Moraitou, John Aliprantis, Yannis Christodoulou, Alexandros Teneketzis, and George Caridakis. 2019. Semantic Bridging of Cultural Heritage Disciplines and Tasks. *Heritage* 2:611–630.

Roman S. Panchyshyn, Frank P. Lambert, and Sevim McCutcheon. 2019. Resource Description and Access Adoption and Implementation in Public Libraries in the United States. *Library Resources & Technical Services (LRTS)* 63(2).

Biblioteche di conservazione e libera fruizione dei manoscritti digitalizzati: la Veneranda Biblioteca Ambrosiana e la svolta inevitabile grazie a IIIF

Fabio Cusimano
Veneranda Biblioteca Ambrosiana, Milano
fcusimano@ambrosiana.it

Abstract

English. The Veneranda Biblioteca Ambrosiana in Milan, thanks to a complex project resulting from the cultural and scientific collaboration and the sharing of resources between the Catholic University of the Sacred Heart of Milan and the University of Notre Dame (IN, USA), has recently inaugurated its new digital library devoted to the open and free access to its digitized manuscript collections: an inevitable turning point made possible thanks to the adoption of innovative technologies such as the IIIF-International Image Interoperability Framework.

Italiano. La Veneranda Biblioteca Ambrosiana di Milano, grazie a un complesso progetto frutto della collaborazione culturale e scientifica e della condivisione di risorse tra l'Università Cattolica del Sacro Cuore di Milano e la *University of Notre Dame* (IN, USA), ha recentemente inaugurato la propria nuova biblioteca digitale ad accesso aperto e gratuito dedicata alla fruizione del proprio patrimonio manoscritto digitalizzato: una svolta inevitabile resa possibile grazie all'adozione di tecnologie innovative quali IIIF-*International Image Interoperability Framework*.

> «As the innovative media of their time, manuscripts generated awe and wonder: illuminated manuscripts […] celebrated and cultivated the human capacity for wonder, and techniques are needed to restore this sense. The Middle Ages is much more aptly described as the Age of Visual Wonder». (Endres 2019, 5)

1 Le funzioni della biblioteca di conservazione e la digitalizzazione

Alla prova della tradizione storica, la *mission* che da sempre caratterizza le biblioteche consiste nel raccogliere, organizzare, ampliare e diffondere la conoscenza tramite l'accesso alle risorse in esse custodite (Bottasso 1999; Gorman 2004; Kempf 2013; Fabian 2015, 55-70), alla costante ricerca di un delicato equilibrio tra fruizione e conservazione.

A tal proposito, come affermano Montecchi e Venuda, «Le due dimensioni dell'attività bibliotecaria, quella orizzontale dell'uso dei libri da parte dei nostri contemporanei e quella verticale della loro conservazione per i posteri, costituiscono i due poli attorno ai quali si strutturano i servizi di ogni biblioteca: al prevalere dell'uno o dell'altro avremo "biblioteche di conservazione" o "biblioteche d'uso", anche se è ben difficile incontrare biblioteche finalizzate unicamente ed esclusivamente all'una o all'altro. Non esiste, infatti, neppure in sede teorica, una netta opposizione tra questi due parametri, essendo la conservazione finalizzata all'uso sia presente che futuro del libro e, sull'altro versante, non potendo l'uso dei libri in biblioteca prescindere da forme di tutela e di conservazione che assicurino loro lunga vita tra gli uomini» (Montecchi e Venuda 2006, 79).

Sebbene ogni epoca abbia vissuto fondamentali momenti di evoluzione e progresso in ogni campo del sapere e della tecnica – spesso inavvertiti (Montecchi e Venuda 2006, 23; Eisenstein 1986; Eisenstein 2004; Barbier 2005; Roncaglia 2010; Febvre 2011; McLuhan 2011; Bertolo 2016; Cursi 2016) dai contemporanei – la nostra società appare dotata di tali e tanti strumenti tecnologici potenzialmente utili a diffondere la conoscenza e la cultura che viene spontaneo chiedersi come sia possibile che tutto questo non abbia coinvolto nativamente il mondo delle biblioteche! Al giorno d'oggi, infatti, non siamo ancora riusciti ad affrancarci da quello che ormai sembra essere divenuto un vero luogo comune, ovvero il rapporto antitetico tra biblioteca e tecnologia: in un simile contesto, la biblioteca di conservazione viene ancora percepita come il luogo refrattario per eccellenza alla tecnologia e all'innovazione, destinato per definizione alla sola tesaurizzazione del proprio prezioso patrimonio.

2 Il XVII secolo e l'innovatività della Veneranda Biblioteca Ambrosiana di Milano

La Veneranda Biblioteca Ambrosiana (Rodella 1992, 121-147; Panizza 2012) di Milano viene tradizionalmente considerata – sin dalla sua solenne inaugurazione avvenuta l'8 dicembre 1609 – come uno dei primi e principali esempi di biblioteca pubblica (IFLA/UNESCO Public Library Manifesto 1994) nell'accezione di un'istituzione creata con il chiaro intento di fornire accesso ai libri (Natale 1995, 1-2) a una comunità di lettori (Galluzzi 2011) quanto più ampia possibile (Serrai 2005, 7-9; Rovelstad 2000, 540-556).

Si ritiene utile approfondire alcuni tratti caratteristici della fondazione della Biblioteca Ambrosiana che il suo ideatore e fondatore – il cardinale Federico Borromeo (Prodi 1971, 33-42; Ravasi 1992, 1-19; Buzzi e Ferro 2005) – fortemente volle aperta a tutti. Per farlo sarà opportuno approcciarsi al modello di biblioteca tipico del tempo (Burke 1992, 391-416; Ghilli 2015, 365-376; DeSeta 2016), soprattutto attraverso la testimonianza di Gabriel Naudé (Rovelstad 2000, 549), autore del celebre *Advis pour dresser une bibliothèque* (Naudé 1627). Egli riserva al IX e ultimo capitolo del suo *Advis*, dal titolo *Quel doit estre le but principal de cette Bibliotheque*, l'aspetto più importante legato alla trattazione teorica sull'allestimento di una biblioteca: quale debba essere lo scopo principale di una biblioteca ben allestita.

In questo modo il Naudé loda senza riserve i particolari servizi che fanno dell'Ambrosiana una vera biblioteca aperta al pubblico, unica nel suo genere:

> Car pour ne parler que de l'Ambrosienne de Milan, & monstrer par mesme moyen comme elle surpasse tante en grandeur & magnificence que en obligeant le public beaucoup de celles d'entre les Romains, n'est-ce pas une chose du tout extraordinaire qu'un chacun y puisse entrer à toute heure presque que bon luy semble, y demeurer tant qu'il luy plaist, voir, lire, extraire tel Autheur qu'il aura agreable, avoir tous les moyens & commoditez de ce faire, soit en public ou en particulier, & ce sans autre peine que de s'y transporter és iours & heures ordinaires, se placer dans des chaires destinees pour cet effect, & demander les livres qu'il voudra fueillerer au Bibliothecaire ou à trois de ses serviteurs, qui sont fort bien stipendiez & entretenus, tant pour servir à la Bibliotheque qu'à tous ceux qui viennent tous les iours estudier en icelle (Naudé 1627, 155-156).

Proprio in relazione al precedente passo, il Serrai puntualizza che «per l'Ambrosiana il riconoscimento di Naudé, evidentemente frutto di esperienza diretta, va ancora oltre per sfociare in un'autentica stupefatta ammirazione» (Serrai 2005, 8).

La descrizione del Naudé, infine, si avvia alla conclusione con altri interessanti spunti che vedono ancora l'Ambrosiana assunta a termine di paragone:

> […] il faudroit premierement observer que toutes les Bibliotheques ne pouvant tousiours estre ouvertes comme l'Ambrosienne, il fust au moins permis à tous ceux qui y auroient affaire d'aborder librement le Bibliothecaire pour y estre introduits par iceluy sans aucune dilation ny difficulté: secondement que ceux qui seroient totalement incognus, & tous autres qui n'auroient affaire que de quelques passages, peussent veoir chercher & extraire de toutes sortes de livres imprimez ce dont ls auroient besoin: tiercement que l'on permist aux personnes de merite & de cognoissance d'emporter à leurs logis les livres communs & de peu de volumes; […] (Naudé 1627, 161-162).

Proprio riguardo agli spunti di cui il Naudé fa esplicita menzione, non si può non rimanere stupiti per quanto essi richiamino concetti e servizi di cui oggi si fa un gran parlare, quali, ad esempio, il servizio di *reference* e il prestito dei volumi: tutto questo può far sovvenire un collegamento tra la prassi descritta dal Naudé – che egli stesso auspica possa diffondersi quale strumento di base per l'utenza presso ogni biblioteca – e la digitalizzazione delle risorse catalografiche/librarie. Quale migliore risposta agli ideali del cardinale Federico Borromeo e del Naudé stesso, di una biblioteca le cui risorse possano essere sempre liberamente accessibili, ricercabili e consultabili proprio attraverso specifici servizi *online* quali, appunto, una nuova biblioteca digitale ad accesso libero?

3 Dalla *Bibliotheca* alla *Digital Library*: cura delle collezioni e *Data Curation*

Un altro documento, stavolta strettamente collegato alla vita della Veneranda Biblioteca Ambrosiana, è di fondamentale importanza a proposito del tratteggio della figura del bibliotecario: si tratta delle *Constitutiones Collegii ac Bibliothecae Ambrosianae* (Bentivoglio 1835; Marcora 1986, 155-164; Annoni 1992, 149-184).

Le *Consitutiones* ambrosiane dedicano un intero capitolo alla figura del bibliotecario, alle sue mansioni e alle tipologie dei cataloghi: si tratta del *Caput X*, *De Bibliothecario et Bibliotheca* (Bentivoglio 1835, 32-39). Presso la Biblioteca Ambrosiana il *Bibliothecarius* è stato affiancato dal *Custos catalogi*, il custode del catalogo (Rodella 2013, 35-36): tale espressione risulta essere etimologicamente molto interessante e, come vedremo, gioca anche un ruolo importante nell'apertura verso funzioni e attività caratteristiche dell'era digitale, quali il *Data Curator* e il derivato *Data Curation*.[1] Il *Data Curator* si ispira ai medesimi principi che guidavano il *Custos catalogi* del XVII secolo e opera per prendersi cura dei cataloghi (oggi prevalentemente OPAC), delle informazioni catalografiche (oggi prevalentemente codificate in formati standard come l'ISO2709), dei metadati (descrittivi, amministrativi, gestionali, tecnici, tutti accomunati dai *tag* e dai metalinguaggi adottati per la loro compilazione, quali, ad esempio, DublinCore e XML), degli oggetti digitali (così come dei diversi formati, specialmente per quanto concerne le immagini digitali), delle svariate procedure tecniche da attivare di volta in volta per avviare la produzione di nuovi oggetti digitali tramite l'utilizzo di differenti apparecchiature (macchine fotografiche digitali, scanner, ecc.), come anche per garantire la conservazione (*storage*) e il perdurare dell'informazione digitale. Altro fondamentale aspetto è quello della progettazione globale degli interventi di digitalizzazione e della messa a punto del necessario flusso di lavoro (*workflow*) ad essi collegati.

3.1. La nuova biblioteca digitale della Veneranda Biblioteca Ambrosiana

Nel percorso d'attuazione della nuova fase di digitalizzazione presso la Veneranda Biblioteca Ambrosiana si è cercato di tenere in debito conto quanto già sperimentato presso altre realtà a livello internazionale, avendo cura di porre le basi per la realizzazione di un progetto necessariamente scalabile e aperto a proficue collaborazioni e condivisioni, nazionali e internazionali, sia a livello tecnico che scientifico.

Il nodo del *data reuse*, per esempio, si è subito imposto in maniera molto concreta: con la precedente attività di digitalizzazione, infatti, è stata prodotta un'ingentissima mole di dati (oltre 1.800.000 immagini in formato .tif non compresso, colori, 24 bit) che rappresentano ancora oggi un prezioso nucleo composto da oltre 2.700 manoscritti integralmente digitalizzati su cui basare l'avvio di una nuova fase di digitalizzazione. Tale ingente quantità di dati, pari a circa 31 Tb di spazio-disco, necessita di cure costanti e va ad aggiungersi alla quotidiana produzione di nuove copie digitali di manoscritti, il tutto con il preciso obiettivo di rendere progressivamente disponibili online, gratuitamente e pubblicamente, le riproduzioni digitali integrali di parte del patrimonio manoscritto ambrosiano.

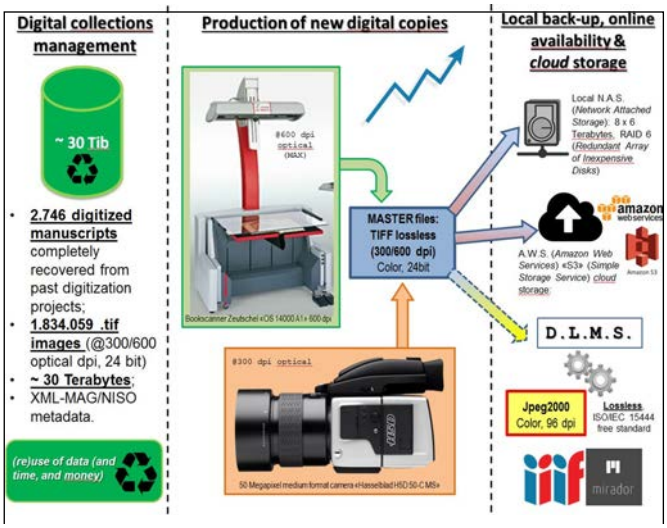


Figura 1: rappresentazione schematica dell'infrastruttura di digitalizzazione della Biblioteca Ambrosiana; in basso a destra: i loghi di IIIF-*International Image Interoperability Framework* e del visualizzatore Mirador.

[1] Proprio nel merito delle funzioni del *Data Curator* emerge evidente il collegamento etimologico al *Custos catalogi* cui ho fatto riferimento in precedenza: l'etimologia del termine inglese *curator* è direttamente derivata dal latino, rispettivamente dal verbo *curo* e dal sostantivo *curator*, ed è proprio per questo motivo che è possibile mettere in relazione tra loro le due figure.

Tale meritorio obiettivo chiama in causa un aspetto fondamentale al giorno d'oggi per la realizzazione di una nuova biblioteca digitale: l'utilizzo di IIIF-*International Image Interoperability Framework* (IIIF 2018) per la visualizzazione di contenuti digitali di qualità via Internet (Snydman 2015, 16-21; Brantl 2016, 10-13; Salarelli 2017, 50-66; Magnuson 2018; Cusimano 2019; Lit, Lambertus Willem Cornelis, van 2020, 160-167).

Viviamo in un'epoca in cui le tecnologie *web based*, la connettività, la diffusione di dispositivi mobili sempre più performanti rendono possibile ciò che solo due lustri fa non era nemmeno immaginabile: ogni biblioteca digitale di nuova generazione dovrebbe pertanto essere predisposta cercando di approfittare di tali condizioni tecnicamente favorevoli, avendo ben chiaro che essa sarà soggetta a diversi livelli di lettura che riguardano l'istituzione-biblioteca che la predispone e gli utenti che ne fruiranno.

L'ecosistema IIIF, dunque, si configura come la "scelta inevitabile" (cui non a caso faccio riferimento nel titolo del presente contributo) poiché consente – dal punto di vista degli utenti – una semplice ed efficace fruizione *online* dei contenuti digitalizzati; e, dal punto di vista dell'istituzione culturale promotrice (nel caso specifico, l'Ambrosiana), garantisce interoperabilità, scalabilità, personalizzazione e condivisione.[2] La nuova biblioteca digitale dell'Ambrosiana[3] implementa l'utilizzo delle APIs <IIIF Image API 2.1.1> (https://iiif.io/api/image/2.1/) e <IIIF Presentation API 2.1.1> (https://iiif.io/api/presentation/2.1/); il visualizzatore Mirador e l'*image server* Cantaloupe; il tutto è interconnesso con l'OPAC dell'Ambrosiana al fine di garantire il collegamento diretto tra il record catalografico/descrittivo del manoscritto ricercato e la relativa risorsa digitale: dalla scheda bibliografica presente nell'OPAC, infatti, tramite l'apposito *link* <Visualizza la copia digitale>, si attiva direttamente all'interno del *browser* il visualizzatore *web* Mirador (Mirador 2018) che consente all'utente *online* un'ottima esperienza di visualizzazione. Il *Manifest* .json relativo a ogni risorsa digitalizzata è pubblicamente reperibile all'interno della scheda informativa contrassegnata dalla "i" posta in alto a destra nell'interfaccia del visualizzatore Mirador.

Figura 2: le risorse digitalizzate sono rese pubblicamente e gratuitamente fruibili grazie all'utilizzo del visualizzatore IIIF *compliant* Mirador, e il punto di partenza della fruizione digitale è proprio il catalogo (OPAC) della biblioteca.

La biblioteca digitale può essere consultata attraverso la *landing page* predisposta (in italiano e in inglese) all'interno del sito *web* ufficiale della Veneranda Biblioteca Ambrosiana, raggiungibile attraverso due differenti percorsi tematici: <Scopri> (italiano: https://www.ambrosiana.it/scopri/biblioteca-digitale/; inglese: https://www.ambrosiana.it/en/discover/the-digital-library/) e <Studia> (italiano: https://www.ambrosiana.it/studia/biblioteca-digitale/; inglese: https://www.ambrosiana.it/en/study/the-digital-library/).

[2] «[…] One of the nicest things about the IIIF approach to shared content is that it lowers the barriers to building light-weight demonstrations like this for teaching and research purposes. The institutions that host the images are on the hook for long-term access and preservation, so it's not necessary to host your own copies of the images. […] There are thousands of manuscripts available now from interoperable repositories that can be used, and – with more institutions joining IIIF each year – thousands more in the offing. As the tools get easier to use and configure, it will be fascinating to see what becomes possible for medieval studies». (https://tinyurl.com/vnhpn3s).

[3] La Veneranda Biblioteca Ambrosiana è stata ufficialmente inserita all'interno della lista delle istituzioni che utilizzano IIIF (https://iiif.io/community/#participating-institutions) quale unica istituzione culturale italiana.

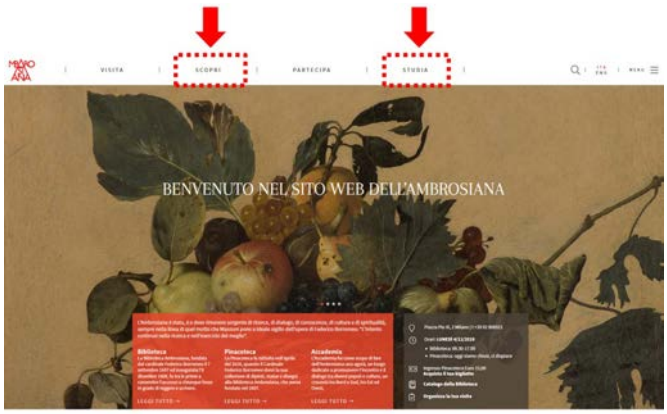


Figura 3: indicazione dei percorsi tematici <Scopri> e <Studia> per la consultazione della biblioteca digitale all'interno del sito *web* ufficiale della Veneranda Biblioteca Ambrosiana (https://www.ambrosiana.it).

Figura 4: dettaglio del percorso tematico <Scopri> all'interno del sito web ufficiale della Veneranda Biblioteca Ambrosiana (https://www.ambrosiana.it).

Figura 5: dettaglio del percorso tematico <Studia> all'interno del sito web ufficiale della Veneranda Biblioteca Ambrosiana (https://www.ambrosiana.it).

La biblioteca digitale dell'Ambrosiana si apre al pubblico attraverso la sezione ad essa dedicata all'interno del proprio OPAC: https://ambrosiana.comperio.it/biblioteca-digitale/. Da qui ogni utente può accedere alla consultazione pubblica e gratuita delle copie digitali seguendo due vie principali:

- attraverso la consultazione diretta della scheda catalografica del manoscritto di proprio interesse a partire dalla segnatura dello stesso: in questo modo l'utente, utilizzando il catalogo per cercare tramite la segnatura il manoscritto cui è interessato, potrà accedere alla visualizzazione pubblica e gratuita della copia digitale seguendo il link <Visualizza la copia digitale> appositamente inserito all'interno della pagina di dettaglio di ciascun *record* catalografico;
- attraverso la consultazione della suddetta pagina riepilogativa (https://ambrosiana.comperio.it/biblioteca-digitale/), sfogliando idealmente la collezione digitale della Veneranda Biblioteca Ambrosiana tramite la lista dei

manoscritti digitalizzati, peraltro riconoscibili grazie all'icona IIIF.

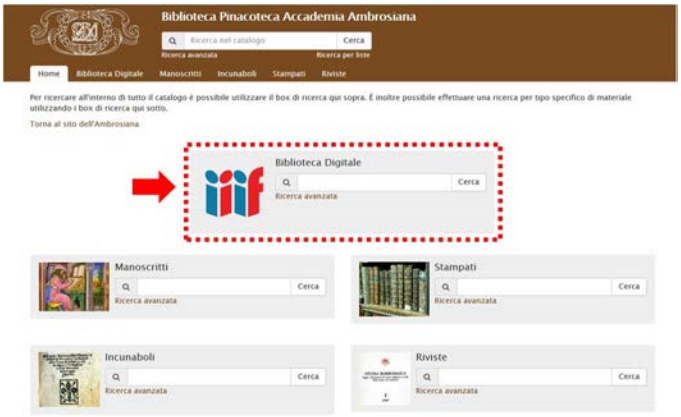


Figura 6: la pagina principale dell'OPAC dell'Ambrosiana con il nuovo riquadro di ricerca dedicato ai manoscritti digitalizzati (https://ambrosiana.comperio.it).

Figura 7: la pagina dell'OPAC dell'Ambrosiana dedicata alla nuova biblioteca digitale (https://ambrosiana.comperio.it/biblioteca-digitale).

Si è anche proceduto a testare le potenzialità di Mirador collegando una porzione di testo trascritto in formato TEI – tratto da un manoscritto della Biblioteca Ambrosiana – alla corrispondente immagine digitale in IIIF dello stesso manoscritto tramite *manifest* .json, il tutto sfruttando le potenzialità dell'*Annotation Tool* integrato nel visualizzatore Mirador (Monella e Cusimano, 2019).

Figura 8: visualizzazione della trascrizione in formato TEI di una porzione del manoscritto digitalizzato grazie all'*Annotation Tool* di Mirador: ms. Ambr. D 23 sup., f. 13v © Veneranda Biblioteca Ambrosiana.

References

IIIF 2018. “About IIIF”, IIIF, accessed November 25, 2019, http://iiif.io/about/;

Annoni, Ada. 1992. “Le Costituzioni e i regolamenti”. In Storia dell’Ambrosiana. Il Seicento, 149-184. Milano: Cariplo;

Barbier, Frédéric. 2005. Storia del libro: dall’antichità al XX secolo. Bari: Dedalo;

Bentivoglio, Francesco (ed.). 1835. Costituzioni del Collegio e della Biblioteca Ambrosiana volgarizzate dal Dottore Francesco Bentivoglio, bibliotecario della medesima, col testo a fronte. Milano: G. B. Bianchi e C.;

Bertolo, Fabio Massimo et al. 2016. Breve storia della scrittura e del libro. Roma: Carocci;

Bottasso, Enzo. 1999. Storia della biblioteca in Italia. Milano: Lampi di stampa;

Braida, Lodovica. 2016. Stampa e cultura in Europa tra XV e XVI secolo. Roma-Bari: Editori Laterza;

Brantl, Markus. 2016. “Das International Image Interoperability Framework (IIIF): Ein neuer Standard für interoperable Bildrepositorien”. Bibliotheksforum Bayern 1/10:10-13;

Burke, Peter. 1992. “L’Ambrosiana e l’Europa del tempo”. In Storia dell’Ambrosiana. Il Seicento, 391-416. Milano: Cariplo-Intesa BCI;

Buzzi, Franco, and Ferro, Roberta (edd.). 2005. Federico Borromeo fondatore della Biblioteca Ambrosiana, Atti delle giornate di studio (Milano, 25-27 novembre 2004). Roma: Bulzoni;

Candela, Leonardo, Castelli, Donatella, Pagano, Pasquale. 2009. “Le biblioteche digitali: origini ed evoluzioni storiche”. DigItalia. Rivista del digitale nei beni culturali 2:36-60. Accessed June 4, 2018. http://digitalia.sbn.it/article/view/277/179;

Cohen, Laura, “A Librarian’s 2.0 Manifesto”, Library 2.0: An Academic’s Perspective, November 8, 2006, accessed November 25, 2019, https://tinyurl.com/ycqssjr7;

Cursi, Marco. 2016. Le forme del libro: dalla tavoletta cerata all’e-book. Bologna: Il Mulino;

Cusimano, Fabio. 2019. “Biblioteche di conservazione & Data Curation: dal Custos catalogi al Digital Librarian. Il caso della Veneranda Biblioteca Ambrosiana”, JLIS 10,1:125-139. Accessed November 25, 2019. http://dx.doi.org/10.4403/jlis.it-12513;

Cusimano, Fabio. Forthcoming. “Medioevo digitale, Medioevo più vicino: il caso della nuova Digital Library ad accesso libero della Veneranda Biblioteca Ambrosiana” (expanded abstract). Proceeding of the International Conference The Middle Ages in the Modern World – MAMO 2018, Rome, 21~24 November 2018. Rome: Collection de l’École française de Rome.

Cusimano, Fabio. Forthcoming. “A ‘cloud’ full of digitized manuscripts: the Veneranda Biblioteca Ambrosiana, from the Custos Catalogi to the Data Curator”. Proceedings of the Conference El’Manuscript 2018, 7th International Conference on Textual Heritage and Information Technologies, Vienna and Krems, Austria, 14-18 September 2018. Sofia: series Scripta & e-Scripta.

De Seta, Ilaria. 2016. “Tre modelli culturali: le biblioteche dei «Promessi sposi»”. In I cantieri dell’italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo. Atti del XVIII congresso dell’ADI – Associazione degli Italianisti (Padova, 10-13 settembre 2014), edited by Guido Baldassarri, Valeria Di Iasio, Giovanni Ferroni, and Ester Pietrobon. Roma: Adi editore, in particolare il paragrafo II. La «realizzazione del suo ideale di cultura»: la biblioteca Ambrosiana, consultabile online, accessed November 25, 2019. http://www.italianisti.it/upload/userfiles/files/DE%20SETA.pdf;

Eisenstein, Elizabeth L. 1986. La rivoluzione inavvertita: la stampa come fattore di mutamento. Bologna: Il Mulino;

Eisenstein, Elizabeth L. 2004. Le rivoluzioni del libro: l’invenzione della stampa e la nascita dell’età moderna. Bologna: Il Mulino;

Emad Isa Saleh. 2018. “Image Embedded Metadata in Cultural Heritage Digital Collections on the Web. An Analytical Study”. In Library Hi Tech 36/1:339-357. Accessed November 25, 2019. https://doi.org/10.1108/LHT-03-2017-0053;

Endres, Bill. 2019. Digitizing Medieval Manuscripts. The St. Chad Gospels, Materiality, Recoveries, and Representation in 2D & 3D, Leeds: Arc Humanities Press; Amsterdam: Amsterdam University Press;

Fabian, Claudia. 2015. Die digitale Renaissance mittelalterlicher Handschriften Aspekte der Erschließung und Digitalisierung - La rinascita digitale dei manoscritti medievali. Catalogazione e digitalizzazione, Lectio Magistralis in Biblioteconomia, Università degli Studi di Firenze, 3 marzo 2015, 55-70. Firenze: Casalini Libri; Febvre, Lucien, Martin, Henri-Jean. 2011. La nascita del libro. Roma-Bari: Editori Laterza;

Galluzzi, Anna. 2011. "Biblioteche pubbliche tra crisi del welfare e beni comuni della conoscenza. Rischi e opportunità". Bibliotime XIV/3. Accessed November 25, 2019. http://www.aib.it/aib/sezioni/emr/bibtime/num-xiv-3/galluzzi.htm;

Ghilli, Carlo, Guerrini, Mauro. 2015. "La Biblioteca Ambrosiana nei «Promessi sposi»". In Biblioteche reali, biblioteche immaginarie. Tracce di libri, luoghi e letture, edited by Anna Dolfi, 365-376. Firenze: Firenze University Press;

Gibson, William. 1990. "Cyberpunk" (Documentary). Directed by Marianne Trench, produced by Peter von Brandenburg, An Intercon Production. Video available in 5 parts on Youtube. Accessed November 25, 2019. https://www.youtube.com/watch?v=xxTuEGEl9EQ;

Gorman, Michael. 2004. La biblioteca come valore. Tecnologia, tradizione e innovazione nell'evoluzione di un servizio. Udine: Forum;

Kempf, Klaus. 2013. Der Sammlungsgedanke im digitalen Zeital;ter - L'idea della collezione nell'età digitale, Lectio Magistralis in Biblioteconomia, Università degli Studi di Firenze, 5 marzo 2013. Firenze: Casalini Libri;

IFLA/UNESCO Public Library Manifesto 1994, accessed November 25, 2019, https://www.ifla.org/publications/iflaunesco-public-library-manifesto-1994;

"IFLA Library Theory and Research Panel, Data Curation Project" 2017, accessed June 6, 2018, https://ifla.wdib.uw.edu.pl/wp-content/uploads/2017/03/LTR-Panel-presentation-on-the-Data-Curation-Project.pdf;

"Library Theory and Research Section", IFLA (International Federation of Library Associations and Institutions), accessed November 25, 2019, https://www.ifla.org/library-theory-and-research;

Lit, Lambertus Willem Cornelis, van. 2020. Among Digitized Manuscripts. Philology, Codicology, Paleography in a Digital World. Leiden-Boston: Brill;

Magnuson, Lauren, "Store and display high resolution images with the International Image Interoperability Framework (IIIF)", ACRL Tech Connect, accessed November 25, 2019, https://acrl.ala.org/techconnect/post/store-and-display-high-resolution-images-with-the-international-image-interoperability-framework-iiif/;

Marcora, Carlo. 1986. "Manoscritti ed edizioni delle 'Constitutiones Collegii ac Bibliothecae Ambrosianae'". In Accademia di San Carlo. Inaugurazione dell'8° Anno Accademico (Milano, 16 novembre 1985), 155-164. Bologna: Cappelli;

McLuhan, Marshall. 2011. La galassia Gutemberg: nascita dell'uomo tipografico. Roma: Armando;
Mirador 2018. "Mirador", accessed November 25, 2019, http://projectmirador.org/;

Monella, Paolo, and Cusimano, Fabio. 2019. "Linking Text and image: TEI XML and IIIF", accessed November 25, 2019, http://www1.unipa.it/paolo.monella/reires2019/;

Montecchi, Giorgio, and Venuda, Fabio. 2006. Manuale di biblioteconomia. Milano: Editrice Bibliografica;

Natale, Maria Teresa. 1995. "Il manifesto UNESCO sulle biblioteche pubbliche". AIB Notizie 7: 1-2;

Naudé, Gabriel. 1627. Aduis pour dresser vne bibliotheque. Presenté à Monseigneur le President de Mesme. Par G. Naudé P. Omnia quae magna sunt atque admirabilia, tempus aliquod quo primum afficerentur habuerunt. Quintil. Lib. 12. A Paris, Chez Francois Targa, au premiere pillier de la grand' Salle du Palais, deuante les Consultations, M.DC.XXVII;

Panizza, Mario. 2012. La storia della Biblioteca Ambrosiana. Novara: De Agostini;

Prodi, Paolo. “Borromeo, Federico”. In Dizionario Biografico degli Italiani, 13, 33-42. Roma: Istituto della Enciclopedia Italiana, 1971. Available also online, accessed November 25, 2019. http://www.treccani.it/enciclopedia/federico-borromeo_(Dizionario-Biografico);

Ravasi, Gianfranco. 1992. “«Federico ideò questa Biblioteca Ambrosiana e la eresse…»”. In Storia dell’Ambrosiana. Il Seicento, 1-19. Milano: Cariplo-Intesa BCI;

Rodella, Massimo. 1992. “Fondazione e organizzazione della Biblioteca”. In Storia dell’Ambrosiana. Il Seicento, 121-147. Milano: Cariplo-Intesa Bci;

Rodella, Massimo. 2013. “Pietro Mazzucchelli (1762-1829) bibliografo ed erudito ambrosiano”. In Frasso, Giuseppe, and Rodella, Massimo, Pietro Mazzuchelli studioso di Dante. Sondaggi e proposte, 35-36. Roma: Edizioni di storia e letteratura;

Roncaglia, Gino. 2010. La quarta rivoluzione: sei lezioni sul futuro del libro. Roma-Bari: Editori Laterza;

Rovelstad, Matilde V. 2000. “Two Seventeenth-Century Library Handbooks, Two Different Library Theories”. Libraries & Culture 35/4: 540-556;

Salarelli, Alberto. 2017. “International Image Interoperability Framework (IIIF): una panoramica”, JLIS 8,1:50-66. Accessed November 25, 2019. http://dx.doi.org/10.4403/jlis.it-12090;

Sardo, Lucia. 2017. La catalogazione: storia, tendenze, problemi aperti. Milano: Editrice Bibliografica;

Serrai, Alfredo. 2005. Angelo Rocca fondatore della prima biblioteca pubblica europea (nel quarto centenario della Biblioteca Angelica), 7-9. Milano: Edizioni Sylvestre Bonnard;

Snydman, Stuart, Sanderson, Robert, Cramer, Tom, 2015. “The International Image Interoperability Framework (IIIF): a community & technology approach for web-based images”. Archiving Conference: 16-21;

Storia dell’Ambrosiana, 4 voll. Milano: Cariplo-Intesa Bci, 1996-2002;

Tammaro, Anna Maria. 2005. “Che cos’è una biblioteca digitale?”. DigItalia. Rivista del digitale nei beni culturali 1:14-33. Accessed November 25, 2019. http://digitalia.sbn.it/article/download/325/215;

Tammaro, Anna Maria. 2008. Biblioteche digitali e scienze umane. Open access e depositi istituzionali, Vol. I. Fiesole: Casalini Libri;

Tammaro, Anna Maria. 2008. Biblioteche digitali e scienze umane. La biblioteca digitale di ricerca per l’apprendimento, Vol. II. Fiesole: Casalini Libri;

Tammaro, Anna Maria. 2011. “Biblioteca digitale co-laboratorio. Verso l’infrastruttura globale per gli studi umanistici”. In Les historiens et l’informatique. Un métier à réinventer, edited by Jean-Philippe Genet, Andrea Zorzi, 11-27. Roma: École française de Rome;

Wondwossen Mulualem Beyene. 2017. “Metadata and Universal Access in Digital Library Environments”. In Library Hi Tech 35/2:210-221. Accessed November 25, 2019. https://doi.org/10.1108/LHT-06-2016-0074;

Wondwossen Mulualem Beyene, Godwin, Thomas. 2018. “Accessible Search and the Role of Metadata”. In Library Hi Tech 36/1:2-17: Accessed November 25, 2019. https://doi.org/10.1108/LHT-08-2017-0170.

Repertori terminologici multilingui fra normatività e uso nella comunicazione istituzionale e professionale[1]

Klara Dankova
Università Cattolica del Sacro Cuore
klara.dankova@unicatt.it

Silvia Calvi
Università degli Studi di Verona
silvia.calvi@univr.it

Abstract

English. This article, after having considered the limits of official multilingual terminology database in the treatment of spontaneous and not-institutionalized terminologies, proposes, through the use of the software *Tedi (Ontoterminology Editor)*, a model for the cataloguing and the publication of these spontaneous terminologies in order to preserve and to enhance a significant cultural heritage. Two case studies, showing examples of spontaneous and not-institutionalized terminologies, will be discussed: the terminology of textile fibres and of climbing. A model of terminology record for the term *polyamide 6* will be described.

Italiano. Il presente articolo, osservando i limiti dei repertori terminologici multilingui istituzionali per quanto riguarda le terminologie in uso ma non ufficializzate, illustra, tramite l'utilizzo del programma *Tedi (Ontoterminology Editor)*, una proposta di catalogazione e divulgazione di queste terminologie spontanee in un'ottica di conservazione e valorizzazione di un patrimonio culturale di grande rilievo. Si approfondiranno due casi di studio di terminologia in uso ma non ufficializzata: la terminologia delle fibre tessili e dell'arrampicata sportiva, illustrando una proposta di progettazione di scheda terminologica per il termine *polyamide 6*.

1 Introduzione

Il presente articolo intende investigare il rapporto tra terminologia e *Digital Humanities* rispetto all'utilizzo di strumenti computazionali per la conservazione e la valorizzazione del patrimonio culturale, veicolato dai termini della comunicazione professionale. La disciplina della terminologia è fin dalle sue origini profondamente legata al trattamento informatico: Wüster, ingegnere elettronico e fondatore di questa disciplina (1931, 1979) considerava infatti la terminologia in un'ottica interdisciplinare ai confini tra linguistica, logica, ontologia e informatica. La presenza di riflessioni di natura informatica negli studi terminologici si intensificò al fine di introdurre nuove metodologie di lavoro che consentissero per la prima volta l'accesso a una enorme quantità di dati da analizzare, raccolti e categorizzati all'interno di risorse digitali. Il trattamento informatico in studi terminologici si arricchì via via nel corso degli anni 1970 e 1980, sia nella fase dell'estrazione di termini, sia in quella della costituzione di repertori terminologici, nella maggior parte dei casi basati sul concetto di *synset* (*set of synonyms*) (Zanola, 2018: 32). A partire dagli anni 1990, la terminologia iniziò ad essere associata all'ontologia, studiando i termini a partire dalla loro dimensione epistemologica e concettuale. Questo percorso di natura onomasiologica diede le basi teoriche per lo sviluppo dell'innovativo approccio introdotto da Christophe Roche[2], basato sul concetto di ontoterminologia, ovvero una terminologia il cui sistema concettuale è un'ontologia formale (Roche, 2012: 2626). In questo studio si è scelto di adottare questo approccio, ritenuto il più appropriato per la rappresentazione dei concetti e la schedatura dei termini che li designano nell'ottica della divulgazione delle terminologie spontanee e non istituzionalizzate, create dai professionisti sul campo o in uso nelle pratiche di gruppi o di precise comunità professionali.

[1]Klara Dankova ha redatto il § 4.1. Silvia Calvi ha redatto i § 1 e 2. I § 3, 4 e 5 sono frutto di una collaborazione delle due autrici.
[2]Le autrici ringraziano il professore Christophe Roche dall'*Université Savoie Mont-Blanc* per la disponibilità ad aver fornito l'accesso al programma *Tedi*, utilizzato ai fini del presente studio.

2 Obiettivi di ricerca

Nel presente articolo si osserveranno i limiti dei repertori terminologici multilingui istituzionali rispetto alla descrizione di terminologie non ufficializzate ma in concreto uso in diversi domini. In un'ottica socioterminologica, si presterà particolare attenzione a come uno stesso concetto possa essere designato da una ricca varietà di termini in funzione del contesto e dell'utente di riferimento.

Alla presentazione delle principali banche dati esistenti, repertori di terminologia ufficializzata – ovvero la terminologia ufficialmente riconosciuta in fonti primarie – seguirà la presentazione di alcuni esempi di terminologie non ufficializzate dei domini delle fibre tessili e dell'arrampicata sportiva. Questi esempi permetteranno di illustrare la varietà terminologica relativa a diversi contesti di utilizzo spesso non menzionati in repertori terminologici istituzionalizzati, lacuna che con la presente proposta si vuole cercare di colmare. Infine, illustrando un caso tratto dal dominio delle fibre tessili, si proporrà una metodologia per la progettazione di una banca dati, costituita da schede terminologiche multilingui.

Obiettivi del presente studio sono quindi:

- dimostrare che la terminologia spontanea deve essere conservata e ufficializzata, in quanto portatrice di un significativo patrimonio culturale, che consente di avvicinarsi anche al mestiere di riferimento per il quale la terminologia diviene custode del relativo saper-fare;
- riflettere sulla rappresentazione e sulla divulgazione dei termini individuati, prestando particolare attenzione alla loro dimensione sociale e interculturale;
- proporre un modello di realizzazione di schede terminologiche multilingui attraverso l'utilizzo di *Tedi, onto**T**erminology **EDI**tor*, software realizzato per progettare ontoterminologie multilingui (Roche, 2007).

3 Banche dati terminologiche e la comunicazione istituzionale e professionale

Le banche dati terminologiche sono strumenti digitali che raccolgono le informazioni sui termini in una o più lingue, appartenenti a più settori e le presentano sotto forma di schede terminologiche redatte in modo standardizzato al fine di permettere la maggior condivisione dei dati raccolti. Diverse sono le informazioni fornite in una scheda terminologica, per esempio il termine, la marcatura morfologica, la definizione, eventuali sinonimi, il contesto di utilizzo e le note enciclopediche. Il beneficio maggiore dell'utilizzo dei termini definiti in modo univoco è la possibilità di comunicare in modo chiaro e preciso, indipendentemente dai soggetti coinvolti (Zanola, 2018: 64).

Il bisogno di costruire una banca dati terminologica nasce prima di tutto nei contesti multilingui per rispondere alle necessità della pubblica amministrazione. Un modello efficiente di catalogazione dei dati terminologici è fornito dalle banche dati canadesi, dal *Grand dictionnaire terminologique* (GDT) dell*'Office québécois de la langue française* e da *Termium Plus* gestito dal *Bureau de la traduction* del governo canadese. L'urgenza di standardizzare l'uso dei termini si manifesta anche nel contesto europeo, in particolare nelle istituzioni dell'Unione Europea: la progettazione di *IATE* (*InterActive Terminology for Europe*) ha portato alla costituzione di una banca dati contenente soprattutto i termini usati nei testi legislativi e amministrativi, pubblicati dalle varie istituzioni dell'Unione Europea. Inoltre, ci sono anche delle banche dati terminologiche che operano esclusivamente a livello nazionale, come per esempio *Termdat*, la raccolta terminologica della Confederazione svizzera (Zanola, 2018: 64-67). Va sottolineato che queste banche dati sono state costruite pensando a un gruppo di utenti ben preciso (i cittadini quebecchesi, dell'Unione Europea, svizzeri ecc.). I termini sono stati individuati all'interno di un contesto specifico, quale il contesto giuridico e amministrativo dell'UE o del Canada, e, di conseguenza, non possono essere sempre utilizzati nei testi relativi ad altre realtà socio-culturali.

Tuttavia, accanto ai termini recensiti e definiti in un contesto istituzionale esistono anche terminologie spontanee utilizzate nella comunicazione professionale, che trovano solo un parziale riscontro nelle banche dati ufficiali, le quali spesso trascurano le variazioni diastratiche. Consideriamo, per esempio, la varietà di termini in francese usati da vari gruppi di persone con riferimento alla fibra acrilica: gli ingegneri chimici useranno probabilmente una denominazione chimica che rivela la composizione della fibra (*polyacrylonitrile*), nella comunicazione tra professionisti in una fiera verrà invece più facilmente utilizzato il codice (*PAN*), mentre nell'ambito della moda sarà più frequente un termine più generico (*acrylique* o *fibre acrylique*). Nell'attuale contesto, il bisogno di disporre di raccolte multilingui di termini a uso dei professionisti di vari settori diventa sempre più forte, rendendo necessario lo sviluppo di un nuovo modello di catalogazione.

3.1. La terminologia non istituzionalizzata: il caso delle fibre tessili e dell'arrampicata sportiva

I termini che designano le fibre tessili sono stati estratti manualmente da un corpus di testi in lingua francese, contenente quattro tipi di fonti: dei cataloghi delle fiere (Première Vision Yarns, 12.02.-14.02 2019, Première Vision Fabrics, 12.02.-14.02 2019), un documento istituzionale (DGE/UBIFRANCE, 2006), un'opera di divulgazione (Fauque e Bramel, 1999) e un manuale tecnico (Weidmann, 2010). Il corpus risultante è costituito da 245 termini, di cui 60 sono nomi generici e 185 nomi di marca. Si è osservato che la terminologia delle fibre tessili è molto complessa sia per le differenze culturali sia per le sue variazioni diastratiche. Quanto alla dimensione culturale, si possono riscontrare delle differenze tra i termini usati in contesti diversi. Infatti, esistono alcuni casi, in cui i termini usati per designare un determinato concetto differiscono da un paese all'altro, anche all'interno di una stessa lingua. Si considerino i termini in francese utilizzati per designare la fibra di elastan: mentre nei paesi dell'UE si usa *élasthanne*, il termine corrispondente negli Stati Uniti è *spandex*, in Giappone *polyuréthane* e in Cina sono in uso i termini *élasthanne* o *spandex* (ISO 2076: 2013). Per quanto riguarda le differenze tra terminologia istituzionale e professionale, si può notare che alcuni termini esclusi dall'uso nella comunicazione istituzionale di un paese possono continuare ad essere usati tra gli esperti del settore. Per esempio, i termini *fibranne* e *rayonne*, che designano in francese rispettivamente le fibre discontinue e i filamenti continui di viscosa, sono stati sostituiti nel 1976 nella comunicazione istituzionale francese dal termine *viscosa* (Browaeys, 2014: 18; Baum e Boyeldieu, 2018: 256). Questo non impedisce però che vengano occasionalmente utilizzati dagli esperti del settore, per mettere in evidenza la differenza tra le due forme della fibra di viscosa[3]. A proposito dei nomi di marca bisogna mettere in evidenza che, anche se designano lo stesso tipo di fibra (la poliammide 6.6), la loro composizione è diversa: mentre *Nylon* è una fibra di poliammide 6.6 convenzionale, *Ultron* presenta delle caratteristiche antistatiche e *Sylkharesse* è un materiale prodotto in forma di microfibra. Infine, nella terminologia delle fibre tessili si riscontra una varietà di termini che designano lo stesso concetto, ma comunque non possono spesso essere usati nello stesso contesto. Nel caso di *Nylon* possiamo individuare le seguenti tipologie di termini: 1) denominazione chimica (*poliesametilenadipamide*), 2) nome generico (*poliammide 6.6*), 3) nome generico istituzionale (nei paesi dell'UE *poliammide* o *nylon*), 4) codice indicato sull'etichetta di composizione (*PA*), 5) codice usato tra i professionisti (*PA 6.6*), 6) nome di laboratorio (*Fibre 6.6*), 7) nome di marca (*Nylon*).

Differenze culturali e variazioni diastratiche possono essere osservate anche nella terminologia dell'arrampicata sportiva, sport antico che tuttavia ha solo recentemente ottenuto un riconoscimento ufficiale, quale l'introduzione tra i nuovi sport olimpici di Tokyo 2020, e che giunge a fissare definitivamente per questa ragione i propri usi terminologici. Nello studio condotto per l'arrampicata sportiva sono stati estratti manualmente 96 termini a partire da un corpus eterogeno in lingua italiana composto da manuali di arrampicata (Commissione Nazionale Scuole di Alpinismo e Arrampicata Libera della Commissione Centrale per le pubblicazioni, 2009; Bressa, Denicu, Capretta, 2010; Ponta, 2016), documenti pubblicati da enti ufficiali quali C.A.I (Club Alpino Italiano) e F.A.S.I. (Federazione Arrampicata Sportiva Italiana), articoli di riviste specializzate come *Montagna 360°*, la rivista ufficiale del C.A.I. Trattandosi di una realtà internazionale in cui il confronto tra esperti in occasione di gare e manifestazioni è all'ordine del giorno, è interessante osservare come le differenze culturali tra i termini individuati siano poche e prevalentemente legate alle scale utilizzate per misurare i gradi di difficoltà e ai prodotti che possono essere messi in commercio in forme e dimensioni diverse da paese a paese, per esempio mentre in Italia la *magnesite* può essere acquistata in forma *granulosa*, non è stato trovato un equivalente nelle fonti canadesi, in cui tale prodotto sembra essere acquistato prevalentemente in formati differenti. Quanto alla variazione diastratica si può osservare come il termine *arrampicata su massi* non venga spesso utilizzato nelle fonti ufficiali che prediligono invece mantenere il termine internazionale *boulder* per agevolare la comunicazione tra professionisti provenienti da realtà culturali differenti. Inoltre, si può constatare che per questa terminologia la rappresentazione visiva degli oggetti e delle tecniche di arrampicata è di grande importanza per la comprensione dei concetti e deve perciò essere presa in considerazione in fase di stesura delle rispettive schede terminologiche.

4 La progettazione di ontologie in prodotti terminologici

La rappresentazione dei concetti e la schedatura dei termini richiede la comprensione dell'organizzazione concettuale del dominio oggetto di studio, attraverso la progettazione di ontologie formali. Diversi sono i

[3] Si veda per es. Daniel Weidmann. 2010. *Aide-mémoire textiles techniques*. Dunod, Paris, p. 64.

programmi attualmente disponibili che permettono questa operazione, tra cui il *Lexicon Model for Ontologies (Lemon)* modello sviluppato dalla *Ontology Lexicon Community* il cui principale obiettivo è la presentazione di informazioni di natura linguistica all'interno di ontologie (McCrae *et al*. 2017); *Protégé* programma realizzato dallo *Stanford Center for Biomedical Informatics Research* per supportare il OWL 2 Web Ontology Language (Tudorache *et al.*, 2013); *Tedi* programma proposto dal *Condillac Research Group in Knowledge Engineering* per la progettazione di ontoterminologie multilingui[4].

Ai fini del presente studio si è scelto di utilizzare il programma che meglio rispecchia la natura della disciplina terminologica, intesa come ontoterminologia: ovvero *Tedi*, programma il cui punto di partenza non è la dimensione linguistica del termine, come avviene per il *Lexicon Model for Ontologies* quanto la sua dimensione nozionale e concettuale. La decisione di prediligere *Tedi* rispetto a *Protégé* è invece giustificata dal fatto che nel primo la distinzione termine-concetto è più immediata in particolare in ottica di una rappresentazione terminologica multilingue.

Un esempio tratto dall'ambito delle fibre tessili consentirà di illustrare una proposta di metodologia di lavoro da adottare per la realizzazione di schede terminologiche basate su un'ontologia formale.

4. 1. *Tedi*, una proposta per la realizzazione di un'ontologia con schede terminologiche multilingui. La terminologia delle fibre tessili: il caso del termine *polyamide 6*

L'editore di ontoterminologie *Tedi* si basa sulla distinzione della terminologia in due dimensioni: 1) la dimensione concettuale extralinguistica, condivisa dalle diverse comunità linguistiche 2) la dimensione linguistica, composta da diversi sistemi lessicali. Le due dimensioni sono strettamente legate, poiché i termini rappresentano i nomi dei concetti in lingua naturale (Roche, 2019: 5). La distinzione delle due dimensioni, permettendo una migliore comprensione del dominio, consente anche di effettuare delle ricerche non soltanto in base alle relazioni linguistiche tra i termini (iperonimia, sinonimia), ma anche in base alle relazioni logiche tra i concetti (concetto generico, concetto specifico) (Roche *et al.*, 2014: 2).

4.2 *Tedi* e la dimensione concettuale dell'ontoterminologia

Per ricostruire il sistema concettuale del dominio, *Tedi* mette a disposizione dell'utente il *concept editor*. Questo editor permette di definire i concetti in un linguaggio formale, che consiste nell'indicazione delle caratteristiche essenziali del concetto, dette anche "differenze", in quanto rappresentano una delle possibilità di realizzazione di una certa caratteristica, predefinita secondo l'asse dell'analisi corrispondente. A titolo di esempio, l'asse dell'analisi "origine della fibra" fornisce due caratteristiche essenziali "naturale" e "chimica". Nel caso del concetto <poliammide 6> (Fig. 1), l'utente definisce l'origine della fibra scegliendo la caratteristica essenziale "chimica".

[4] Si veda: http://new.condillac.org/projects/tedi

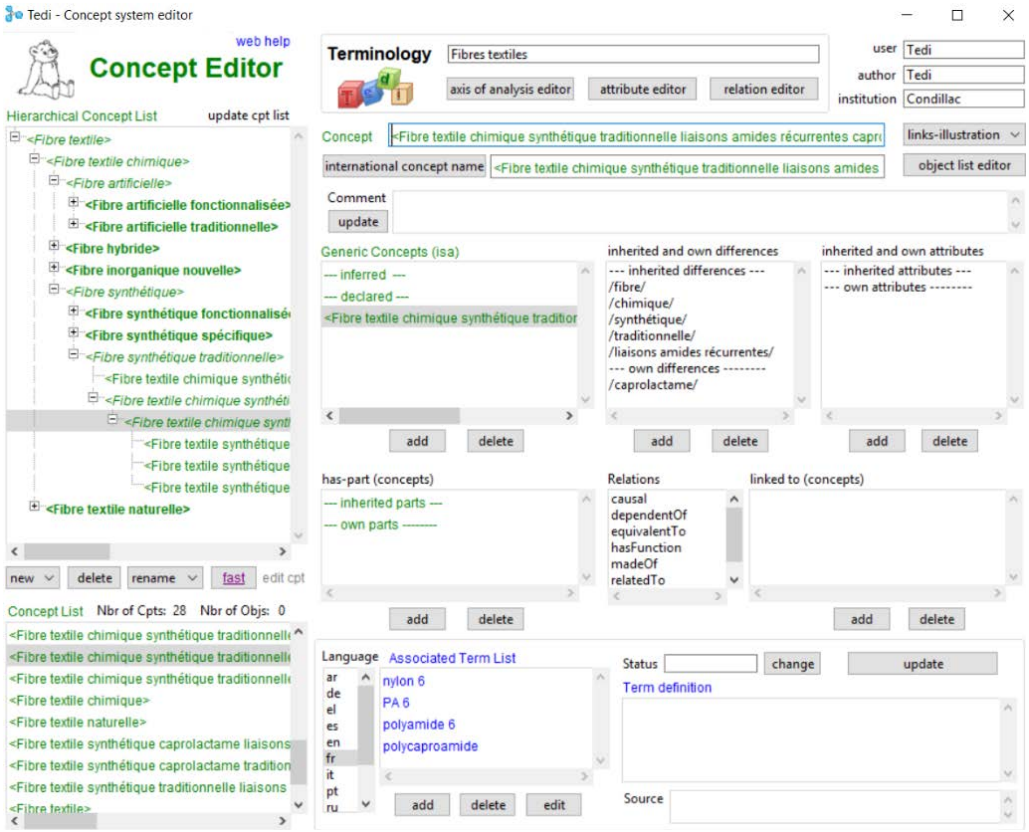


Figura 1: Il concetto <poliammide 6> definito nel *concept editor* di *Tedi*

Una volta identificate le caratteristiche essenziali del concetto, il sistema genera in funzione di esse il nome del concetto, che permette di comprendere la natura degli oggetti che rientrano sotto il concetto stesso (es. il concetto <poliammide 6> viene denominato < Fibre textile chimique synthétique traditionnelle liaisons amides récurrentes caprolactame>). In seguito, l'utente inserisce il concetto nella rete di relazioni tra i concetti del sistema, indicando il suo concetto generico (il concetto <poliammide>: <Fibre textile chimique synthétique traditionnelle liaisons amides récurrentes>) e, eventualmente, i suoi concetti specifici (es. il concetto <Lilion>: <Fibre textile synthétique caprolactame traditionnelle chimique liaisons amides récurrentes société Snia Viscosa>). In questo modo, si ricostruisce l'organizzazione concettuale del dominio, che viene visualizzata nel *concept editor* di *Tedi* nell'angolo in alto a sinistra (vedi Fig. 1). Nel caso di alcune terminologie, quali quella dell'arrampicata sportiva, un ruolo importante nella comprensione del concetto è svolto dalla sua rappresentazione visiva. Per venire incontro a questa esigenza, *Tedi* è dotato della funzione *link-illustration* che consente di associare al concetto non solo immagini, ma anche video e collegamenti ipertestuali.

4.3 Creazione di una scheda terminologica in *Tedi*: la dimensione linguistica

I termini vengono definiti nel *term editor* che propone degli editor indipendenti per una serie di lingue (es. francese, italiano, inglese). La definizione del termine si inserisce nella lingua naturale, con la possibilità di utilizzare un modello di definizione, elaborato da *Tedi* in base alla definizione formale del concetto. Oltre alla definizione del termine, *Tedi* consente di inserire altri dati relativi al termine, quali la fonte della definizione, l'informazione morfologica, lo status del termine ("preferenziale", "alternativo", "obsoleto"), il suo contesto di utilizzo, le note enciclopediche, le varianti ortografiche e le forme flesse. La sezione note enciclopediche è di particolare interesse per la terminologia non istituzionalizzata in quanto in questa sezione è possibile presentare sia delle note di carattere culturale sia delle indicazioni circa la variazione diastratica. Nel caso delle fibre tessili per esempio si può indicare se il termine oggetto di studio è connotato culturalmente e se esso si riferisca alla denominazione chimica, al nome di laboratorio o al suo codice. Inoltre, nel *term editor* appaiono in modo automatico anche gli equivalenti, gli iperonimi, gli iponimi e i sinonimi del termine, recensiti nella banca dati e associati a un concetto definito nel linguaggio formale a sua volta in relazione con il concetto designato dal termine in questione (Fig. 2).

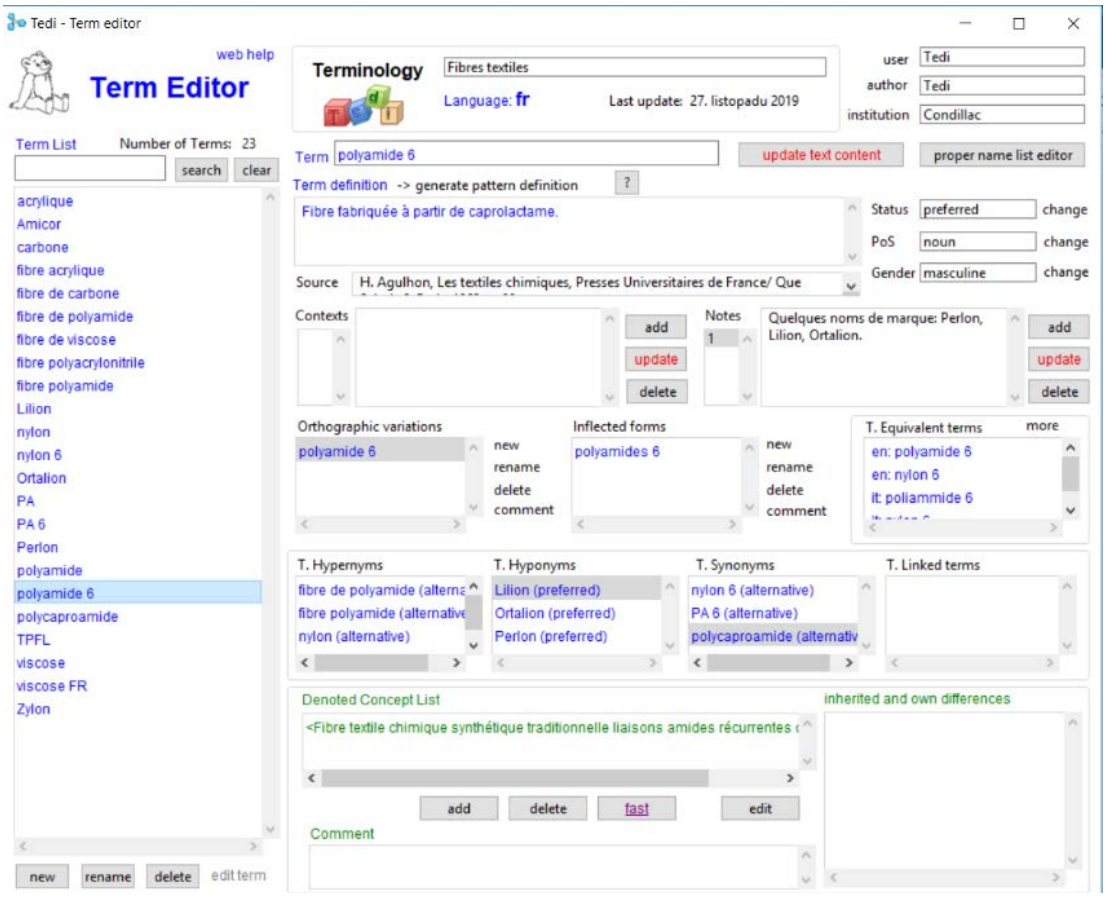


Figura 2: Il termine *polyamide* 6 definito nel *term editor* di lingua francese di *Tedi*

I termini equivalenti in varie lingue (per esempio *polyamide* 6 (fr), *poliammide 6* (it), *polyamide* 6 (en)) sono associati a un unico concetto, definito in un linguaggio formale e quindi extralinguistico, il che permette di creare una banca dati terminologica multilingue, contenente una rappresentazione delle relazioni tra i concetti, che può essere visualizzata esportando i dati presenti nel *concept editor* in altri programmi, quali *Cmap Tools*, *Protégé*. Inoltre, l'esportazione delle informazioni sui termini in una lingua (nel nostro caso in francese) in formato HTML consente la progettazione di un dizionario elettronico, composto dalle schede terminologiche che forniscono diverse indicazioni tra cui la definizione del termine in lingua naturale, gli equivalenti in altre lingue e l'elenco delle caratteristiche essenziali del concetto (Fig. 3).

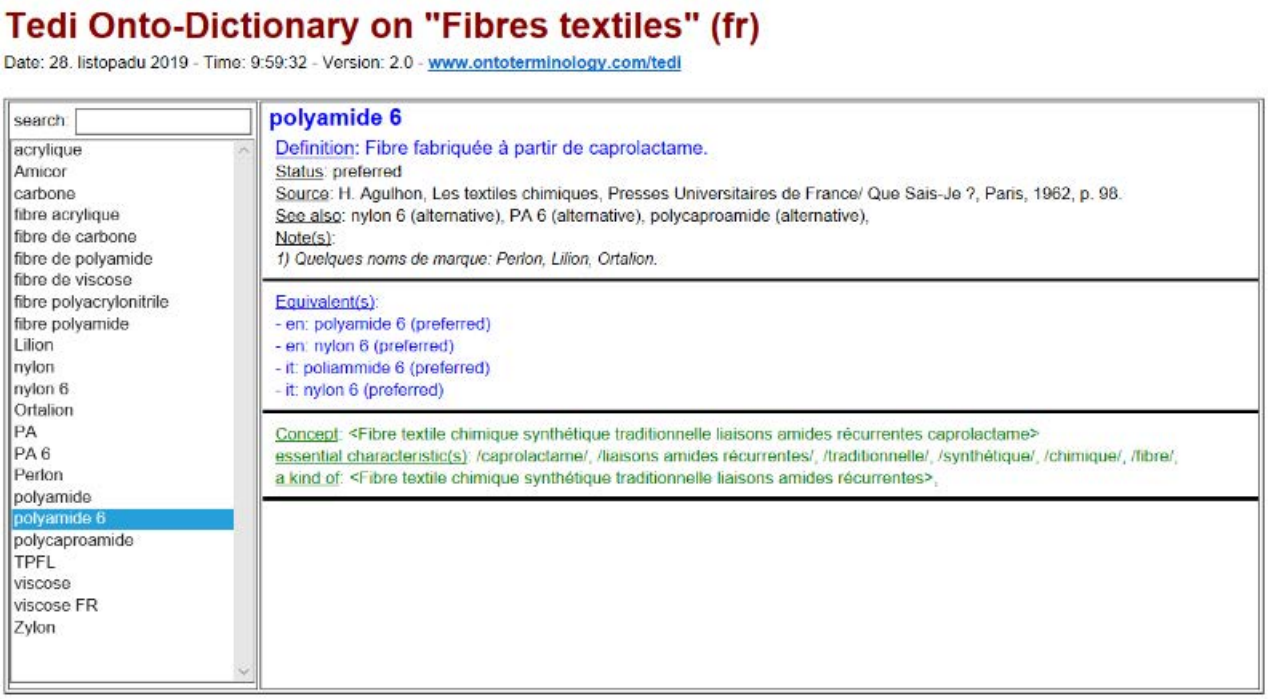


Figura 3: Il dizionario elettronico: la scheda terminologica di *polyamide* 6 (fr)

5 Conclusioni

Disponendo attualmente delle banche dati terminologiche di carattere istituzionale o nazionale, ci si trova di fronte a un nuovo bisogno, ossia quello di progettare un ricco repertorio terminologico digitale, facilmente utilizzabile nella comunicazione professionale multilingue. È necessario tenere in considerazione il fatto che i bisogni dei professionisti di un settore differiscono spesso da quelli dei principali destinatari delle banche dati fornite dalle istituzioni (per esempio traduttori, legislatori, giuristi, redattori di testi). Come dimostrato dalla proposta illustrata nel presente articolo, l'utilizzo del programma *Tedi*, basato su un approccio ontoterminologico, permetterà quindi di progettare delle innovative banche dati terminologiche, attente alla dimensione socio-culturale della terminologia spontanea e in uso nella comunicazione professionale. Questo percorso, applicabile a più domini, permetterà quindi di conservare e ufficializzare un ricco patrimonio terminologico che, finora, non ha goduto dello stesso trattamento della terminologia istituzionalizzata.

Bibliografia

Maggy Baum et Chantal Boyeldieu. 2018. *Dictionnaire encyclopédique des textiles*. Eyrolles, Paris.

Nicoletta Bressa, Bruno Capretta, Gian Pietro Denicu e Sandro Neri. 2010. *Dall'arrampicare all'arrampicata: tra spontaneità e tecnica*. Calzetti Mariucci, Torgiano.

Christine Browaeys. 2014. *Les enjeux des nouveaux matériaux textiles*. EDP Sciences, Les Ulis.

Commissione Nazionale Scuole di Alpinismo, Sci Alpinismo e Arrampicata Libera della Commissione Centrale per le pubblicazioni (eds). 2009. *Manuale di arrampicata: arrampicata e allenamento*. C.A.I., Milano.

DGE/UBIFRANCE. 2006. *Textiles Techniques. Le futur se tisse en France*. 1-24 : https://www.entreprises.gouv.fr/files/files/directions_services/secteurs-professionnels/etudes/textileF.pdf

Claude Fauque et Sophie Bramel. 1999. *Une seconde peau : fibres et textiles d'aujourd'hui*. Éditions Alternatives, Paris. ISO (2076: 2013) *Textiles — Man-made fibres — Generic names.*

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar and Philipp Cimiano. 2017. *The Ontolex-Lemon model: development and applications*. Proceedings of eLex 2017 conference: 19-21. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

Angelo Ponta (eds). 2016. Walter Bonatti. Il sogno verticale: cronache, immagini e taccuini inediti di montagna. Rizzoli, Milano.

Christophe Roche. 2007. *Le terme et le concept : fondements d'une ontoterminologie*. TOTh 2007 : Terminologie et Ontologie : Théories et Applications, Jun 2007, Annecy, France : 1-22. https://hal.archives-ouvertes.fr/hal-00202645/document

Christophe Roche. 2012. *Ontoterminology: how to unify terminology and ontology into a single paradigm*. LREC 2012, Eighth international conference on Language Resources and Evaluation. Istanbul, Turkey: 2626-2630. http://ontologia.fr/Bibliographie/567_Paper_Header.pdf

Christophe Roche, Luc Damas and Julien Roche. 2014. *Multilingual Thesaurus: The Ontoterminology Approach*. CIDOC 2014 - Access and Understanding – Networking in the Digital Era, CIDOC (Comité International pour la DOCumentation), Sep. 2014, Dresden, Germany: 1-14. https://hal.archives-ouvertes.fr/hal-01272725/document

Christophe Roche. 2019. *Tedi : ontoterminology editor. Manuel Utilisateur*. Christophe Roche, Le Bourget du Lac.

Tania Tudorache, Csongor Nyulas, Natalya F. Noy and Mark Musen. 2013. *WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web*. Semantic Web Journal, Volume 4, Number 1/2013. IOS Press: 89-99. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691821/pdf/nihms373006.pdf

Daniel Weidmann. 2010. *Aide-mémoire textiles techniques*. Dunod, Paris.

Eugen Wüster. 1931. *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik. Die nationale Sprachnormung und ihre Verallgemeinerung*. VDI, Berlin.

Eugen Wüster. 1979. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Springer, Wien.

Maria Teresa Zanola. 2018. *Che cos'è la terminologia*. Carocci, Roma.

Sitografia

C.A.I. [Club Alpino Italiano]. https://www.cai.it

Condillac Research Group. Terminology and Ontology. Tedi. http://new.condillac.org/projects/tedi.

F.A.S.I. [Federazione Arrampicata Sportiva Italiana]: http://www.federclimb.it/

GDT [Le Grand dictionnaire terminologique]. http://www.granddictionnaire.com/

IATE [Interactive Terminology for Europe]. https://iate.europa.eu/

Protégé [Ontology editor and framework for building intelligent systems]. https://protege.stanford.edu/

TERMDAT [La banca dati terminologica dell'Amministrazione svizzera]. https://www.termdat.bk.admin.ch/Search/Search

Termium Plus [La banque de données terminologiques et linguistiques du gouvernement du Canada]. https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html

The Digital *Lexicon Translaticium Latinum*: Theoretical and Methodological Issues

Chiara Fedriani
University of Genoa
chiara.fedriani@unige.it

Irene De Felice
University of Genoa
irene.defelice@edu.unige.it

William Michael Short
University of Exeter
w.short@exeter.ac.uk

Abstract

English. In this paper, we present some theoretical and methodological issues involved with the creation of the *Lexicon Translaticium Latinum*, a new digital resource for the study of Latin metaphors. This resource is based on the ontology of the Latin WordNet '2.0' (https://latinwordnet.exeter.ac.uk) but extends this specification so as to be able to capture the kinds of large-scale metaphorical patterns that are becoming increasingly documented in Latin's semantic system. In particular, we discuss 1) the theory and method underpinning the revised and expanded ontology; 2) the tagset adopted for classifying metaphors and their relations; and 3) the procedure for annotating conceptual metaphors and linking these with other semantic structures within the WordNet (synsets). In line with the recognition in the 'second wave' cognitive sciences that metaphor is a fundamental mechanism of human cognition (as well as key to the structuring of cultural understanding), our aim is to establish the Lexicon not merely as a comprehensive repository of figurative usage in Latin, but as an accurate model of the conceptual system that its speakers relied upon in thinking, speaking and behaving in diverse contexts of symbolic expression.

Italiano. In questo articolo vengono presentati gli assunti teorici e metodologici alla base del *Lexicon Translaticium Latinum*, una nuova risorsa digitale per lo studio della metafora nella lingua latina, che si basa su Latin WordNet "2.0", un'ontologia semantica della lingua latina liberamente accessibile dal web (https://latinwordnet.exeter.ac.uk), ma estende significati-vamente i dati in esso contenuti rappresentando gli schemi metaforici documentati nel sistema semantico latino a diversi livelli della struttura lessicale. In particolare, si presentano: 1) gli assunti teorici e metodologici alla base della progettazione e della realizzazione dell'ontologia revisionata ad ampliata; 2) il tagset adottato per classificare le metafore e le loro relazioni; 3) la procedura per annotare le metafore concettuali e per collegarle con la struttura semantica dei synset coinvolti. Riconoscendo nella metafora, in linea con le scienze cognitive della "seconda ondata", un meccanismo fondamentale della cognizione umana, sia nella concettualizzazione delle esperienze che nella strutturazione di categorie culturali, il nostro obiettivo è quello di mostrare come il *Lexicon* non sia da intendersi come un mero archivio degli usi figurati attestati in latino, ma piuttosto come un accurato modello di rappresentazione del sistema concettuale alla base del pensiero e del comportamento, anche linguistico, dei parlanti di questa lingua antica.

1 Introduction

The *Lexicon Translaticium Graecum et Latinum* is a collaborative international project aimed at developing an on-line, extensible, open-access lexicon of metaphors in the ancient languages – beginning, in reverse chronological order, with Latin. Unlike existing electronic dictionaries for Latin, which simply re-create their printed counterparts in machine-readable form, the *Lexicon Translaticium* incorporates

insights from up-to-date theories of meaning and, in particular, the view developed in cognitive linguistics of metaphor as a key structuring device of language and thought. In capturing deeply entrenched and highly conventionalized metaphoric and metonymic patterns that organize meanings pervasively throughout this language and at different orders of linguistic encoding, the *Lexicon Translaticium* is meant as a psychologically realistic model of the conceptual system underpinning Latin. Built on top of the ontology provided by the Latin WordNet, the *Lexicon* will be interoperable with existing electronic corpora and thus capable of delivering rich figurative data for integration into natural language processing applications. The project, directed by William Short and Chiara Fedriani and staffed by an international five-member team, is currently on-going. However, its intellectual rationalization is well established and its technical design and implementation have progressed to the point where preliminary 'test' data is already publicly available, with about 20 metaphors presently annotated. We aim to launch the *Lexicon* officially in Spring of 2020 with a fuller dataset consisting of about 100 conceptual metaphors.

2 Theoretical background

Recognizing the all-pervasive character of certain metaphorical patterns in language, George Lakoff and Mark Johnson (1980) argued that the frequent clustering of metaphorical linguistic expressions around abstract, intellectual or otherwise intangible concepts in fact reflects the inherently metaphorical workings of cognition itself. People talk about most abstract concepts metaphorically, that is, because – it is claimed – they actually conceive of them metaphorically in terms of other (usually more concrete) concepts. On this view, metaphors are the projections of conceptual structure and content from one domain to another that occur as a way of mentally representing and reasoning about experiences not directly grounded in the physico-spatial world. Cognitive linguists argue, moreover, that it is the systematic nature of metaphors – in other words, that metaphors characterize regular mappings between organized domains of knowledge – that allows people to think and reason (and therefore also to speak) meaningfully about experiences that may be difficult to comprehend in and of themselves.

In line with this theory, the 'entries' of our metaphor dictionary – unlike those of a traditional lexicon – will therefore consist of large-scale patterns of metaphorical understanding that link together *concepts*, rather than the semantic structures of words *per se* and so structure the meanings of words across the lexicon and at different levels of linguistic encoding. The kinds of metaphors that constitute the data of our *Lexicon* are those that are so conventionalized and so entrenched in the shared linguistic and cognitive habits of Latin speakers that they seem not to have been perceived as figurative at all – and indeed deliver Latin speakers' entirely regular, 'everyday' ways of conceptualizing certain experiences. Of course, as 'imaginative' or 'creative' (or more narrowly 'literary') metaphors most often derive in some way from more conventionalized metaphors, these kinds will also be represented. In this way, the *Lexicon Translaticium Latinum* will form a comprehensive catalogue of the range of metaphorical themes that structure meaning in Latin.

3 Technical implementation

Technically, the *Lexicon* will be realized as a computerized relational database, whose data model combines aspects of the architecture of the MetaNet Project of the International Computer Science Institute in Berkeley, California, with the WordNet framework. The Berkeley MetaNet is an electronic repository, viewable interactively on the Internet as a Wiki, that contains records for hundreds of attested conventional and imaginative metaphors in English, including time metaphors, mind metaphors, and emotion metaphors, as well as metaphors relating to government, disease, and violence. Most importantly for our purposes, the MetaNet provides a set of high-level ontologies for annotating and organizing figurative language data under the theory of conceptual metaphor in cognitive linguistics. In particular, the MetaNet provides a theoretically-grounded formal specification for encoding *kinds* of conceptual metaphors as well as *kinds of relations* between metaphors. For example, the 'type' of a metaphor can be tagged with values such as 'primary', 'composed', or 'entailed', which correspond to well defined theoretical categories. A primary metaphor is one that emerges directly from correlations in experience, as in MORE IS UP or PURPOSES ARE DESTINATIONS, while complex metaphors are those built up out of at

least two more basic primary ones. Entailed metaphors are specialized submappings that can be inferred through experiential knowledge from a primary or complex metaphor, and which often form the basis of coherence between metaphors.

Likewise, metaphors can be organized into hierarchies through simple relations of super- or sub-ordination, or into more intricate systems according to different kinds of (again theoretically grounded) relationships, such as 'extension' (where one mapping takes advantage of conceptual material left unused by another), 'elaboration' (where one mapping embellishes another with additional conceptual material), 'combination', or 'questioning'. 'Reciprocality' is another common feature of metaphor systems and is available to capture 'orientational' metaphors that involve body-based experiential polarities such UP *vs* DOWN, LEFT *vs* RIGHT, CENTER *vs* PERIPHERY, IN *vs* OUT, and so on.

Whereas the MetaNet specification provides the foundation for encoding metaphors (as mappings between concepts) and their relations, the ontologies and data structures of the WordNet deliver the core repertoire of concepts that participate in these relations. As a semantic database, the WordNet represents lexical meaning in terms of *synsets*, which are uniquely identifiable 'definitions' for hypothetically all the senses capable of being expressed in a given language (thus organizing the lexicon into discrete 'synonym sets'). In other words, a WordNet synset – which pairs a unique identifier, consisting of a part-of-speech tag and a string of between six and eight integers, with a descriptive gloss and possibly higher-order 'domain'-level tags – should be seen as representing a distinct *concept* that may constitute the meaning of a word or words in the language under scrutiny. A WordNet for Latin was developed by Stefano Minozzi for the Fondazione Bruno Kessler's MultiWordNet Project (see Minozzi, 2008), consisting of about 9,000 lemmas tagged with synsets drawn from English and Italian. This is now being expanded through an international collaboration directed by the University of Exeter, to include over 70,000 words covering the archaic through classical periods of this language, as well as language-specific synsets defining meanings that are peculiar to Latin and not represented among the 100,000 or so synsets originally defined for English.

4 Innovations of design

Because the Latin WordNet (and indeed the WordNet specification generally) does not presently distinguish between literal and figurative sense attributions, it is being re-architected to accommodate the encoding of metonymic and metaphoric as well as literal senses of words. Annotation *at the level of the lemma* of specific sense (synset) assignments as being either literal, metonymic, or metaphorical is in fact one of the major new 'layers' at which figurative information is represented within the *Lexicon*. Consider, for example, the database entry for the word *baculum*, which can be accessible and marked-up by project participants through our bespoke on-line curation and annotation interface. In classical Latin, this word meant 'walking stick' and thus has been tagged with synset n#03585559, 'a stick carried in the hand for support in walking' as one of its literal senses (and indeed also its prototypical sense). Over time, however, and particularly in the early Christian period, the word came to be used more abstractly in the sense of any 'support' and in ecclesiastical texts regularly exhibits this meaning. This chronologically circumscribed figurative meaning of the word (n#04399253, 'something providing immaterial support or assistance to a person or cause or interest') is therefore annotated as a metaphorical sense. Differentiating between literal, metonymic, and metaphorical signification introduces an entirely new dimension of semantic structure into the WordNet framework, validated by modern linguistic theory.

Along with annotations at the level of lexical semantic structure distinguishing between a word's literal, metonymic, and metaphorical senses (represented by synsets), conceptual metaphors themselves will be coded as a relationship between synsets, understood as discrete concepts. For example, the FEAR IS A WEAPON metaphor, known in Latin in expressions such as the one in (1), is represented as a mapping between the synset that means 'fear' (n#05590260) and the one that means 'weapon' (n#03601056). In turn, the ANXIETY IS A SUBSTANCE metaphor, again illustrated by the passage in (1), is structured as a mapping between the synsets meaning 'anxiety' (n#04491326) and 'substance' (n#00010572), respectively.

1. *ipsius regis non tam subito pavore perculit pectus, quam anxiis inplevit curis* (LIV. 1, 56) 'As for the

king himself, his heart was not so much struck with sudden terror as filled with anxious forebodings'

Accordingly, any lemma annotated with one of these synsets as a literal, metonymic, or metaphorical sense is automatically linked (and accessible) via the metaphor by virtue of those sense attributions. In other words, as the theory posits, the metaphor operates as a *supralexical* structuring device of meaning in Latin: it helps determine, and motivate, the specific semantic developments of words and explains why the vocabulary of 'weapons' (not only the word corresponding to *weapon* but the whole conceptual domain relating to weapons and their use) can be used to talk about FEAR. Without the conceptual metaphor, there is no way to explain why WEAPON concepts are so regularly used to represent FEAR concepts and these would have to remain isolated, and – worse – arbitrary – facts of Latin's semantics. Crucially, moreover, the layer of more global conceptual-metaphorical information is tightly integrated with the more local layer of lexical-semantic information. In other words, the two layers of annotation – 1) the conceptual metaphor itself, as a mapping between synsets (concepts) and 2) the attribution of synsets to lemmas as specifically *metaphorical* senses – work hand in hand. When a lemma is tagged as 'having' a synset as one of its literal, metonymic, or metaphorical sense, the annotator is also able to indicate the specific metaphor that underpins the given sense.

This is to recognize within the relational structure of the database – and thus of the organization of Latin's semantic system – the theoretical claim that metaphors operate supra-lexically and provide motivating conceptual frameworks for the figurative extension of word meaning. In other words, rather than belonging to the semantic structure of any particular word (or determining, wholesale, the possible figurative meaning of a word), metaphors provide the *specific* pathways of figurative development that *specific* word senses may undergo in the course of a language's history. For instance, *baculum*'s metaphorical sense of 'something providing immaterial support or aid', would be tagged with the metaphor AN EMOTIONAL SUPPORT IS A PHYSICAL SUPPORT (or even more generally, THE EMOTIONAL IS THE PHYSICAL). This metaphor operates independently of this word's semantic structure – it very likely also determines the metaphorical usage of, e.g., *fulcio* – literally, 'to prop up' – in the sense of 'to uphold (emotionally)', as in CIC. Rab. 16, 43, *veterem amicum suum (. . .) labentem excepit, fulsit et sustinuit re, fortuna, fide* ('he supported his old friend – who was slipping downward – with his goods, his fortune and his confidence') – and so provides a powerful mechanism of bringing together otherwise disparate aspects of Latin's semantic system and discovering relationships that otherwise might remain hidden, obscured by outmoded principles of lexicographic organization.

Finally, the ability to organize metaphors into highly articulated networks or groupings via different kinds of mapping relations recognizes that, at a higher level of conceptual structure, metaphors participate in systems. Besides the relations mentioned above, another 'organizing' mechanism of metaphors is that of the image schema. In conceptual metaphor theory, an image schema is "a recurring dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience" (Johnson 1987: xiv). Metaphorical mappings are usually encoded at a quite specific level of semantic granularity, and can be seen as detailed instantiations of more superordinate metaphors relying on general image schemas (e.g., FORCE, CONTAINER, OBJECT). In turn, mappings can give rise to further subordinate figurative patterns, with more semantic details filled in. These hierarchical relationships are all annotated within each metaphor record and give rise to a dense network of interconnected figurative meanings.

5 Annotation procedures and tagging scheme

Annotators first identify (a set of) documented metaphor(s) used by Latin writers to express an abstract concept, corresponding to a given synset, by analysing all occurrences of a relevant (set of) lemma(s) included in the synset within a selected corpus of literary texts. Encoding of metaphors, conceived as mappings between two synsets, is manually conducted through an annotation layer which has been designed expressly for this purpose. Very specifically, a metaphor is annotated according to its **status** (conventional, literary, or imaginative), **type** (primary, complex, orientational, ontological, one-shot image) and **period** of documentation. Moreover, it is labelled with a **shorthand expression** (e.g. 'ideas are food') and an **adjectival descriptor** (e.g. 'alimentary') following conventions in cognitive linguistics. The mapping itself is represented as a unidirectional relationship between two synsets, identified as the

source and **target**. Additional information includes relationships between two or more metaphors at higher or lower levels of semantic specificity, namely through **superordinate** and **subordinate** mappings. Annotators can also catalogue relations between mappings (e.g. extension, elaboration, reciprocity, derivation, combination, and entailment) that may characterize complex metaphor systems.

To exemplify this methodology, we present a case study of metaphor annotation pertaining to the semantic field of fear. A preliminary step identified synset n#05590260, 'an emotion experienced in anticipation of some specific pain or danger' (which pertains to five lemmas pointing to the concept of fear in Latin: *formido*, *metus*, *pavor*, *terror*, *timor*), as the primary target domain of the mapping. We scrutinized all occurrences of these lemmas (4,995) in the 'Antiquitas' section of the *Bibliotheca Teubneriana Latina* (3 BCE to 4 CE), distinguishing between literal (ex. 2) and figurative (ex. 3) usages. We counted but discarded literal usages, and further subclassified figurative usages into more fine-grained metaphorical subschemas.

2. *prae metu ubi sim nescio* (PLAUT. *Cas*. 413) 'I don't know where I am for fear'

3. *huic aliquem in pectus iniciam metum* (PLAUT. *Cas*. 589) 'I'll inject some fear into his heart'

Through careful analysis of the literal wording of the contexts in which these words appear, we identified 23 metaphorical mappings which instantiate three main superordinate image schemas, namely FORCE, CONTAINER, and OBJECT. An example of a metaphor actualizing the FORCE schema is FEAR IS A MILITARY FORCE (ex. 4); whereas FEAR IS A SUBSTANCE THAT FILLS THE EXPERIENCER (ex. 5) exemplifies the OBJECT schema.

4. *tum vero ingens metus nostros invadit* (SALL. *Iug*. 106, § 6) 'at last a great fear assailed the Romans'

5. *vidi hominem XIIII Kal. Febr. plenum formidinis* (CIC. *Att*. 9, 10) 'I saw him on January 17, thoroughly cowed [lit. filled up with]'

Once the catalogue of subschemas appeared to cover all possible metaphorical expressions involving the relevant lexical field, a generalized annotation template was used to record details about each mapping. For example, the annotation record for FEAR IS A MILITARY FORCE is as follows:

status ‹conventional›
type ‹ontological›
period Naev.+ ‹*Pun*. fr. 57, *magnae metus tumultus pectora possidit*›
shorthand expression ‹fear is a military force›
adjectival descriptor 'military'
source ‹n#06088783 | 'an opposing military force'›
target ‹n#05590260 | 'an emotion experienced in anticipation of some specific pain or danger'›
derives from ‹fear is a hostile force›

And it is annotated as follows in the *Lexicon* interface (Figure 1):

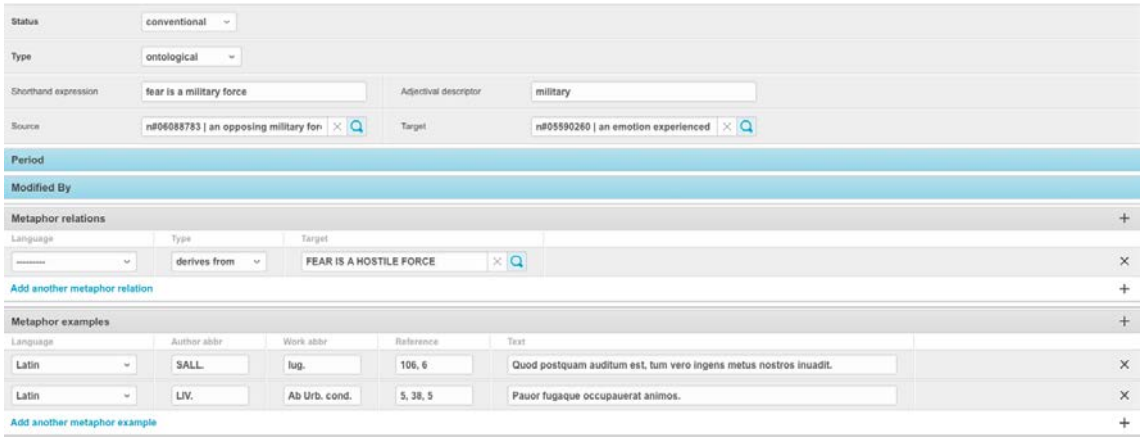


Figure 1. The annotation layer of the FEAR IS A MILITARY FORCE metaphor.

Finally, the metaphor entry is enriched with illustrative examples drawn from literature (ex. 6).

6. *olim iam adversus hunc metum emunivit animum* (SEN. *Con*. 3, 17, 10) 'but he has long since fortified his mind against fear of that'

According to this annotation procedure, users will be able to search the database using a variety of query types. For example, it will be possible to search for a single lemma (like *amor*), for a specific figurative source (like 'fire') or target domain ('love'), for an image schema (COUNTERFORCE), and thus to view all the metaphorical concepts built up from any of these elements. This will make it straightforward to discover certain features of figurative structuration within Latin's semantic system, such as the set of source domains that characterize the understanding of a given concept (what cognitive linguists called the 'range of the target') or, conversely, the set of target domains that are structured by a concept (the 'scope of the source'). It could also help shed light on the ways in which presumably human-universal aspects of cognition (sensorimotor gestalts) provide the scaffolding for culture-specific conceptualizations. What is more, because the metaphorical information of the *Lexicon Translaticium Latinum* piggybacks on the ontology provided by the WordNet, users will automatically be able to take advantage of the rich lexical and semantic knowledge already present in this database, enabling highly complex figuratively-aware queries. The *Lexicon* therefore portends to have significant implications for corpus search, text-processing and other natural language understanding applications.

6 Conclusions

The theoretical and methodological underpinnings of this project, along with the practical annotation procedure it has implemented, suggest that the *Lexicon Translaticium Latinum* could contribute significantly not only to cognitive and semantic approaches and to metaphor theory, but also to linguistic, literary, and cultural research in Classical Studies, especially as part of this field's wider ecosystem of natural language understanding applications. Indeed, we hope to position the Lexicon not merely as a repository of figurative usages in Latin, but as an interface to the system of knowledge itself that Latin speakers relied upon in thinking and speaking in diverse contexts of symbolic expression, and thus as a resource for better understanding how members of Roman society 'made sense' in, and of, their world.

References

Vyvyan Evans and Melanie C. Green. 2006. *Cognitive Linguistics: an Introduction*. Edinburgh University Press, Edinburgh.

Chiara Fedriani. 2016. Ontological and Orientational Metaphors in Latin: Evidence from the Semantics of Feelings and Emotions. In William M. Short (ed.), *Embodiment in Latin Semantics*, 115–140. John Benjamins, Amsterdam.

Dirk Geeraerts and Hubert Cuyckens (eds). 2010. *The Oxford Handbook of Cognitive Linguistics*. Oxford University Press, Oxford.

Raymond W. Gibbs. 2005. *Embodiment and Cognitive Science*. Cambridge University Press, Cambridge.

Evelyn Gius and Janina Jacke. 2017. The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. *International Journal of Humanities and Arts Computing* 11(2):233–254.

Mark Johnson. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. University of Chicago Press, Chicago.

George Lakoff. 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press, Chicago.

Jerome McGann. 2016. Marking Texts of Many Dimensions. In Susan Schreibman, Ray Siemens, and John Unsworth (eds.), *A New Companion to Digital Humanities*, 358–376. Wiley Blackwell, Malden, MA.

Stefano Minozzi. 2008. La costruzione di una base di conoscenza lessicale per la lingua latina: Latin-Wordnet. In Giuseppe Sandrini (ed.), *Studi in Onore di Gilberto Lonardi*, 243–258. Fiorini, Verona.

Egle Mocciaro and William M. Short (eds.). 2019. *Towards a Cognitive Classical Linguistics: The Embodied Basis of Constructions in Greek and Latin*. De Gruyter, Berlin.

Samuel D. Neill. 1987. The Dilemma of the Subjective in Information Organisation and Retrieval. *Journal of Documentation* 43(3):193–211.

Douglas Raber and John M. Budd. 2003. Information as Sign: Semiotics and Information Science. *Journal of Documentation* 59(5):507–522.

William M. Short. 2008. Thinking Places, Placing Thoughts: Spatial Metaphors of Mental Activity in Roman Culture. *Quaderni del Ramo d'Oro* 1:106–29.

William M. Short. 2013. 'Transmission' Accomplished? Latin's Alimentary Metaphors of Communication. *American Journal of Philology* 134(2):247-275.

William M. Short (ed.). 2016. *Embodiment in Latin Semantics*. John Benjamins, Amsterdam.

Selling Autograph Manuscripts in 19th c. Paris: Digitising the *Revue des Autographes*

Simon Gabay
Université de Neuchâtel / ILF
Neuchâtel, Suisse
simon.gabay@unine.ch

Lucie Rondeau du Noyer

Mohamed Khemakhem
INRIA / ALMAnaCH
Université Paris Diderot, Paris 7
mohamed.khemakhem@inria.fr

Abstract

English. In Paris, the manuscript market appears in the early 20's of the 19th c. Fixed-price catalogues and auction catalogues are regularly published, describing each document in detail. Such descriptions being highly formalised, it is possible to extract and structure them (almost) automatically, and thus create a database of sold manuscripts in 19th c. Paris.

Italiano. Il mercato dei manoscritti appare a Parigi all'inizio degli anni '20 del XIX se-colo. In questo contesto, mercanti specializzati vendono regolarmente cataloghi d'asta a prezzo fisso che descrivono minuziosamente ogni documento acquistabile. Tale testi hanno delle strutture ricorrenti che è possibile estrarre e strutturarle (quasi) automaticamente, creando così un database di tutti i manoscritti venduti nella Parigi del XIX secolo.

1 Introduction

The number of projects dealing with French related objects circulating on the private market is increasing: research is currently being carried out in art history (Saint-Raymond, 2018), book history (Montoya, 2018), medieval manuscripts (Wijsman, 2017)... Following recent trends, the latter project, that is the most relevant to our own work, is now sharing its data with other teams (Burrows et al., 2019) in the USA[1] and in the UK [2] to trace the history of manuscripts over time and places, across national and linguistic borders.

Unfortunately, no similar survey has been conducted yet on modern French autographs. If Renaissance manuscripts and their history are better known thanks to the Biblissima project (Turcan-Verkerk and Bertrand, 2014), no systematic work has been carried out on 17th, 18th and 19th c. materials. However, sale catalogues are recognised as being useful, since they are, for instance, regularly used as sources for critical editions (Sévigné, 1978, p. 158) (Voltaire, 1960, p. 18) (Lamartine, 2001, p. 348).

Such a hole in our knowledge is due to the fact that it remains extremely tedious to extract information, because this task is either performed manually or with imperfect digital solutions (Cuadra and Michels, 2013; Barman, 2019). In the present paper, we therefore want to propose an (almost) automated workflow for the retroconversion of catalogues to transform images into structured information and create a database of sold items.

2 The corpus

2.1 The manuscript market

Since the beginning 19th c., rich collectors have been selling manuscripts on the private market (Bodin, 2000). Archives and libraries still keep sales catalogues that have been published on a regular basis (fixed price catalogues) or for special sales (auction catalogues) by dealers. Most of the manuscripts sold in these catalogues are modern and contemporary autograph manuscripts.

The information contained in these catalogues is crucial for at least four different reasons.

[1] https://sdbm.library.upenn.edu
[2] http://mappingmanuscriptmigrations.org

- It helps assessing the authenticity of autographs: it is unlikely that a document sold repeatedly on the private market, and therefore authenticated each time by an expert, is a forgery.
- It documents the reception of authors, via the history of collections (*i.e.* who collected what?) and prices (*i.e.* who costs how much?).
- It informs us on the distance between what has been sold and what is available in libraries (*i.e.* are there autographs we do not know about?).
- It provides us with images of documents which are still in private hands, because catalogues sometimes offer either facsimiles or pictures of autographs sold.

For a first test phase, we have concentrated our efforts on the *Revue des Autographes*, a journal published since 1860's in Paris by Gabriel Charavay.

2.2 The *RDA* collection

In the second half of the 19th c., the autograph market is mature and the first generation of dealers begins to retire. In 1865, Gabriel Charavay (1818-1879) cease the opportunity of Auguste Laverdet's (1809-1867) retirement to take over his business, following the example of his elder brother Jacques (1809-1867), who opened his own shop in 1830. At that moment, Gabriel abandons his role of editor for *L'Amateur d'autographes*, a journal about the autograph market in Paris created in 1862, which keeps being published by his brother Jacques.

Realising the importance of a publication attached to his activities, Gabriel creates another journal one year after his installation, in 1866: the *Revue des autographes, des curiosités de l'histoire et de la biographie* (*RDA*). Two journals for such a small market is however too much: in December 1868, after eight months of interruption, the price of the publication is divided by two and part of the content consists now of a list based on the autographs for sale in Gabriel's stock. Over time, the proportion of articles keeps diminishing and the *RDA* becomes first a hybrid publication mixing news and items to be sold, and eventually a fixed-price catalogue with the name of a journal published (almost) monthly until 1936. In the meantime, Gabriel's shop is taken over by Gabriel's son Eugène (1879-1892), and then by Eugène's widow (1892-1918) and by Eugène's daughter (1918-1936).

The transformation of an hybrid journal into a disguised fixed-price catalogue under a journal's name is confirmed by a modification of the format: Eugène Charavay opts for a two-columns layout and a smaller font (cf. figure 1), harder to read but easier to browse for readers, who are now buyers, looking for the autograph of their dreams.

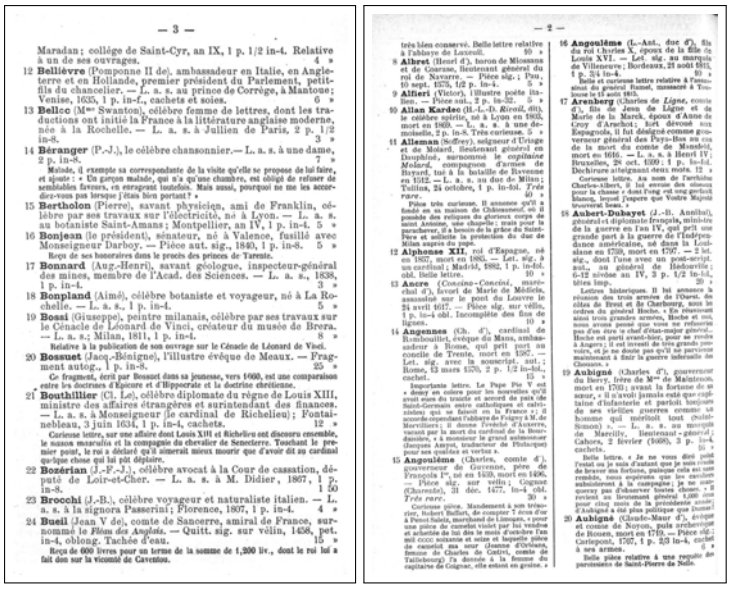
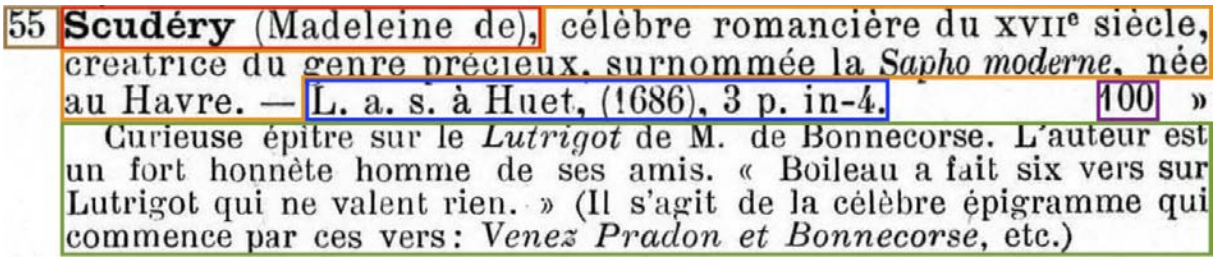


Figure 1: One-column layout (1873) vs two-columns layout (1893).

3 Encoding

3.1 Entries

It is the images of these catalogues that we want to transform into minable data. Each of them generally contains a minimum of c. 200 entries, all of them being extremely dense in information and always following the same structure:


55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

Figure 2: *RDA*, n°67 (March 1881), lot N°55.

We can clearly see the lot number (in brown), the name of the author (in red), a short biography (in orange), the material description of the autograph (in blue), the price (in pruple) and an additional description (in green).

To render the structure of the document, we propose the following encoding in XML-TEI:

Such an encoding allows a simple disambiguation of the entry and increases the accuracy of the search. In our example, we have the names of three major 17th c. French writers: the novelist Madeleine de Scudéry (1607-1701), the bishop Daniel Huet (1630-1721), and the satirist and poet Nicolas Boileau-Despréaux (1636-1711). The three names are enclosed in three different tags (`name`, `desc` and `note`) reflecting their status in the document (author, addressee, mention): we can therefore easily narrow down our query to a name depending on its role.

3.2 Workflow

The presented encoding can be compiled semi-automatically through a simple three steps workflow:

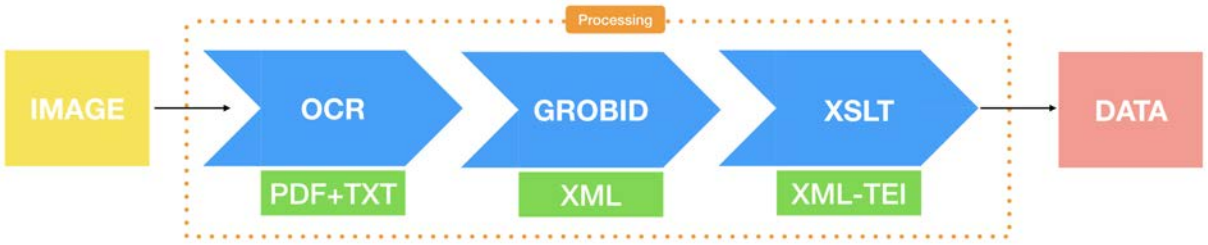


Figure 3: Workflow.

The scan is OCRised with Transkribus (Kahle et al., 2017), for which a substantive model of 125,000 words has been created (CER of 0.59%). The pdf with a text layer is then processed with GROBID-Dictionaries (Khemakhem et al., 2018b), a tool relying on text and layout features (cf. Figure 4) to perform a supervised classification of the parsed text and generate a TEI compliant encoding where the various segmentation levels are associated with an appropriate XML tessellation (Khemakhem et al., 2017). Preliminary designed for the retroconversion of dictionaries (Bohbot et al., 2018), it is also used to parse large bibliographical collections (Lindemann et al., 2018) or address directories (Khemakhem et al., 2018a).

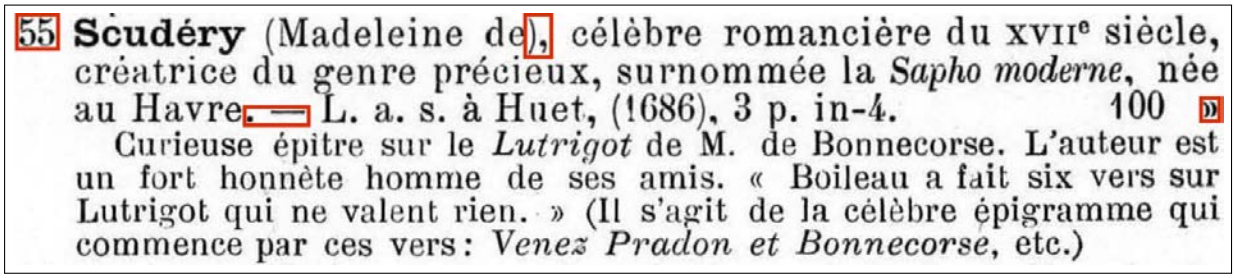

55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, créatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100
Curieuse épitre sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

Figure 4: Features.

GROBID-dictionaries provides an answer to important issues left open by previous attempts, that do not produce standardised data (*e.g.* in XML-TEI), are not open source (Cuadra and Michels, 2013) or do not offer fine-grained encoding (Barman, 2019). On top of this, GROBID is a free, language agnostic, easily trainable solution compatible with other sub-projects of the GROBID galaxy, which leaves the door open to further analysis with complementary tools, such as GROBID-NERD (Named-Entity Recognition and Disambiguation).

After preliminary tests ensuring the compatibility of GROBID-Dictionaries with sale catalogues (Khemakhem et al., 2018c) models have been created for the *RDA* and many other catalogues (Rondeau du Noyer et al., 2019) with a new bigram template for the GROBID-Dictionaries models (Rondeau Du Noyer et al., 2019) to reinforce the parsing of the structure of each entry.

With GROBID-Dictionaries, the document is annotated using a cascading approach: several Conditional Random Fields (CRF) models are applied one after the other, each of them corresponding to a granularity level in the final XML hierarchy:

Levels	**Tag(s)**	**Task**
1	body	separates the content from running titles, page numbers, . . .
2	entry	separates the entries in the <body>
3	num, form and sense	separates the lot n°, information on the author and the MS in <entry>
4	name and desc	separates the name of the author and its biography in <form>
5	subsense and note	separates the MS description and the additional note in <sense>

Table 1: GROBID-Dictionaries Segmentation levels

The GROBID-Dictionaries output for our example is therefore the following:

GROBID-Dictionaries being developed for lexicographic purposes, its results are encoded in a TEI compliant output, but with tags reflecting the content of dictionaries rather than catalogues. Therefore, we automatically convert the output into a second TEI document whose tags are dedicated for the described catalogue elements. The consistency of the transformation output is controlled with a specific schema, prior to its final publication via an XML database.

4 Future work

As for future work, four tasks will be undertaken. First, we will move towards a fully open source workflow and therefore abandon Transkribus for Kraken (Kiessling, 2019). Second, we will increase the size of our database by retroconverting the entire *RDA* collection (c. 500 catalogues), but also another important series of catalogues: the *Lettres autographes et documents* published by the other branch of the Charavay family up to the First World War (c. 500 catalogues). Third, we will improve our modeling and increase the granularity of our data collection to capture more informations regarding the document (size, format, length) and named entities (people, places). Fourth, we will use this additional information to reconcile entries and share our work with similar projects via an RDF export.

Ideally, we should eventually be able to go from the TEI digital edition of sales catalogues to a semantic dataset, described using controlled vocabularies where authors, places and manuscripts would be referred to using unique identifiers (ISNI, ISMI. . .). It would allow federated search with other databases of sold manuscripts, but also with catalogues of libraries in France and abroad.

References

Raphael Barman. 2019. Aucase - auction catalog segmentation. Type: dataset. https://github.com/raphaelBarman/aucase-inha/.

Thierry Bodin. 2000. Les grandes collections de manuscrits littéraires. In *Les Ventes de livres et leurs catalogues: XVIIe-XXe siècle*, École des chartes, Paris, pages 169–190.

Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem, and Laurent Romary. 2018. Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. In *GLOBALEX 2018 - Globalex workshop at LREC2018*. Miyazaki, Japan, pages 1–6. https://hal.archives-ouvertes.fr/hal-01728328.

Toby Burrows, Eero Hyvönen, Lynn Ransom, and Hanno Wijsman. 2019. Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts. *Manuscript Studies* 3(1). https://repository.upenn.edu/mss_sims/vol3/iss1/13.

Ruth Cuadra and Suzanne Michels. 2013. Publishing German Sales, A Look under the Hood of the Getty Provenance Index. https://blogs.getty.edu/iris/publishing-german-sales-a-look-under-the-hood-of-the-getty-provenance-index/.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Kyoto, Japan, volume 04, pages 19–24. https://doi.org/10.1109/icdar.2017.307.

Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet, and Jean-Luc Pinol. 2018a. Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. In *JADH2018 "Leveraging Open Data"*. Tokyo, Japan. https://hal.archives-ouvertes.fr/hal-01814189.

Mohamed Khemakhem, Luca Foppiano, and Laurent Romary. 2017. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. eLex, Leiden, the Netherlands. https://hal.archives-ouvertes.fr/hal-01508868v2.

Mohamed Khemakhem, Axel Herold, and Laurent Romary. 2018b. Enhancing Usability for Automatically Structuring Digitised Dictionaries. In *GLOBALEX workshop at LREC 2018*. Miyazaki, Japan. https://hal.archives-ouvertes.fr/hal-01708137.

Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, and Giancarlo Luxardo. 2018c. Automatically Encoding Encyclopedic-like Resources in TEI. Tokyo, Japan. https://hal.inria.fr/hal-01819505.

Benjamin Kiessling. 2019. Kraken - an universal text recognizer for the humanities. Utrecht, The Netherlands. https://dev.clariah.nl/files/dh2019/boa/0673.html.

Alphonse de Lamartine. 2001. *Correspondance*, volume 3. Honoré Champion, Paris.

David Lindemann, Mohamed Khemakhem, and Laurent Romary. 2018. Retro-digitizing and Automatically Structuring a Large Bibliography Collection. In *European Association for Digital Humanities (EADH) Conference*. EADH, Galway, Ireland. https://hal.archives-ouvertes.fr/hal-01941534.

Alicia C. Montoya. 2018. The MEDIATE project. *Jaarboek voor Nederlandse Boekgeschiedenis / Yearbook for Dutch Book History* 25:229 – 232.

Lucie Rondeau Du Noyer, Simon Gabay, Mohamed Khemakhem, and Laurent Romary. 2019. Scaling up Automatic Structuring of Manuscript Sales Catalogues. Graz, Austria. https://hal.inria.fr/hal-02272962.

Lucie Rondeau du Noyer, Simon Gabay, Mohamed Khmakhem, and Laurent Romary. 2019. Automatic TEI encoding of manuscripts catalogues with GROBID-Dictionaries. Type: dataset. https://doi.org/10.5281/zenodo.3383658.

Léa Saint-Raymond. 2018. Le pari des enchères : le lancement de nouveaux marchés artistiques à Paris entre les années 1830 et 1939. Corpus bibliographique des ventes aux enchères publiques considérées Type: dataset. https://doi.org/10.7910/DVN/MZIBKB.

Marie de Rabutin-Chantal Sévigné. 1978. *Correspondance*, volume 3. Gallimard, Paris.

Anne-Marie Turcan-Verkerk and Paul Bertrand. 2014. BIBLISSIMA: Bibliotheca bibliothecarum novissima, an observatory for written cultural heritage of the Middle Age and the Renaissance. In *Heritage and Digital Humanities. How should training practices evolve?*, Berlin, pages 129–139.

Voltaire. 1960. *Correspondence*. 59. Institut et musée Voltaire, Genève.

Hanno Wijsman. 2017. The Bibale Database at the IRHT. *Manuscript Studies* 1(2). https://repository.upenn.edu/mss_sims/vol1/iss2/10.

Enriching a Multilingual Terminology Exploiting Parallel Texts: An Experiment on the Italian Translation of the Babylonian Talmud

Angelo Mario Del Grosso, Emiliano Giovannetti, Simone Marchi
Istituto di Linguistica Computazionale "A. Zampolli"
{name.surname}@ilc.cnr.it

Abstract

English. Parallel texts can represent an extremely useful source of information in a number of text and linguistic processing tasks. In this work we show an experiment conducted on the Italian translation of the Babylonian Talmud, a text we have analyzed and processed to support in the construction of a multilingual Hebrew/Aramaic/Italian terminological resource. The approach we adopted comprised: i) the TEI encoding of the text, ii) the automatic extraction of the Italian terms, iii) the addition of Hebrew/Aramaic terms via word-by-word alignment, iv) the revision of the obtained results.

Italiano. I testi paralleli possono costituire una fonte estremamente utile di informazioni per numerosi task di elaborazione del testo e della lingua. In questo lavoro illustriamo un esperimento condotto sulla traduzione italiana del Talmud babilonese, un testo che abbiamo analizzato ed elaborato per supportare la costruzione di una risorsa terminologica multilingue in Ebraico, Aramaico e Italiano. L'approccio adottato comprende: i) la codifica TEI del testo, ii) l'estrazione automatica dei termini italiani, iii) l'aggiunta dei termini ebraici e aramaici tramite tecniche di allineamento parola per parola, iv) la revisione dei risultati ottenuti.

1 Introduction

Translation is the only way of making a text accessible to people that do not understand the language the original text is written in. Translation, in other words, allows to build bridges between peoples and cultures. It is no coincidence that it has been through a translation, contained in the well-known Rosetta Stone, that Egyptian hieroglyphs could be deciphered. The work we here describe is based upon a similar principle: how to exploit the translation in a "known" language of a text written in an "unknown" language to derive some linguistic information from the latter. In our case, the "known" language is a language for which tools and resources are available to automatically extract information from a text written in that language. Viceversa, the "unknown" language is the one that poses analytical problems, as it typically happens in projects involving ancient texts and languages. In particular, as detailed in the following section, we wanted to experiment a way of supporting the construction of a multilingual terminology by exploiting an existing translation.

The use of parallel texts in support to lexicon construction is a field known as bilingual lexicon extraction, and it has a wide scientific literature (see for example (Fung, 1998), (Tufiş et al., 2004), (Gutierrez-Vasques, 2015)). From a more applicative point of view, tools and software libraries have been implemented to assist developers in implementing the word-by-word text alignment necessary to process parallel texts. Giza++[1] and the Berkeley aligner[2], for example, have been largely adopted for these tasks. More in general, and in the context of Digital Humanities, the idea of exploiting parallel texts has been adopted in a number of initiatives, among which we point out the Perseus project, where the project team, together with the Von Humboldt professorship G. Crane within the Global Philology

[1] http://www.statmt.org/moses/giza/GIZA++.html
[2] https://code.google.com/archive/p/berkeleyaligner/

project, analyzed and implemented a collection of technologies and tools to envisage the "complexities of working with a historical record that contains far more languages than any individual could study, much less master" (Crane, Gregory et al., 2019).

2 Objectives and motivation

The experiment we here illustrate, still in progress, was conducted on the Babylonian Talmud Italian translation, in the context of the homonymous project[3]. The project, in addition to the development of the software Traduco used to support in the translation of the Talmud (Giovannetti et al., 2016), envisages the construction of a multilingual (Hebrew-Aramaic-Italian) terminological resource to support a number of activities, such as boosting the Translation Memory System with terminological information and creating an ontology of the talmudic domain. As described in Section 3, the Italian portion of the resource was built with the aid of a terminology extractor exploiting linguistic analysis tools for Italian. However, no tool or linguistic resource was available to automatically process the three main ancient languages appearing in the Talmud, namely, mishnaic Hebrew, biblical Hebrew and babylonian Aramaic. To obviate to this issue, and to the difficulty of automatically detecting the source terms through standard extraction processes, we chose to exploit the data produced in the last seven years of project activities, i.e. the available translated tractates of the Talmud. The results of the experiment suggested more ways of exploiting the obtained list of term-pairs in addition to the enrichment of the terminological resource, for example, as it will be discussed in the final version of the paper, to help in the lemmatization of semitic languages.

3 Methodology

Basically, the proposed approach makes use of a word-by-word alignment technique applied to a text in translation. The overall extraction process, leading to the enrichment of the terminological resource, followed a four step approach: i) encoding of the parallel text in TEI, ii) extraction of the Italian terms using a customized term extractor, iii) application of a word-by-word alignment technique to the parallel textual segments of the Talmud, iv) manual revision of the obtained alignment for the detection of the Hebrew/Aramaic terms corresponding to the Italian ones.

3.1 TEI encoding of the parallel text

We have first modeled and encoded the available parallel text (i.e. the Talmud and its Italian translation) by means of the best practices dictated by the Text Encoding Initiative (TEI), whose schema is currently the *de facto* standard to encode text-bearing objects (TBO) within the most authoritative scholarly projects involving literary inquiries. Actually, the choice to adhere to TEI environment provides benefits both to scholars, offering a standard model for the digital representation of critical texts, and to technicians, concerning modularity, data management, and, in particular, independence related to specific development choices. We have adopted the hierarchical text-group technique in order to encode the basic textual segments in three different modalities: 1) the original talmudic text; 2) the Italian translated text; 3) the literal Italian translated text. Moreover, the linkage among the different textual fragments has been conducted by means of the linkgroup technique [4]. Section 3.3 will illustrate the word-by-word alignment task that has been developed.

3.2 Extraction of the target terms

As mentioned before, given the lack of NLP tools and resources for Ancient Hebrew and Aramaic we could carry out the automatic extraction of the terminology only on the italian translation of the Talmud. For this purpose we used T2K^2 (Dell'Orletta et al., 2014), a platform for linguistic analysis available at the Institute of Computational Linguistics (ILC) of the Italian National Research Council (CNR). T2K^2 includes a stochastic module for terminology extraction which appeared adequate for our experimental purposes. We applied the extractor to four of the already translated and revised tractates of the Talmud,

[3] https://www.talmud.it

[4] Module number 16 of the TEI guidelines - Groups of Links. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SAPTLGen/html/SA.html#SAPTLG

namely: Berakhòt, Rosh haShanà, Ta'anìt and Qiddushìn. The corpus made of textual (plain-text UTF-8) documents was analyzed with T2K^2 and the obtained output was furtherly processed in order to remove erroneous terms deriving from Part-Of-Speech tagging errors, and to sort the extracted terms by means of the TF-IDF (Term Frequency-Inverse Document Frequency) statistical measure. An important outcome of the TF-IDF is to permit to measure the relevance of each term for each tractate in which it appears: a high value of TF-IDF represents a high degree of relevance in the context of a specific tractate. The Table 1 shows some examples of term relevance by tractate.

Berakhòt			Qiddushin		
Terms	tfidf	freq	Terms	tfidf	freq
Birkàt haMazòn	0.0239	120	documento	0.0159	123
emissione di seme	0.0209	105	perutà	0.0155	120
Shemà	0.0205	206	qiddushìn	0.0142	55
sogno	0.0166	167	schiava	0.0119	46
gabinetto	0.0076	38	terra di Israele	0.0107	83
frutto della terra	0.0072	36	divorzio	0.0104	40
benedizione sul vino	0.0062	31	rapporto sessuale	0.0101	78
tipi di cibi	0.0060	30	padrone	0.0100	186
pane dalla terra	0.0056	28	schiava ebrea	0.0088	34
bisogni	0.0047	47	trovatello	0.0080	31

Table 1: The first ten Italian terms extracted from two of the four analyzed tractates and ordered by tf-idf.

3.3 Extraction of the source terms via alignment

Word-by-word text alignment is a very useful technique to help understanding cross-lingual properties of parallel texts while processing only one half of the whole resource (Tiedemann, 2011). In order to add the Hebrew and Aramaic terms to the terminological resource we are building up from the Talmud, we set up the alignment process at token granularity. Specifically, we used an open source library realized by the Berkeley University (Liang et al., 2006) to develop a tool for the linking of Hebrew/Aramaic textual segments with the corresponding Italian translations.

Italian terms	most likely Hebrew term	other candidates Hebrew terms
benedizione (2.1)	בְּרָכָה (0.41)	מְבָרְכִין (0.29), מְבָרֵךְ (0.09)
Shemà (1.1)	קְרִיאַת (0.53)	שְׁמַע (0.44)
preghiera (2.2)	תְּפִלָּה (0.30)	תְּפִלַּת (0.15), תְּפִלָּה (0.13), תְּפִלָּתוֹ (0.16)
pane (1.9)	לֶחֶם (0.27)	פַּת (0.16), רִיפְתָּא (0.16), הַפַּת (0.22)
anno (2.14)	שָׁנָה (0.32)	שָׁנָה (0.15), הַשָּׁנָה (0.10), הַשָּׁנָה (0.19)
mese (1.93)	חֹדֶשׁ (0.25)	הַחֹדֶשׁ (0.21), לַחֹדֶשׁ (0.24)
giorno (1.90)	יוֹם (0.36)	הַיּוֹם (0.14), בַּיּוֹם (0.09), יוֹמָא (0.19)
shofàr (0.87)	שׁוֹפָר (0.77)	בְּשׁוֹפָרוֹת (0.08)
obbligo (2.1)	יָצָא (0.34)	חוֹבָה (0.09), חוֹבָתוֹ (0.22)
schiavo (0.82)	עֶבֶד (0.80)	וְעֶבֶד (0.09)

Table 2: Some examples of Italian-Hebrew/Aramaic aligned terms. Italian terms with high entropy (such as "preghiera") have been aligned with multiple Hebrew/Aramaic terms: the confidence that the term "תְּפִלָּה" (the one with the highest likelihood) is the actual translation of preghiera is low.

To carry out the word-by-word alignment, the tool implements generative models that have been studied during the last decades by the IBM researchers and by the Machine Translation community

(Brown et al., 1993). In particular, it adopts the IBM Model-1 with the extension of the Hidden Markov Model paradigm (Östling and Tiedemann, 2016). The alignment task employed non-supervised machine learning algorithms adopting probabilistic models to calculate the likelihood estimation of aligning a term in a known language to a term in a foreign language. The expressiveness of these kinds of alignment models is particularly suitable in the literary domain, where translations tend to be more interpretative and less literal. Eventually, for each Italian term the computed probabilistic alignment model provided a list of Hebrew/Aramaic candidate words. In Table 2, the numbers reported next to the Italian terms represents the entropy measure, which indicates the confidence of the translated word. The numbers next to the Hebrew/Aramaic words indicate the likelihood that word is the translation of the corresponding Italian word.

3.4 Manual revision

The aligner developed so far is based on statistical approaches which are, inherently, prone to errors. For this reason, the alignment environment requires a tool to validate and manually annotate the obtained outputs. We are thus developing a Web application able to manage and process the aligned text segments. As shown in 1, we have provided the proofreader with the possibility to annotate each word with a number of language and textual traits, namely lemma, Part-of-Speech, type of text, and language.

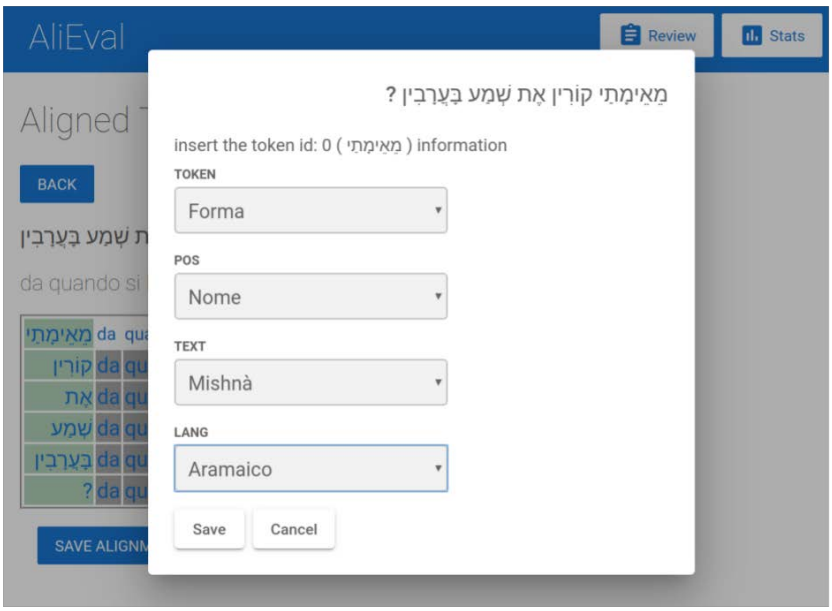


Figure 1: The annotation component of the proofreader.

The output of the aligner is formatted as a sequence of strings like `0-0 4-6 2-5 3-4 1-2 1-1` representing the word pairs that have been aligned. The order of the pairs is not significant, while the number within the pair represents the position of the word within the source-target strings; for example, in the two strings "עַד סוֹף הָאַשְׁמוּרָה הָרִאשׁוֹנָה" and "fino alla fine della prima veglia" the pair 0-0 would indicate the word pair "עַד-fino" (Hebrew is read from right to left). More details about the proofreader will be provided in the final version of the paper. Eventually, the revision process will allow to build a ground truth and/or a gold training set and consequently put in place a complete validation process of the alignment results.

4 Preliminary results, discussion and next steps

As it was shown, a parallel text can be exploited fruitfully via text alignment techniques to help in the construction of a multilingual terminology. Our reference scenario was the Italian translation of the

Babylonian Talmud, carried out in the context of the homonymous project. At the current stage of the work, 219.000 tokens have been analyzed, distributed on 42.000 textual segments extracted from the four aforementioned tractates which have been translated so far.

In addition to their use in populating the terminological resource, the obtained term-pairs may be also exploited in other ways. The first two applications we are going to investigate are: the boosting of text search, as recently experimented also in (Andonovski et al., 2019), and the support in the automatic processing of the source language.

Concerning the next steps of this research, once a significant number of segments (and, thus, of the terms appearing in the segments) will have been revised by the expert of the Talmud, a formal evaluation of the accuracy of the approach will be carried out. Fig. 2 shows an example of revision of the alignment.

Figure 2: An example of use of the proofreader: the output of the automatic alignment (at the top) and the relative revision (at the bottom).

Besides, we intend to improve the performance of the approach by taking into account the variety of texts and languages that coexist inside the Talmud before the application of the aligner. As a matter of fact, the Babylonian Talmud is constituted by two (macro) texts, i.e. the Mishna and the Gemara, which, in turn, incorporate portions of other texts, such as, for example, quotes from the Tanakh (the Hebrew Bible). In the particular case of the Talmud, each text is written in a specific language: the Mishna in Mishnaic Hebrew, the Gemara in Babylonian Aramaic and the Tanakh in Biblical Hebrew. The idea is to automatically classify each segment of the Talmud on the basis of the text it belongs to and, after that,

to apply the aligner on each textual class composed of linguistically homogeneous segments. By doing this, we expect a better accuracy from the aligner and, ideally, no need from the revisor to indicate the language of each segment.

Acknowledgement

This work was conducted in the context of the TALMUD project and the scientific cooperation between S.ca r.l. PTTB and ILC-CNR.

References

Jelena Andonovski, Branislava Šandrih, and Olivera Kitanović. 2019. Bilingual lexical extraction based on word alignment for improving corpus search. *The Electronic Library* .

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311. https://www.aclweb.org/anthology/J93-2003

Crane, Gregory, Jovanovic, Neven, Sklaviadis, Sophia, De Luca, Margherita, Šoštarić, Petra, Foradi, Maryam, Cottrell, Kate, Tauber, James, Shamsian, Farnoosh, and Palladino, Chiara. 2019. Confronting Complexity of Babel in a Global and Digital Age. In *Complexities*. ADO, Utrecht, Netherlands. https://dev.clariah.nl/files/dh2019/boa/0611.html

Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2K^2: a system for automatically extracting and organizing knowledge from texts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Machine Translation and the Information Soup*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 1–17.

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2016. Traduco: A collaborative web-based cat environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities* 32(suppl_1):i47–i62.

Ximena Gutierrez-Vasques. 2015. Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. pages 154–160.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-NAACL '06, pages 104–111. https://doi.org/10.3115/1220835.1220849

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics* 106:125–146. https://doi.org/10.1515/pralin-2016-0013

Jöho Tiedemann. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Dan Tufiş, Ana Maria Barbu, and Radu Ion. 2004. Extracting Multilingual Lexicons from Parallel Corpora. Computers and the Humanities 38(2):163–189. https://doi.org/10.1023/B:CHUM.0000031172.03949.48

Towards a Lexical Standard for the Representation of Etymological Data

Fahad Khan
ILC-CNR/Pisa
fahad.khan@ilc.cnr.it

Jack Bowers
Austrian Center for Digital Humanities/Vienna
Inria - Team ALMAnAC/Paris
Jack.Bowers@oeaw.ac.at

Abstract

English. We introduce a new standard (currently under development), LMF Diachrony-Etymology, which is intended to constitute a common model for the creation of diachronic lexical data, and in particular etymologies, as computation lexical resources. Having situated it whtin in the context of previous developments in this area, we outline the content of the new standard, and in particular the core classes of the model as well as describing our overall approach to modelling etymological data. Finally, we give an example encoding of an entry taking from an etymological dictionary of Latin.

Italiano. Questo paper presenta LMF Diachrony,-Etymology un nuovo standard, (attualmente in fase di sviluppo), volto a costituire un modello comune per la rappresentazione di dati lessicali diacronici e in particolare di etimologie, sotto forma di risorse lessicali computazionali. Situeremo lo standard nel contesto delle tendenze attuali del settore e ne presenteremo il contenuto, con particolare riferimento alle principali classi del modello e alla descrizione dell'approccio complessivo alla modellizzazione diei dati etimpolgici. Infine, presenteremo la modellizzazione formale di una entrata tratta da un dizionario etimologico del latino.

1 Introduction

In this submission we will introduce a new standard, currently in an advanced stage of development[1], for the modelling and publication of etymological data in computational lexical resources[2]. The standard in question, which we will refer to it as LMF-Ety, is the third part of a new multi-part revision of the Lexical Markup Framework (LMF), ISO 24613-4, originally published by the International Standards Organisation (ISO) as a single standard in 2008[3]. We will motivate the need for LMF-Ety by describing some of the main challenges of modelling etymological data in computational lexical resources and showing how our new standard meets these challenges as well as how it differs from other previous models. Subsequently, we will describe the core concepts which we have so far established in our model and illustrate them through the use of an extended example taken from an etymological dictionary. Our intention is both to summarise the work we have carried out in the development of the LMF-Etymology standard as well as to showcase our broader approach to modelling etymologies. This approach entails the representation of etymologies as formal graphs describing simple narratives relating to a given lexicon phenomena; it is an approach that takes account of and consolidates previous attempts at modelling etymologies computationally but that also seeks to extend them in various different directions with a view to obtaining a more robust and expressive model. Additionally both authors are also involved in other initiatives for modelling etymologies in two other frameworks/standards, the Text Encoding Initiative (TEI) (Bowers and Romary, 2016) and Linked Data (Khan, 2018). We will end the submission by

[1]At the time of writing, December 2019, the standard is being prepared for submission to an ISO ballot as a Draft International Standard (DIS). The classes and the approach which we present are now therefore fairly stable.

[2]The two authors are the joint project leaders of this standard.

[3]Note that the original version of LMF did not contain specific provision for modelling etymological/diachronic information.

describing how LMF-Etymology may be rendered inter-operable with work being carried out in these two latter frameworks. This will also be relevant for understanding the practical details of how LMF-Etymology can actually be used (SPOILER ALERT: Part 4 of the LMF standard is a serialisation of all the previous parts in TEI-XML).

2 Background

The importance of standards for the publication of scientific and scholarly datasets and resources for rendering them more findable, accessible, interoperable and reusable is by now well understood across the board. There have been a number of initiatives for promoting such standards and best practices in the field of language resources. Three of the most notable of these as applied to the case of lexical datasets are: the original version of the Lexical Markup Framework described below; the Dictionaries chapter of the Text Encoding Initiative (TEI) guidelines[4]; and finally the RDF-based Ontolex-Lemon guidelines (McCrae et al., 2017). These standards not only help to ensure a greater measure of interoperability between different computational lexicons, but they also facilitate the representation of lexical information in a way that makes it more amenable to advanced kinds of machine processing. Up until recently all three of these standards have dealt almost exclusively with synchronic lexical data[5]. This neglect of diachronic data is due, in part, to the awkwardness associated with the addition of extra temporal parameters to statements in data frameworks such as UML or RDF, and partly due to the (relatively) slow pace of development in the three standards overall – even if this would seem to constitute a missed opportunity, particularly in the case of etymological data since, at an abstract level, etymologies traditionally describe graph structures. They would therefore be ideally suited for representation in formalisms where this underlying structure can be rendered explicit, making such data easier to query and process. Moreover standards like LMF and especially the RDF-based Ontolex-Lemon would potentially make it easier to link together and query across different etymological datasets and to therefore create extended etymological networks. LMF-Ety is intended both for the creation of etymological datasets *ex novo* as well as for the conversion of legacy print resources as structured data. In this latter respect it should be noted that although our initial use cases have so far been largely concerned with the conversion of legacy dictionaries into structured resources, descriptions of etymological graph structures can be found in, and therefore potentially extracted from, numerous different kinds of texts. These include both scholarly works in linguistics, especially in the sub field of historical linguistics (articles, book chapters, monographs, etc), along with other genres of texts, literary, religious and philosophical[6]. It is clear then that the computational modelling of etymologies stands firmly at the intersection of computational linguistics, e-lexicography and the digital humanities. This inter-disciplinarity can also be appreciated in the fact that etymologies for languages with a sufficiently extensive written tradition will often contain attestations to coprora of historic texts; these texts can sometimes be reconstructions or have disputed interpretations as to particular word senses, (bringing to bear issues concerned with textual criticism and philology/literary criticism more generally).

2.1 The Lexical Markup Framework

The original 2008 version of LMF, ISO 24613: 2008, was intended as a "standardized framework for the construction of computational lexicons" (Francopoulo, 2013) an was conceived of as a common model both for lexicons for use in NLP applications as well as for computational versions of print or legacy dictionaries. Regarded as one of the most important standards in the field of lexical resources, LMF was enormously influential in the definition of the Ontolex-Lemon model and its predecessor*lemon*. A review of the Lexical Markup Framework undertaken by ISO in 2016 resulted in a decision to revise the standard, and to publish it as a multi-part standard (Romary et al., 2019) to render it more modular; the decision was also made to broaden the applicability of the new version of LMF to capture more kinds of lexical information. In consequence it was decided that one of the new parts of LMF should

[4] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

[5] The TEI guidelines do permit the representation of etymological data in a structured way but in a relatively shallow way. A recent proposal to allow for more salient kinds of etymological annotation can be found in (Bowers and Romary, 2016)

[6] See (Khan, 2018) which presents an example taken from Hobbes'*Leviathan*.

be a specialised module dealing with diachronic lexical information. The different parts of the new LMF standard are: the **Core Model** (ISO 24613-1); the **Machine Readable Dictionary Module** (ISO 24613-2); the **Diachrony-Etymology Module** (ISO 24613-3); Serialisations in both TEI (ISO 24613-3) and LBX (ISO 24613-3). Two new modules dealing with syntax and semantics and morphology have also been proposed.

2.2 Related Work and Standards

A number of proposals have been made in the past to try and redress the lack of provision for encoding structured etymological data in lexical resources. These have included suggestions for extensions of both for TEI (Bowers and Romary, 2016) and LMF (Salmon-Alt, 2006), as well as proposals for new Ontolex-lemon classes and properties(Chiarcos et al., 2016), (Khan, 2018). The standard described in the current work is influenced by the aforementioned works, and in particular it is informed by the approach in taken in (Khan, 2018) while abstracting away from specific details mentioned in that work and which pertain to the Resource Description Framework (RDF). Moreoever, it is the result of an attempt to converge towards a set of high level concepts, abstractions, that are sufficiently expressive to encode the main kinds of phenomenon and information which tend to be included in etymologies, as well as being simple enough to be usable by a wide community of potential users. The main concepts which we have determined upon are described in the next section. As we mentioned above the fourth part of the new standard is a TEI serialisation of all the other parts (due to be published at the same time as LMF-Etmyology). This should ensure that at a high level both TEI and LMF are interoperable and in particular the etymological apprpach taken by LMF-Etymology is compatible with TEI. It also means that LMF-Etymology should ultimately be accessible to digital humanists who are more used to working with TEI.

3 LMF Etymology

The definition of LMF Ety (ISO Standard 24613-3) is dependent on the two preceeding ISO standards in the new multi-part LMF standard, i.e., the Core Module, recently published by ISO, and the Machine Readable Dictionaries Module, due to be published in 2020 (Romary et al., 2019). Both of these standards contribute foundational concepts (for modelling lexicons) to LMF Ety such as **Lexical Entry**, **Lemma**, **Form**, and **Sense**: all of which keep their (fairly intuitive) meaning from the previous version of LMF and all of which share the meaning of similarly titled concepts in TEI and Ontolex-Lemon. On the basis of these foundational concepts then LMF Ety defines a number of additional classes which enable us to associate temporal/historical information with lexical data encoded in LMF. The strategy we adopt is that suggested in (Khan et al., 2014) of modelling linguistic elements such as words, senses, forms, etc as *perdurants*, that is, as entities associated with a lifespan, which in the present case represents the interval of time in which they are considered to have been part of common usage within a given linguistic community[7]. This enables us to situate lexical entries etc in a temporal dimension and also to relate them together via diacrhonic linguistic processes. Our model then represents etymologies as simple narratives, or as rather simple *narrative graphs*, in which different linguistic phenomena (each of which can be potentially associated with a lifespan and situated on a timeline) are linked together using special etymological link elements, individuals of the class **EtyLink**, which can represent different kinds of historical linguistic processes such as *inheritance* or *borrowing* or *semantic shift*. The other new elements in the standard are the described below:

- **Etymon** and **Cognate**: Two elements modelled as subtypes of **Lexical Entry**. What differentiates them from other lexical entries in a lexicon is their (specialised) role: they are used in describing the etymologies of other lexical entries: **Etymon**s are lexical entries from which a given lexical entry is derived via some historical process; **Cognate**s are lexical entries which share a common ancestor with a given a lexical entry; Additionally **Cognate Set** represents the reification of a set of cognates.

- **Etymology**: An element that represents a single history of a lexical entry or other element. We associate **Etymology** individuals with an ordered series of **EtyLink** instances; this allows us to

[7]This approach makes it easier to represent such information in RDF.

forum 'market place, public space; place where the fruit was laid for pressing (Cato+)' [n. *o*; *forus* Lucil., Pompon., CIL] (Lex XII+)
Derivatives: *forus* 'deck (on a ship); passage (in a beehive); rows of benches (in a stadium)' (Enn.+), *forēnsis* 'of the forum, public' (Varro+).
PIt. **fworo-* '(room) near the door'. It. cognates: U. **furu**, *furo* [acc.sg.] 'forum'.
PIE *dʰuor-o- '(room near the) door'. IE cognates: Skt. *dvāram* [n.] 'door, gate, passage', Lith. *dvāras* [m.] 'estate; court', OCS *dvorъ* 'court', PTo. **twere* 'door'.
WH interpret *forum* as 'fenced area' to the root of *forāre*, but Pokorny 1959 rejects this. *Forum* is generally regarded as a derivative of PIE 'door', and connected with other IE forms from **dʰuor-o-*. The required semantic development is 'area at the doors' > 'entrance room, vestibule' > 'public room' > 'public space'; this is not so problematic as to overrule the formal correspondences with Lith. *dvāras*.
Bibl.: WH I: 537f., EM 250, IEW 278f., Meiser 1986: 116, Schrijver 1991: 471f., Sihler 1995: 180, Untermann 2000: 305. → *foris*

Figure 1: Entry for *forum*.

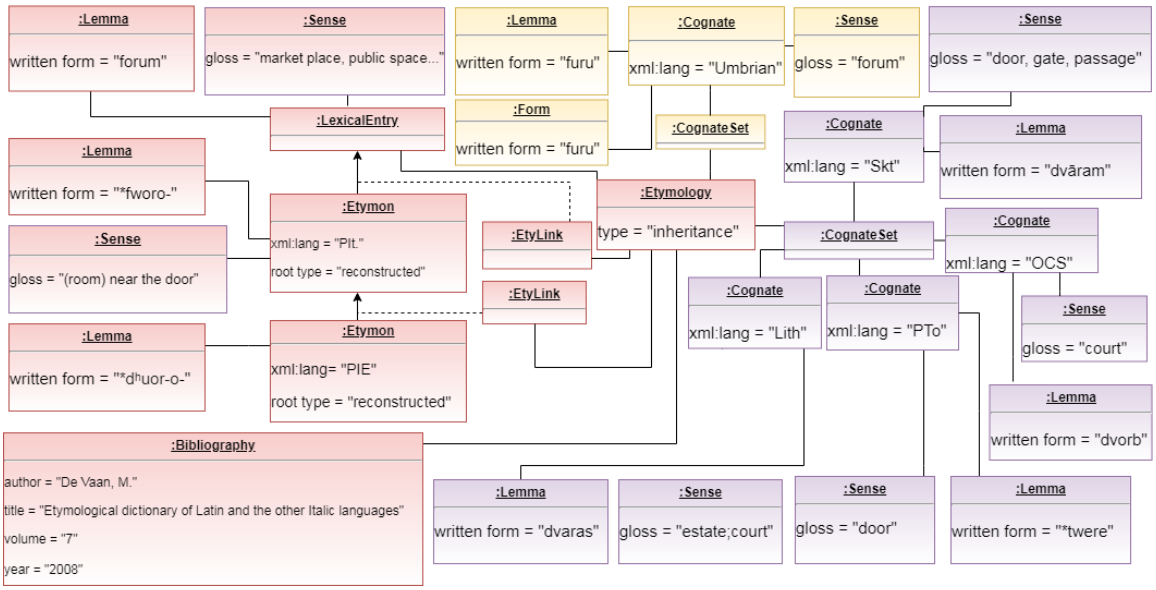


Figure 2: Encoding of the entry for *forum*.

define different etymologies featuring shared elements. In addition **Etymology** instances can be recursive, they can also be typed to define the changes undergone according to any number of linguistic processes.

Although we have had to leave out a number of details in this brief summary, the classes which we have enumerated above are the fundamental ones for understanding and using the standard. We were able to establish these classes over the course of numerous iterative design cycles during which draft proposals were reviewed against a large and diverse number of use cases: evaluating them on the basis of their salience, expressivity, and understandability. In Figure 1 we present the entry for the Latin word *forum* from De Vaan's *Etymological dictionary of Latin and the other Italic languages* (De Vaan, 2018), and in Figure 2 a partial encoding of this entry using LMF. Here we have focused chiefly on the information in the written entry which concerns etymons and cognates[8]. Note the relationship between the lexical entry and its two etymons (both of which have been categorized as reconstructed lexemes). We have also added two **Cognate Set** elements which, although we haven't shown it in the diagram, can be linked to their associated etymons. Note that these elements are linked to the LMF lexical entry for *forum* via an **Etymology** element which is in reality a container for an ordered set of **EtyLink** elements. It is important also to note that not all of the information in an etymology can be easily represented in a graph like structure and this we can instead represent in additional textual elements.

[8]There exists provision in LMF-Ety for representing information concerning attestations, references to secondary literature and for adding textual information as notes attached to entries or etymologies although we haven't presented it here. We also haven't added explicit temporal important to this example either. Full details will be available in the final version of the standard.

4 Acknowledgements

The first author was supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure).

References

Jack Bowers and Laurent Romary. 2016. Deep encoding of etymological information in TEI. *Journal of the Text Encoding Initiative* 10. https://jtei.revues.org/1643

C. Chiarcos, F. Abromeit, C. Fäth, and M. Ionov. 2016. Etymology meets linked data. a case study in turkic. In *Digital Humanities 2016. Krakow*.

Michiel De Vaan. 2018. *Etymological dictionary of Latin and the other Italic languages*, volume 7. LEIDEN·BOSTON, 2008.

Gil Francopoulo. 2013. *LMF lexical markup framework*. Wiley Online Library.

Anas Fahad Khan. 2018. Towards the representation of etymological data on the semantic web. *Information* 9(12):304.

Fahad Khan, Federico Boschetti, and Francesca Frontini. 2014. Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources. Proceedings of the Workshop on Linked Data in Linguistics 2014 (LDL-2014).

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. *The OntoLex-Lemon Model: Development and Applications*, pages 587–597. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Laurent Romary, Mohamed Khemakhem, Monte George, Jack Bowers, Fahad Khan, Mandy Pet, Stephen Lewis, Nicoletta Calzolari, and Piotr Banski. 2019. Lmf reloaded. In *Asialex 2019*.

Susanne Salmon-Alt. 2006. Data structures for etymology: towards an etymological lexical network. .

Workflows, Digital Data Management and Curation in the RETOPEA Project

Ilenia Eleonor Laudito
Leibniz Institute for European History (IEG)
Mainz, Germany
laudito@ieg-mainz.de

Abstract

English. The main aim of the RETOPEA project is to give insight into political and religious peace-making history offering a prism for interpreting contemporary social issues related to religious diversity and tolerance. The project emphasises a reflective learning from history, where teenagers with different cultural, national and ethnic backgrounds will actively interpret historical and contemporary information about past religious conflicts and the present representation of similar conflicts, in contrast to the classical schooling methods where students are merely passive receptors. To achieve this purpose the students will create a series of short films that will be published on the project's website among the background informations developed by the historical researchers. The project's aim demands a multilingual dataset to guarantee a proper understanding of the produced research data by the students residing in eight different European countries. Due to this requirement, not only all data but also its metadata necessitated a proper translation from English in seven different European languages. This paper describes the workflow and the procedures used in this on-going project and depicts the possibilities and the necessity of multilingualism and automatic translations, as well as the technical issues encountered concerning these topics.

Italiano. Il principale obiettivo del progetto RETOPEA è quello di fornire informazioni storiche sulla pace politica e religiosa, offrendo un prisma per l'interpretazione di attuali questioni sociali legate alla diversità religiosa e alla tolleranza. Il progetto pone l'accento su un apprendimento riflessivo della storia in contrasto ai classici metodi scolastici, in cui gli studenti sono meri recettori passivi. A tal fine gli adolescenti partecepanti al progetto, residenti in otto diversi paesi europei e dunque aventi diversa identità culturale, nazionale ed etnica, realizzaranno una serie di cortometraggi, rappresentando e confrontando attivamente conflitti religiosi passati e presenti. Sia i cortometraggi che il materiale elaborato dal gruppo di ricerca storica saranno pubblicati sul sito web del progetto. Al fine di garantire una corretta comprensione dei dati di ricerca, è necessario un dataset multilinguale. Perció tutti i dati di ricerca con i relativi metadati saranno tradotti dall'inglese nelle sette diverse lingue europee del progetto. Questo articolo descrive l'organizzazione del flusso di lavoro, le metodologie e le procedure utilizzate nel progetto e illustra la necessità di dati multilinguali e di traduzioni automatiche in progetti di umanistica digitale, soffermandosi sulle ulteriori possibilità di sviluppo di tali funzioni e sulle problematiche techniche incontrate nel corso del progetto.

1 Description and aim of the RETOPEA project

Funded by the European Commission under the program Horizon 2020, RETOPEA (REligious TOleration and PEAce) aims at creating a modern understanding of religious conflicts and peace-making history among youngsters and students throughout Europe. The intention is to teach in a comprehensible and appealing way, complex aspects of the past and present society. The project's target group are students between 12 and 18 years old, attending schools as well as non-academic institutions in European countries partnered with the project (which are Spain, UK, France, Belgium, Germany, Finland, Estonia and Macedonia).

Characteristic to this project is its mixture of a historical corpus of peace treaties and agreements – spanning from settlements prior to the anno domini to the most recent Charter of Fundamental Rights of the European Union – and contemporary political discourses, popular culture among teenagers, new spiritual initiatives and heritage. The materials selected and processed by the historical research groups (called "clippings"[1]) will be disclosed on the RETOPEA official website and will serve as primary resources and background information for the creation of short documentary films about the different aspects of tolerance and religious coexistence

[1] A clipping is a piece of information with a length of ca. 200 – 500 words, about a specific subject, possibly containing different media types and formats.

by the students participating to the project. These short films (called "docutubes") will be published on the project's online platform and can be used in the future for further teachings.

It is essential to the project's aim and purpose that all research data is correctly translated in the seven languages corresponding to the above-mentioned partnered European countries. This fundamental requirement assures that all students, independently from their knowledge of the topics presented and their level of understanding of the English language, are able to comprehend solidly the informations provided by the researchers.

2 Technical environment, workflow and data management

The Data Management Plan guarantees the storage sustainability and the long-term preservation of all relevant research data produced and processed by the historical research groups. The collected data will be publicly available via a Virtual Research Environment (VRE). Additionally, the designed workflow establishes a proper coordination between the historical research groups and the technical requirements of the project.

2.1 Technical enviroment

The technical specifications for the storage, preservation and visual presentation of the collected data consists of three main components: a collective access database (with an API to link material from other databases and platforms), a digital repository (TENEO) and a publishing tool (Omeka) that also serves as the project's official website.

Crucial for the publication platform's selection was a user-friendly environment for students, teachers and researchers. Omeka is a web publication platform for sharing digital collections and creating media-rich online exhibits (Omeka S User Manual, 2019). This tool is mainly used by universities, archives, museums and galleries and fits the project's specific demands, due to the heterogeneity of its data.

2.2 Workflow

The workflow divides the clipping's production process in four main stages: creation of the source clippings in English, automatic translation of the clippings, review of the automatic translations and upload by the research data manager into the VRE.

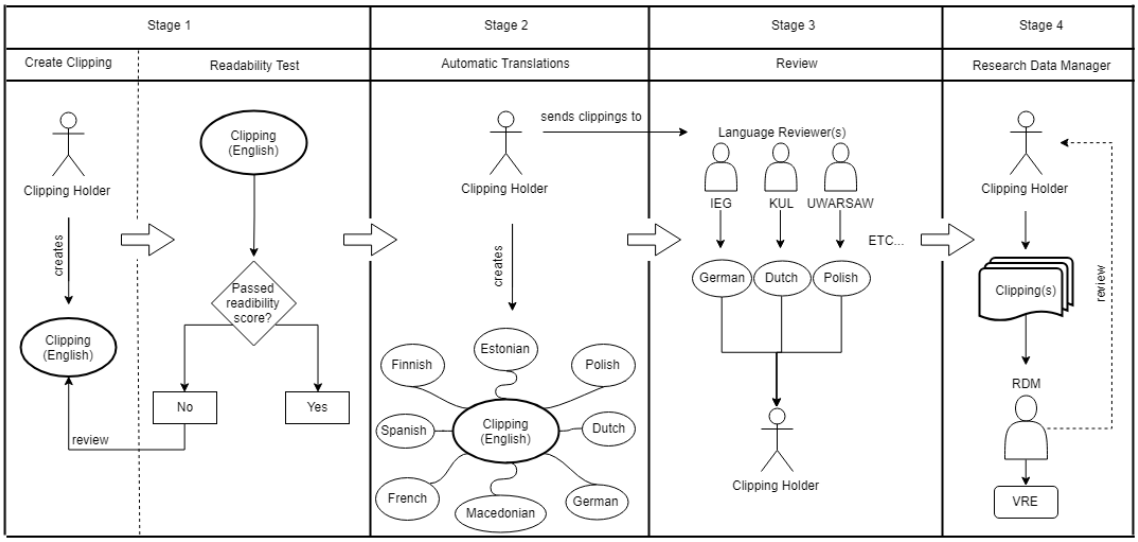


Figure 1: The four stages of the workflow.

All clippings require to be written in English to facilitate the translation of its content afterwards. Additionally, the clippings need to pass a readability test – which is based on the English language – to match the student's reading and grade level. The selected material differs widely in subject matter, field and case study. According to the analysis made on the clippings reading ease and grade level, RETOPEA safeguards an appropriate level of understanding of the complex topics selected by the researchers. The pedagogical importance of this step underlines once again the project's priority concerning the didactical value of the data produced.

The second stage concerns the automatic translation of the clippings, whereas the third stage concentrates on the review of the automatically translated content of the clippings. As explained, it is fundamental that all clippings are grammatically, linguistically and thematically comprehendible by youngsters. The current translation tools available are fallible, thus the need for a reviewing process by native language speakers. Nonetheless, the possibility to accelerate the workflow and lighten the workload of the translation process through automation is a major advantage. At the present state, the tools used for the automation were tested only on a small number of clippings, since the research groups must work on its production first. From the various web translators tested, the following tools will be used in the project: DeepL for Dutch, German, Spanish, French and Polish, Google Translate for Macedonian and Finnish and Tilde for Estonian. The most arduous languages to translate result to be Finnish and Macedonian, not only for the complexity of the language *per se*, but mostly for the lack of tools available.

In the final step the research data manager will upload the complete dataset in the VRE and make small adjustments concerning the clippings' visual representation based on the resources and formats used (text, video and/or audio), in concordance with the clipping's creators.

Albeit this workflow is efficient and functional to RETOPEA's work organisation, it ought to be adapted to fit other requisites and necessities, if used in different projects. There are mainly two things that should be taken into consideration: first, the volume of the research data produced; second, the financial aspect of a reviewing process should not be underestimated and adequately evaluated.

RETOPEA has a relatively small dataset of short text snippets (approximately 400 clippings) which can be uncomplicatedly handled and humanly reviewed. Due to this factor, the costs and extent of a reviewing process are limited. If a project has a broader dataset, then the human involvement and financial aspects should be carefully scrutinised, especially in the case of minor European languages. For example, as for the Finnish language review, it may be more convenient in terms of financial resources and time management to directly translate the data from the source language without going through an automatic translation.

This notwithstanding, one major advantage of this workflow is that it can be easily managed, exploited and followed by researchers, independently from their computer skills and know-how.

2.3 Metadata structure

Every clipping necessarily includes a written part and may contain different media types in various formats beyond its written content (e.g. additional textual resources, images, video and/or audio material, YouTube or other URLs, etc…). To facilitate both the researchers' and the research data manager's work, the researchers will fill the clipping's metadata in two simple spreadsheets, containing the tags selected by the research data manager in the column header. The spreadsheets will be pasted and exported to CSV format and uploaded in the Omeka environment. Given the dataset's heterogeneity in terms of composition of resources, the development of a project-wide metadata scheme to normalise the metadata records was mandatory.

The metadata standards and controlled vocabulary used in the project are the *Dublin Core Metadata Standard* (DC), the *Bibliographic Ontology* (BIBO), *GeoNames* (GN) and the *Canadian Writing Research Collaboratory Ontology* (CWRC). Omeka provides an automapping module that maps the metadata terms of the spreadsheet's header to the imported vocabularies (Omeka S User Manual, 2019).[2] The module also allows to associate a media source (e.g. HTML, URL, YouTube link, etc…) to a selected metadata tag. In RETOPEA's case, the clippings' contents are ingested as HTML-code through the Bibliographic Ontology Term <Content>. The tag will be hidden afterwards and will not be displayed as metadata, yet the content will be visible as media attachment.

[2] For example, "dcterms:title" is automatically mapped to the Dublin Core <Title> property.

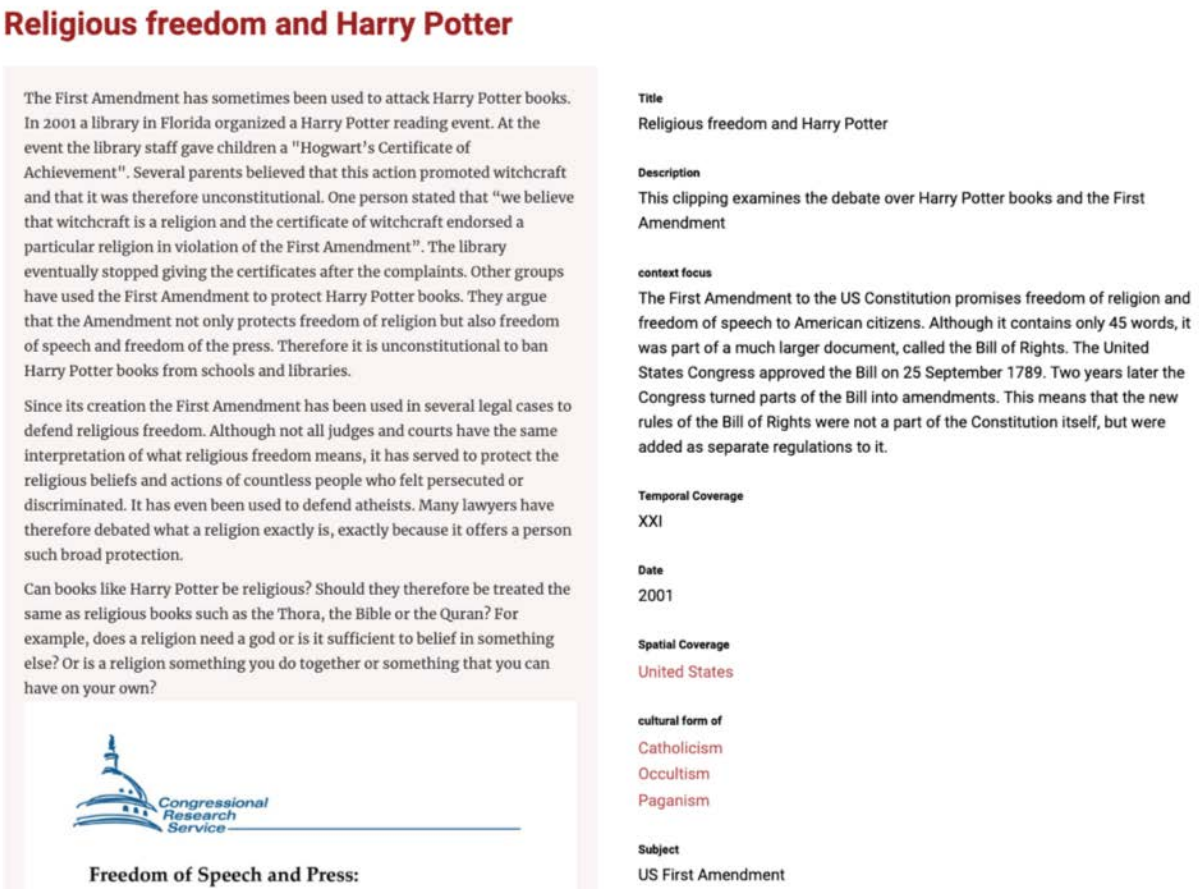


Figure 2: Draft of a clipping as displayed on the RETOPEA website.

Further, Omeka allows to aggregate the ingested data in user-made collections. The twelve collections used in this project aim at grouping the clippings into abstract thematic classifications of project-relevant keynotes. The collections used in RETOPEA represent generic topical focuses (e.g. "Religious practice", "Gender and Sexuality", "Propaganda and stereotyping", etc…), to which clippings can belong independently of their subject. Subjects differ from the topical focuses, for the latter have a broader thematic range and may apply to an indefinite number of clippings, whereas the intent of the subject's list is to bundle a relative small number of clippings into strictly defined subjects (e.g. "Peace of Augsburg", "Edict of Nantes", "YouTube channels", "Political speeches", etc…). The same clipping can appear in more than one collection, depending on how many relations the researcher associated to the clipping. This type of arrangement constructs an intricate entanglement between clippings in order to create vast links and relations between clippings that do not share the same subject matter. These relations and clusters belong to and are part of the metadata description, creating both a vertical and a horizontal hierarchical structure. The main purpose of this structure is to drive the website users and the teenagers using the provided clippings to produce their docutubes to discover as many clippings as possible, independently of the clipping's affiliation or subject matter.

Besides the clipping's title, description, contextual focus and content, the implemented BabelNet API will automatically extract and translate all other metadata tags and keywords through an HTTP interface that returns JSON (Navigli and Pozzetto, 2012). BabelNet not only functions as an online translator, but also recognises synonyms, word sense and (multilingual) semantic relatedness, shaping the possibility to generate linked data and semantic networks.

The described metadata structure was designed specifically for RETOPEA and determined by the intensive

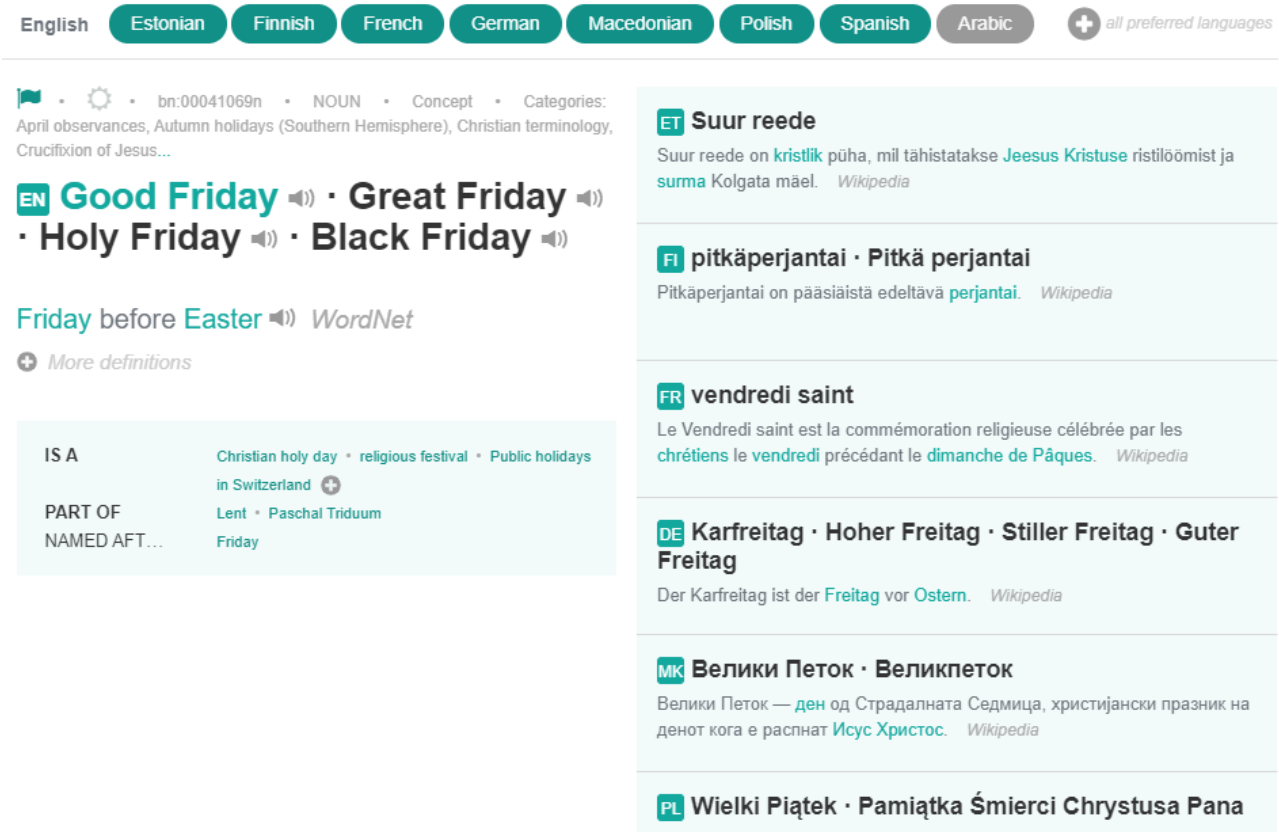


Figure 3: BabelNet translation of "Good Friday".

collaboration, confrontation and discussion with the two historical research groups. Due to the project's distinctive didactical purpose and sundry data, it contains peculiar arrangements that are not always feasible or desirable in other DH-projects. This notwithstanding, this structure could be readily adopted and accordingly modified to fit other requirements.

Considering RETOPEA's didactical and pedagogical purpose, the project's resources will be disclosed under the Creative Commons Licenses (i.e. CC BY-NC-SA). Most of the external materials used in the project can be likewise used and remixed by third parties. Additionally, future tasks will concern the implementation of external databases, like the IEGs "Maps" and "European History Online" (EGO) Databases, and the "On site, in time" project.

Further developments in RETOPEA will give a more precise evaluation about the translation tools used and the metadata structure. Moreover, the controlled vocabularies used leave open the possibility to connect, organise, retrieve and interlink the project's resources in Linked Open Data.

3 Innovation possibilities and DH importance

The methodology and the workflow described in this paper aims at giving a suggestion on how humanistic projects can organise and arrange the digital data produced and the immense possibilities that Natural Language Processing tools and approach may offer.

In the last years the availability and growing amount of data extremely increased, also through the thriving of Digital Humanities related projects. The need for automatically translated documents, data and metadata will increase as more DH-projects arise worldwide. This need does not only apply to major and minor European languages, but also to Arabic and Asiatic languages as well as dialects.

Acknowledgements

The research was supported by the Leibniz Institute for European History (IEG) and the Religious Toleration and Peace (RETOPEA) research project. This paper is based upon work supported and funded by the European Commission under the funding program Horizon 2020, Grant CULT-COOP-05-2017. Special thanks go to Marco Büchler, who provided insight and expertise and collaborated to the development of the workflow. Further gratitude goes to Bram De Ridder, who wrote the clipping shown in "Figure 2", for granting and permitting the publication of his draft example.

References

Navigli Roberto and Ponzetto Simone Paolo. 2012. *Multilingual WSD with Just a Few Lines of Code: the BabelNet API*. Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Korea, July 9-11, pp. 67-72. http://wwwusers.di.uniroma1.it/~navigli/pubs/ACL_2012_Navigli_Ponzetto.pdf

"Omeka S User Manual". *Omeka S*, Corporation for Digital Scholarship, Roy Rosenzweig Center for History and New Media, George Mason University. Accessed 15.09.2019. https://omeka.org/s/docs/user-manual/

Il confronto con Wikipedia come occasione di valorizzazione professionale: il case study di Biblioteca Digitale BEIC

Lisa Longhi
Biblioteca digitale BEIC
lisa.longhi@BEIC.it

Abstract

Italiano. Il confronto è un'occasione imprescindibile per la crescita del bibliotecario. Se il dialogo con i colleghi permette di affinare le proprie competenze nei flussi di lavoro tradizionali, la cooperazione con un universo che un tempo appariva distante, come quello di Wikipedia, è in grado di produrre uno scambio di abilità e punti di vista, utili all'apertura dei reciproci orizzonti e quindi alla circolazione di conoscenza e contenuti culturali. La partnership tra BEIC e Wikimedia Italia ha prodotto nel 2014 un progetto GLAM-wiki (Galleries, Libraries, Archives and Museums), tuttora attivo che costituisce un case study significativo: lo staff bibliotecario, già caratterizzato al suo interno da profili formativi diversi, ha avuto la possibilità di comprendere la filosofia dell'enciclopedia libera, di Wikidata e degli altri progetti basati sull'open content, entrando nell'ottica della contribuzione e della valorizzazione del proprio patrimonio digitale.

1 *Cosmetica del nemico* o delle diverse comunità

L'attendibilità dei contenuti di Wikipedia continua ad essere oggetto di accesi dibattiti che vedono come protagonisti su un fronte i bibliotecari coinvolti nella redazione delle voci dell'enciclopedia, sull'altro i bibliotecari che ancora se ne tengono a distanza[1]. Esiste ancora un solco che si pensava colmato. *Cosmetica del nemico* è il titolo di un breve romanzo di Amélie Nothomb, dove la parola 'cosmetica' va intesa nel suo significato semantico, ovvero «la morale suprema che determina il mondo». Il problema è che ogni mondo ha la sua.

La prima reazione, una decina di anni fa, all'idea di una collaborazione tra bibliotecari e wikipediani è stata infatti quella di collocare le due realtà agli antipodi: da un lato una professione con tradizioni secolari e normative condivise per il trattamento delle risorse bibliografiche; dall'altro una comunità giovane con pochi principi operativi ed etici - i cinque pilastri - che nascono dall'esigenza di un sapere libero[2]. Negli ultimi anni è apparso sempre più chiaro il fatto che questa impressione derivasse da un pregiudizio: sebbene infatti uno dei pilastri reciti che «Wikipedia non ha regole fisse», esistono molti punti di contatto fra le tradizioni bibliotecarie e le prassi di Wikipedia, laddove le intenzioni programmatiche di garantire l'accesso libero alla conoscenza si riflettono nell'utilizzo di strutture a schema fisso - come template e legami con URI - per organizzare le informazioni. Inutile sottolineare che si tratta di strumenti assai noti a chi si occupa di information literacy e di cataloghi bibliografici[3].

2 *Lo straniero* o dei diversi bibliotecari

Questo contributo prende tuttavia le mosse da un processo di crescita più lontano. Negli ultimi anni si è discusso a lungo sui percorsi formativi utili a diventare bibliotecari e sulla necessità di acquisire competenze che sviluppino capacità storico-umanistiche, ma anche una formazione tecnica e tecnologica. L'obiettivo è quello di aderire a un modello di bibliotecario ampio e complesso, per essere riconosciuti come bibliotecari in Italia e in Europa. A tal fine, si sta cercando di delineare un iter formativo unico e accreditato[4].

[1] Uno degli ultimi dibattiti è apparso a inizio 2019 sul forum di discussione *Humanist*, poi diffuso dalla lista AIB-CUR: i messaggi riguardanti Wikipedia sono raccolti nel thread *the question on Wikipedia*, <https://dhhumanist.org/volume/32/>.

[2] Cinque pilastri, <https://it.wikipedia.org/wiki/Wikipedia:Cinque_pilastri>.

[3] Esiste una nutrita bibliografia che attesta il dialogo e i risultati di una virtuosa collaborazione realizzata e possibile, dimostrando vicinanza sul piano operativo ma anche affinità metodologica. Questi argomenti sono raccolti per esempio in: L. Catalani-P. Feliciati, *Wikipedia, le biblioteche e gli archivi / Wikipedia, Libraries and Archives*, «JLIS.it», 9, 3 (September 2018), p. I-III, <https://www.jlis.it/issue/view/789>.

[4] L'emanazione della l. 14/1/2013, n. 45 ha sancito in Italia l'obbligo di rintracciare dei criteri univoci, relativi alle conoscenze e alle competenze, che ogni professionista intellettuale. Per una disamina della legge e dei suoi effetti in ambito bibliotecario, si veda: R. De Magistris, *Il riconoscimento delle professioni non regolate e la legge n. 4 del 14 gennaio 2013*, «AIB studi», 53, 3 (2013), p. 239-260, <http://aibstudi.aib.it/article/view/9074>.

Tuttavia, benché esistano pubblicazioni che descrivono quali siano le tappe di questo percorso[5], benché l'Associazione Italiana Biblioteche si sia fatta carico di rilasciare l'attestato di qualità dei servizi presso il Ministero dello sviluppo economico[6], benché l'ICCU abbia intensificato il piano di formazione e aggiornamento professionale[7], ci troviamo ancora di fronte a numerosi singoli bibliotecari che mettono a frutto le loro diverse competenze, risultato di percorsi di studio eterogenei e non sempre pensati nell'ottica della professione bibliotecaria. A questo proposito, è significativo il caso dei 'bibliotecari-studiosi', vale a dire di quei, non pochi, bibliotecari provenienti dal campo della ricerca, che prestano le loro competenze di paleografi, musicologi, storici dell'arte, al lavoro di catalogatore, facendosi portavoce all'interno della biblioteca delle reali necessità degli utenti più specialisti. L'utente-studioso può trovare in questo tipo di bibliotecari delle alleanze: il catalogatore esperto infatti, seppur 'straniero di nascita', a differenza del protagonista di Camus, che non si integra nella sua stessa vita, sposa la missione del bibliotecario come facilitatore dell'apprendimento del lettore. Da ricercatore, conosce i manoscritti, i libri antichi o le partiture musicali e sa quali siano i pochi dati certi che non devono cambiare mai in un catalogo (titoli uniformi, voci di autorità normalizzate), ma è anche in grado di fornire quei dati sensibili che spesso si trovano fuori dai campi di ricerca, rispondendo così a un bisogno informativo specifico[8].

3 *I custodi del libro* o della collaborazione con altre istituzioni

I custodi del libro di Geraldine Brooks è un romanzo che racconta la storia di un libro - un codice ebraico del XV secolo - ma è anche la storia delle persone che hanno confezionato il libro nelle sue parti e soprattutto delle persone, i bibliotecari (cui il libro è dedicato), che l'hanno tutelato e tramandato nei secoli. E' una storia che invita alla collaborazione tra bibliotecari e tra biblioteche, come stimolo alla crescita e arricchimento, nella salvaguardia della cultura.

Aprirsi al confronto in un panorama di cooperazione nazionale e internazionale è infatti una grande occasione per chi lavora nel settore dei beni librari, poiché reca giovamento non solo ai singoli bibliotecari, attraverso il confronto coi colleghi, ma anche alle istituzioni di appartenenza, che donano maggiore visibilità e fruibilità al loro patrimonio. Per le collezioni digitali poi il confronto e la cooperazione diventano necessari, poiché il rischio di confondersi nella rete e di rendere i dati poco reperibili diviene esponenziale.

I progetti di cooperazione sono molteplici: possiamo citare Internet Culturale[9], Europeana[10], WorldCat[11], il CERL[12]. Far parte di queste reti offre ai bibliotecari il vantaggio di favorire la circolazione delle informazioni, la visibilità dei dati e l'opportunità del confronto e della collaborazione con gli altri professionisti della rete. Tuttavia per favorire dialogo e cooperazione è necessario parlare la stessa lingua, ovvero adottare preventivamente standard aggiornati che permettano l'interoperabilità e il riuso di metadati affidabili in un'ottica internazionale.

4 *Il filo del rasoio* o del confronto con il mondo Wiki

"Camminare sul filo del rasoio è difficile", dice un passo delle Upanishad: in quel contesto - come nel romanzo di Somerset Maugham, che ne trae titolo e ispirazione - ci si riferisce allo stare in equilibrio tra senso pratico e spiritualità. Tra due piani diversi, tra due realtà. Se il bibliotecario-studioso già deve esercitare il proprio equilibrio, cimentandosi, da ricercatore, negli scenari più 'tradizionali' della creazione e dello scambio di

[5] Associazione italiana biblioteche, *Il portfolio delle competenze: un nuovo strumento per il professionista dell'informazione*, a cura dell'Osservatorio Formazione (coordinatore Patrizia Lùperi); contributi di Manuela De Noia, Matilde Fontanin, Patrizia Lùperi, Roma, Associazione italiana biblioteche, 2017.

[6] Ministero dello sviluppo economico, Associazioni professionali che rilasciano l'attestato di qualità dei servizi. 2016,<http://www.sviluppoeconomico.gov.it/index.php/it/cittadino-e-consumatori/professioni-non-organizzate/associazioni-che-rilasciano-attestato-di-qualita>.

[7] Il piano di formazione e aggiornamento professionale ha previsto, per il 2019, ben due cicli di corsi di aggiornamento organizzati dall'ICCU indirizzati al personale bibliotecario specializzato, <https://www.iccu.sbn.it/it/attivita-servizi/formazione-e-didattica/Corsi-di-formazione-attivi/>.

[8] Chiara Cauzzi [et al.], *Conoscersi per riconoscersi: la partecipazione come specchio del bibliotecario,* «AIB studi», 56, 2 (maggio/agosto 2016), p. 219-239, <http://aibstudi.aib.it/article/view/11462>.

[9] Internet Culturale, <http://www.internetculturale.it/>.

[10] Europeana Collection, <https://www.europeana.eu/portal/it>.

[11] What is WorldCat?, <http://www.worldcat.org/whatis/default.jsp>.

[12] Consortium of European research libraries, <https://www.cerl.org/>. Il Consorzio promuove l'attività di ricerca attraverso gruppi di lavoro internazionali come Heritage of the Printed Book Database (<https://www.cerl.org/resources/hpb/main>), CERL Thesaurus, <https://data.cerl.org/thesaurus/_search>), Material evidence in incunabula (<http://data.cerl.org/mei/_search>).

metadati bibliografici, la difficoltà aumenta quando il bibliotecario si trova a lavorare in ambienti apparentemente estranei.

I progetti GLAM (acronimo per Galleries, Libraries, Archives, Museums, analogo internazionale dell'italiano MAB) sono nati formalmente nel 2010, proprio con l'intenzione di promuovere la collaborazione tra istituzioni culturali e Wikipedia, integrando bibliotecari, archivisti e curatori museali tra i propri contributori, ma alcuni tra questi professionisti del settore culturale avevano già contribuito - individualmente, come utenti - ai contenuti dell'enciclopedia libera fin dalla sua nascita nel 2001[13]. Da quindi una decina di anni il mondo GLAM collabora con le comunità wikipediane di tutto il mondo per l'arricchimento dell'enciclopedia libera sui temi del patrimonio culturale: collaborazioni che vanno dalla semplice "donazione" di dati (scansioni, fotografie d'archivio, dati bibliografici) a partnership più strutturate che prevedono il "wikipediano in residenza", figura di mediazione che collabora a stretto contatto con l'istituzione culturale, formando il personale su come contribuire a Wikipedia e agli altri progetti Wiki[14].

Abbiamo detto che l'interazione tra le due comunità può essere utile alla crescita di entrambe, se la si sa osservare senza pregiudizi: Wikipedia è ormai una delle principali fonti di accesso all'informazione per molti utenti del web, ma l'enciclopedia ha bisogno di colmare le lacune nel settore dei beni culturali, proponendosi nel contempo di dare risalto al lavoro di archivisti e bibliotecari, altrimenti poco visibile in rete (con materiale digitale visibile solo dal sito dell'istituzione, ma non rintracciabile con ricerche sui motori di ricerca): in pratica sono i bibliotecari stessi, gli archivisti e gli operatori di musei a creare le voci relative ai propri patrimoni e ai relativi ambiti disciplinari, oltre a condividere le proprie immagini digitali con licenze libere e aperte (Creative Commons CC-BY e CC BY-SA) o nel pubblico dominio (CC0). La collaborazione è infatti finalizzata a condividere contenuti e risorse in un'ottica di promozione della conoscenza libera e le biblioteche digitali possono contribuire ulteriormente, fornendo ad esempio fonti bibliografiche verificabili alle voci dell'enciclopedia.

5 *I doni della vita* o dei beni comuni digitali

I doni della vita è un romanzo di Irène Némirovsky e i doni cui fa riferimento l'autrice sono i solidi valori e le certezze di una famiglia che, in un momento di difficoltà materiale, si trasformano in forza. Fuori di metafora, occorre riconoscere l'opportunità di miglioramento offerta dalla contribuzione, bisogna avere la consapevolezza del fatto che aggiungere qualità alle voci create o arricchite, e contemporaneamente beneficiare di nuove conoscenze tecniche, è fonte di aggiornamento professionale e crescita personale. Una volta fugato il pregiudizio che porta ad arroccarsi nella contezza della propria professionalità e scegliere di non contribuire, può verificarsi il rischio, ugualmente insidioso, di contribuire, usando la rete come una cattedra.

Essendo Wikipedia uno dei siti d'informazione più visitati al mondo ed essendo un progetto no profit aperto a tutti, ha inevitabilmente attirato a sé coloro che cercano visibilità sul web, trasformando la sua apertura in vulnerabilità. Per arginare il rischio dell'autopromozione da un lato i wikipediani si sono dotati di una burocrazia "difensiva"[15], dall'altro occorre che i professionisti dell'informazione - insegnanti, docenti universitari, bibliotecari, archivisti - si premurino di partecipare in un'ottica neutrale perché le informazioni di Wikipedia siano il più accurate possibile. È auspicabile quindi una sempre maggiore contribuzione da parte degli addetti ai lavori: un bibliotecario, ancor più un un bibliotecario studioso, dovrebbe dare una mano a questo progetto di condivisione nella sua globalità, secondo il concetto di *filiera del dato open*, ben descritto da Andrea Zanni in un suo recente contributo[16], in cui si dimostra quanto "la partecipazione a un bene comune digitale diventa un mezzo per il bibliotecario di adempiere alla sua missione, andando a fornire le informazioni direttamente dove gli utenti della rete sono presenti, attraverso le competenze, le collezioni e i valori propri della professione bibliotecaria". Del resto l'utilizzo di protocolli e formati aperti, di licenze e policy aperte garantisce l'interoperabilità necessaria fra macchine e macchine, ma anche fra umani e umani.

[13] GLAM (cultura). In: Wikipedia: l'enciclopedia libera, <https://it.wikipedia.org/wiki/GLAM_(cultura)>; in particolare Progettto:GLAM/biblioteche, <https://it.wikipedia.org/wiki/Progetto:GLAM/Biblioteche>.

[14] Progetto:GLAM/Wikipediano in residenza, <https://it.wikipedia.org/wiki/Progetto:GLAM/Wikipediano_in_residenza>.

[15] Wikipedia:Contenuti promozionali o celebrativi, <https://it.wikipedia.org/w/index.php?title=Wikipedia:Contenuti_promozionali_o_celebrativi&oldid=88978752>.

[16] A. Zanni, *Le biblioteche e la filiera dell'open*, «JLIS.it», 9, 3 (September 2018), p. 75-91, <https://www.jlis.it/article/view/12486>.

6 *Le affinità elettive o* del metodo di lavoro

Le affinità elettive di cui parliamo sono quelle tra Fondazione BEIC e Wikipedia. Anche Goethe aveva desunto il titolo del suo romanzo dall'affinità chimica, cioè la tendenza di alcuni elementi chimici a legarsi tra loro. Nel caso di BEIC e Wikipedia l'affinità nasce da una comunione di intenti: le due istituzioni hanno infatti l'obiettivo comune di promuovere e condividere contenuti e risorse.

Dal 2008 la Fondazione BEIC[17] sta realizzando una biblioteca digitale che raccoglie collezioni tematiche, selettive e multidisciplinari[18]. Per le digitalizzazioni si è affidata a studiosi e a istituzioni di prestigio che, attingendo a raccolte italiane e straniere, hanno selezionato gli esemplari in base a specifici criteri, legati alla rilevanza degli autori e delle opere e al carattere internazionale delle fonti. E' stata dedicata molta attenzione alla creazione di un catalogo amichevole che ponesse particolare cura nella presentazione dei dati e nella loro integrazione con altri cataloghi, così come nella rapida visualizzazione e nella facile manipolazione delle immagini[19]. In questo modo la Biblioteca Digitale BEIC si propone di rendere liberamente accessibile un vasto complesso di opere umanistiche e scientifiche, in un arco temporale che va dal medioevo all'età contemporanea[20]. Lo staff di Biblioteca Digitale BEIC, eterogeneo già al suo interno, ha una routine di lavoro che prevede il confronto quotidiano con cataloghi nazionali (SBN OPAC, Edit16) e internazionali (Karlsruhe, GW, ISTC), con database per il controllo delle voci di autorità (VIAF, CERL Thesaurus) e con strumenti enciclopedici (Dizionario biografico degli italiani, Wikipedia). Questa attività di consultazione ha portato alla consapevolezza di dover contribuire alle reti che si ritengono efficaci: la Biblioteca Digitale BEIC collabora quindi da tempo ad alcuni progetti, attraverso il riversamento - possibile grazie all'interoperabilità degli standard MARC - di registrazioni della biblioteca digitale in WorldCat, Europeana e ISTC.

L'obiettivo è quello di valorizzare le proprie collezioni in un contesto internazionale e di permettere il massimo riutilizzo dei metadati, per questo rilasciati con licenza CC0 (pubblico dominio). Queste caratteristiche rendono la Biblioteca Digitale BEIC particolarmente adatta alla collaborazione con Wikimedia Italia. Dal 2014 è stato così avviato un progetto GLAM[21], che, nell'ottica di una condivisione aperta dei contenuti e di diffusione della conoscenza libera, ha il fine di disseminare le risorse della Biblioteca Digitale all'interno di Wikipedia e dei progetti fratelli. Dall'inizio del GLAM a oggi, le attività svolte dalla BEIC insieme al Wikipediano in residenza - inizialmente Federico Leva, poi Marco Chemello - hanno spaziato dall'arricchimento delle voci dell'enciclopedia, al caricamento di immagini digitalizzate in Wikimedia Commons, alla collaborazione con i progetti di Wikisource[22]. Negli ultimi anni, tuttavia, è cresciuta la consapevolezza che uno in particolare fra i progetti sarebbe stato di grande interesse dal punto di vista bibliotecario: Wikidata[23]. Wikidata è un database libero che si basa su principi storici e fondamentali della teoria della catalogazione, come il controllo di autorità, e li affianca alla più recente ottica Linked Open Data (LOD). L'affinità di Wikidata con i modelli per la costruzione di registrazioni di autorità nei cataloghi è confermata dal fatto che le entità descritte in Wikidata sono arricchite dagli identificativi persistenti provenienti da dataset bibliografici autorevoli, come quello della Bibliothèque nationale de France o del CERL Thesaurus.

Nel 2017, dunque, si è cominciato a procedere al riversamento in Wikidata dei metadati sottoposti a controllo di autorità della Biblioteca digitale BEIC, a partire dagli autori persona che ricorressero come intestazioni principali dei record bibliografici (circa 5000 nomi), seguiti poi dagli autori persona con

---

[17] Biblioteca europea di informazione e cultura, Biblioteca digitale, <http://www.BEIC.it/it/articoli/biblioteca-digitale>. Per le tappe del progetto: *La Biblioteca europea di Milano (BEIC): vicende e traguardi di un progetto*, a cura di Antonio Padoa-Schioppa, Milano, Skira, 2014.

[18] Biblioteca europea di informazione e cultura, Le collezioni, <http://www.BEIC.it/it/pagina/le-collezioni>.

[19] Il protocollo catalografico BEIC ha come primo punto di riferimento le Regole italiane di catalogazione: REICAT, integrate con norme per il trattamento di risorse particolari (ad esempio, per la musica, ci si attiene a: Istituto centrale per il catalogo unico, Titolo uniforme musicale: norme per la redazione, Roma, ICCU, 2014, allo standard MARC21 e alle indicazioni di RDA.

[20] Attualmente la Biblioteca Digitale BEIC conta circa 39.569 oggetti digitali per un totale di oltre 99.936 registrazioni, che spaziano per tipologia dagli incunaboli alle opere d'arte, dai libri antichi ai periodici, dai manoscritti alle risorse audio-video.

[21] Progetto:GLAM/BEIC, <https://it.wikipedia.org/wiki/Progetto:GLAM/BEIC>.

[22] Sono state caricate in Wikimedia Commons le immagini di migliaia di opere presenti nella Biblioteca Digitale BEIC in formato TIFF a 300 o 400 dpi; le immagini sono state inserite in migliaia di voci esistenti di Wikipedia in italiano e di altre 200 versioni linguistiche; sono stati aggiunti riferimenti bibliografici a oltre 1600 voci in italiano, aggiungendo i link agli oggetti digitali consultabili nella Biblioteca Digitale BEIC; sono state create oltre 650 nuove voci in italiano e inglese di autori; infine, sono stati creati in Wikisource gli indici di numerose opere in italiano, provenienti da Internet Archive o dalla stessa Biblioteca Digitale BEIC. Si veda: C. Consonni-F. Leva, *Progetto GLAM/BEIC*, «Biblioteche Oggi», 33 (marzo 2015), p. 47-50, <http://www.bibliotecheoggi.it/rivista/article/view/24/265> ; F. Leva-M. Chemello, *The effectiveness of a Wikimedian in permanent residence: the BEIC case study*, «JLIS.it», 9, 3 (September 2018), p. 141-147, <https://www.jlis.it/article/view/12481>.

[23] <https://www.wikidata.org/wiki/Wikidata:Main_Page>

intestazione secondaria (quasi 15000 nomi)[24]. La fase preliminare al riversamento dei dati ha comportato la preparazione dei metadati di partenza: la Biblioteca Digitale BEIC utilizza lo standard MARC21, pertanto tutte le informazioni relative agli autori di tipo persona sono state rintracciate all'interno dei tag 100 (per le intestazioni principali) e 700 (per le intestazioni secondarie). Il fatto che il protocollo catalografico BEIC preveda in ogni caso l'inserimento delle date nel record di autorità (e non soltanto per disambiguare gli omonimi) ha permesso un'identificazione più certa nel corso del riversamento in Wikidata e ha assicurato la presenza di un dato stabile e di tipo *enciclopedico* in tutte le registrazioni. I metadati associati agli autori persona sono stati esportati dal catalogo, rielaborati e riversati in Mix'n'match[25], un tool wiki che permette di confrontare i record del catalogo con gli elementi esistenti in Wikidata e abbinarli nel caso si riferiscano alle stesse entità. Gli elementi *abbinati*, cioè i casi in cui i nomi del catalogo BEIC corrispondevano a un elemento già esistente in Wikidata, sono stati arricchiti di una nuova proprietà che evidenziasse la corrispondenza con le singole entità del Catalogo BEIC (proprietà 'Descritto nella fonte'); per gli elementi *unmatched*, ovvero i nomi presenti nel catalogo BEIC che non avevano alcun abbinamento con entità di Wikidata (circa il 30% del totale), si è resa necessaria la creazione da zero di elementi Wikidata. In entrambi i casi, vista la mole di dati e la difficoltà di procedere manualmente alle modifiche e alle creazioni, si è deciso di utilizzare QuickStatements, un altro tool che permette di intervenire in Wikidata in maniera semiautomatica: partendo da un elenco di elementi Wikidata, questo strumento è in grado di inserire lo stesso tipo di informazione in tutti con un'unica operazione[26]. Le nuove entità sono state poi arricchite di nuove informazioni, come il genere o l'occupazione, attingendo le informazioni da fonti diverse. Nel caso dell'occupazione, il dato talvolta era già stato inserito nel record di autorità (per esempio, la qualifica di santo, papa, imperatore o elementi disambiguanti per le forme omonime)[27], talaltra, il dato si è basato sull'analisi dei relator code associati ai tag 100 e 700 del record bibliografico[28].

Il risultato di questo lavoro vede per ora l'entità "Biblioteca digitale BEIC" associata a quasi 15000 elementi Wikidata e fotografa lo stato attuale del progetto che, come si è detto, è ancora molto ampio. Una volta concluso il progetto sugli autori di tipo persona, l'obiettivo è quello di estendere la prassi descritta anche agli altri metadati sottoposti al controllo di autorità (autori ente, editori, luoghi e titoli): si sta già lavorando sulle forme normalizzate degli editori antichi e moderni – includendo anche quelli presenti nel catalogo dell'Archivio della Produzione Editoriale Lombarda, sempre allestito dalla BEIC – e sui titoli. Una volta raggiunto l'obiettivo di esportare i dati di autorità dal catalogo a Wikidata, la prospettiva sarà quella di integrare le informazioni presenti nelle entità di Wikidata all'interno dei record di autorità del catalogo, a partire dagli identificativi persistenti e dalle forme varianti del nome. Un'ulteriore prospettiva sarà quella di migliorare la struttura dei record di autorità del catalogo BEIC, creando per ciascuno di essi un identificativo persistente, che diventerà il riferimento stabile al di fuori del catalogo, in primo luogo negli elementi Wikidata, nel rispetto di un'ottica Linked Open Data.

L'incontro (e talvolta scontro) tra catalogo BEIC e Wikidata ha dato risultati molto positivi: l'esposizione dei dati di autorità della Biblioteca digitale BEIC al di fuori del contesto del catalogo ha permesso il loro arricchimento, ha messo alla prova la loro struttura profonda in termini di interoperabilità e possibilità di riuso. In generale, la contribuzione ai contenuti di tutti i progetti Wikimedia ha portato notevole visibilità agli oggetti digitali, sia nei siti Wikimedia (con 23 milioni di visualizzazioni mese) sia nel sito di provenienza: a novembre 2019 il 75% delle visite al portale BEIC proveniva dai progetti Wikimedia. Quindi il progetto GLAM-wiki, come CERL, Europeana, ISTC, continua a essere per la Biblioteca Digitale BEIC un modo per guardare oltre i propri confini e per amplificare le competenze delle persone che lavorano al suo interno.

---

[24] L'eterogeneità delle collezioni della Biblioteca Digitale BEIC si riflette sull'insieme di autori che le rappresentano: autori classici greci e latini, sconosciuti commentatori medievali, incisori e disegnatori Seicenteschi, fotografi del Novecento, direttori d'orchestra e musicisti contemporanei.

[25] Creato da Magnus Manske, uno dei membri della comunità wiki internazionale.

[26] I dati sono stati impostati secondo una struttura che accoppia la proprietà "Descritto nella fonte" (P1343) al valore "Biblioteca digitale BEIC" (Q51955019), a sua volta arricchito da due riferimenti: la forma del nome così come ricorre nel Catalogo BEIC e l'URL che lancia una ricerca corrispondente nel Discovery tool della biblioteca digitale. Nel secondo caso, gli elementi da creare non avrebbero dovuto contenere solo la coppia proprietà-valore relativa alla presenza nella Biblioteca digitale BEIC, ma anche alcune informazioni canoniche usate per descrivere le persone: nome, cognome, date di nascita e morte. Queste informazioni sono state desunte con facilità dai metadati bibliografici presenti nel catalogo e sono state successivamente normalizzate in una struttura adatta all'immissione in Wikidata.

[27] Emblematico il caso di "Francesco Rossi", nome presente nel Catalogo BEIC che si riferisce a quattro diverse persone, distinte proprio in base alla loro professione: chirurgo, egittologo, filologo, giurista.

[28] Il relator code è un codice standard che permette di associare al nome dell'autore il ruolo che tale autore ha rivestito nell'ambito della risorsa descritta, ad esempio curatore, editore, illustratore, regista.

7 *L'opera struggente di un formidabile genio* o della crescita del bibliotecario

In conclusione, ho scomodato uno dei titoli più evocativi della letteratura contemporanea, perché il protagonista cerca di capire le regole delle cose, di smontare i pezzi e di metterli in ordine, per mostrarli in un modo diverso e più semplice. E lo fa in una logica classificatoria, corredata di indici tematici, semplificati in un elenco di parole-chiave. E questo mi sembra un buon modo per parlare la lingua dei bibliotecari.

Quello che mi stupisce della discussione apparsa in rete, e di cui abbiamo parlato in apertura, è la posizione di chi, tra i bibliotecari, vede a priori poca qualità in uno strumento di pubblica utilità, un'enciclopedia che può essere scritta da tutti, dimostrando una certa nescienza su cosa sia realmente Wikipedia. La parola *enciclopedia* viene dal greco *ἐγκύκλιος παιδεία*, istruzione circolare, insieme di dottrine che formano un'educazione completa. In questo senso Wikipedia ha lo stesso obiettivo che le biblioteche (soprattutto quelle pubbliche) si sono sempre poste: il servizio alla comunità, la circolarità della conoscenza. Nel caso di Wikipedia, una conoscenza a portata di tutti e perfezionabile da tutti. Circolante, quindi democratica.

L'enciclopedia libera si basa sulle competenze che ciascun utente può mettere a disposizione e sul confronto che ne nasce per arrivare a una versione condivisa. Può sembrare una sfida, ma questo è l'atteggiamento che i bibliotecari dovrebbero tenere: non si contribuisce alla comunità semplicemente aggiungendo link ai propri fondi o a pagine dedicate alle istituzioni di appartenenza, occorre comprenderne la filosofia e adeguarsi alle regole. Il case study di Biblioteca Digitale BEIC dimostra quanto un approccio collaborativo, aperto al confronto e all'arricchimento, possa allargare le prospettive di tutta la comunità, in armonia - nel senso più olistico del termine - con la quinta legge di Ranganathan, laddove la biblioteca cresce insieme ai propri tempi, alla tecnologia, alle esigenze dei propri lettori.

Bibliografia

Catalani L.-Feliciati P., *Wikipedia, le biblioteche e gli archivi / Wikipedia, Libraries and Archives*, «JLIS.it», 9, 3 (September 2018), pp. I-III. DOI: 10.4403/jlis.it-12510: https://www.jlis.it/issue/view/789

Cauzzi C. [et al.], Conoscersi per riconoscersi: la partecipazione come specchio del bibliotecario, «AIB studi», 56, 2 (maggio/agosto 2016), pp. 219-239. DOI 10.2426/aibstudi-11462

Consonni C.-Leva F., Progetto GLAM/BEIC, «Biblioteche Oggi», 33 (marzo 2015), pp. 47-50, http://www.bibliotecheoggi.it/rivista/article/view/24/265

De Francesca V.-Viazzi F., Come gestire una collezione di libri digitalizzati, Editrice Bibliografica, 2019 (Library Toolbox)

Leva F.-Chemello M., The effectiveness of a Wikimedian in permanent residence: the BEIC case study, «JLIS.it» 9, 3 (September 2018), pp. 141-147. DOI:10.4403/jlis.it-12481

Giaccai S., Come diventare bibliotecari wikipediani, Editrice Bibliografica, 2015 (Library Toolbox)

Zanni A., Le biblioteche e la filiera dell'open, «JLIS.it», 9, 3 (September 2018), pp. 75-91, DOI:10.4403/jlis.it-12486

Making a Digital Edition: The Petr*archive* Project

Isabella Magni
Rutgers University
isabella.magni@rutgers.edu

Abstract

English. This short essay will discuss current issues and future potentials of editing texts in the digital domain, while presenting a concrete case study: the Petr*archive* (ed. Storey, Magni and Walsh), an open access "rich-text" digital edition of Francesco Petrarca's songbook *Rerum vulgarium fragmenta*. The website proposes a new digital way of visualizing, studying and teaching Petrarch's work by offering different levels of visualization of the texts (facsimile high-quality images of all the *chartae* of the partial holograph Vaticano Latino 3195, its complete diplomatic transcriptions and edited forms and a nine-section commentary including a new English translation), as well as multiple indices and tools to access the diverse strata of the work's composition. Particular attention in this poster presentation will be given to new features currently in development, among which a new visual index that will allow users to navigate the material composition of Petrarch's manuscript and to analyze and visualize its fasciculation and the history of its composition.

Italiano. In questo breve saggio si discutono questioni e potenzialità dei metodi di edizione digitale, presentando nel contempo un caso specifico: il progetto Petr*archive*, una edizione digitale *open access* del canzoniere di Francesco Petrarca *Rerum vulgarium fragmenta*. Il sito si propone come un nuovo modo di visualizzare, studiare e insegnare l'opera petrarchesca offrendo diversi livelli di visualizzazione dei testi: immagini a colori ad alta risoluzione dell'olografo parziale (manoscritto Vaticano Latino 3195), nuove complete trascrizioni, sia in forma diplomatica che normalizzata, e un commentario in nove parti, che include una nuova traduzione in inglese. Grazie a un attento lavoro di encoding, Petr*archive* propone anche numerosi indici e strumenti per accedere ai diversi strati di composizione del canzoniere. Particolare attenzione in questa presentazione (poster) verrà riservata ai nuovi strumenti che stiamo sviluppando, tra cui un nuovo *visual index* che permetterà agli utenti di visualizzare la composizione materiale del manoscritto petrarchesco nei suoi undici fascicoli (nella loro composizione visiva e materiale) e di visualizzare le varie fasi di composizione del testo.

The *Rerum vulgarium fragmenta* is an icon of the Italian and Western literary tradition. Unlike other canonical works, we still possess authorial drafts of several poems and a partial holograph – MS Vaticano Latino 3195 – transcribed over roughly a decade (1366ca.-1374) in part by Giovanni Malpaghini[1], a young copyist from Ravenna working under Petrarch's strict supervision (Dotti 1987). After completing the first four quaternions, part of the fifth, the seventh and part of the last quires, around 1368, for unknown reasons Malpaghini decided to leave Petrarch's employment and protection. After his departure, while transcribing the remaining poems, Petrarch began a long, difficult and often interrupted process of transcription and revision of the entire songbook. From its intended status as a fair copy, Vaticano Latino 3195 soon became a service copy in which the poet experimented his visual poetics as the basis for a potential but never realized final fair copy. The basis of any textual research on the *Fragmenta* is accepting that it is not a perfect work and that it was unbound and unfinished when Petrarch died in 1374. Copies of the *Fragmenta* often partial and unauthorized – at times even corrupted – were already circulating during the poet's lifetime. Petrarch himself often lamented in his letters that his youthful vernacular poems were disseminated without his consent among the "multitude"[2]. Centuries of textual transmission and cultural mediation have progressively altered

[1] In a 2015 publication ("Malpaghini copista di Petrarca?" in *Cultura neolatina* LXXV: 2015 205-16) Monica Berté proposed to separate the historical figure of young Giovanni Malpaghini from that of Petrarch's scribe.

[2] In a letter to Giovanni Boccaccio Petrarch writes: "those brief and scattered vernacular works of my youth are no longer mine, as I have said, but have become the multitude's, I shall see to it that they do not butcher my major ones" (Sen. V 2, transl. Bernardo). And

the way we visualize, read and ultimately interpret Petrarch's poems, and the way we reconstruct the history of the work, often more conjectural than factual[3]. The starting point of our work on the Petr*archive* – and that of material and digital philology – is therefore to go back to the material sources and to re-examine original documents in order to dig below the surface of a "modernized Petrarch" and to re-construct the forms and contexts in which the work was produced.

The Petr*archive* does not reproduce in OCR other editions of the *Fragmenta* but offers in XML-TEI and John Walsh's TEIBoilerplate, high-definition images of each *charta* of manuscript Vaticano Latino 3195, new diplomatic transcriptions and edited forms of the entire songbook, and its discoverable palimpsests. Through carefully structured text encoding, the site aims at re-constructing Petrarch's texts maintaining their most overlooked aspect: their visual and material forms[4]. The basic authorial principles that characterize Petrarch's 366 texts and his carefully and authorially constructed visual poetics in MS Vaticano Latino 3195 are:

1. 31-line per *charta* organized in two columns;
2. thematic and visual integrity of the *charta*, in which the poems are often not simply juxtaposed but carefully selected to form groupings of pomes deeply linked by meaning, thematic unity and contrast;
3. contrasting visual structures to distinguish the five different poetic genres of which the *Fragmenta* is composed: two-column horizontal reading strategy for sonnets (transcribed over 7 lines-two verses per line), madrigals, ballata and canzone, as opposed to the two-column vertical reading strategy of sestina;
4. use of space as organizational device, signaling, for example, the subdivision of the collection in two parts (cc.49-52v), or a new trajectory of the macro-text with the transcription on c.22v of the canzone *Mai non vo' più cantar com'io soleva* anticipated by blank space (eight transcriptional lines) at the bottom of c. 22r.

The Petr*archive*'s first task is therefore to re-visualize these basic authorial principles while maintaining a simple interface and ease of use. The encoding of *charta* 1v, which presents four sonnets, serves as an example of how the digital code translates textual and prosodic features together with visual aspects of the façade of the *charta*:

```
<pb n="charta 1 verso" facs="../images/vat-lat3195-f/vat-lat3195-f-001v.jpg" />
<lg xml :id="rvf005" type="sonnet" n="5">
<lg type="octave">
<lg type="dblvrs" corresp="#canvasline">
<l n="1"><hi rendition="#red #fs24pt">Q</hi><hi rendition="#small-caps">u</
hi>ando io <choice><orig>mouo</orig><reg>movo</reg></choice> i sospiri a chiamar
<choice><orig>uoi</orig><reg>voi</reg></choice><supplied>,</supplied></l>
<l n="2"><choice><orig>El</orig><reg>E 'l</reg></choice> nome che nel cor mi scrisse
<choice><orig>amore</orig><reg>Amore</reg></choice>&v2c ;</l>
</lg> […][5]
```

Petrarch's visual poetics is maintained in the digital code: every pair of verses is translated in the strip of encoding as a <lg> (line group) of two verses (type="dblvrs") corresponding to one canvas line (corresp="#canvasline"). The result of the transformation of the encoding onto the web page is a new

in a letter to Pandolfo Malatesta, responding to the request by the *signore* of Rimini to receive a copy of his letters, Petrarch writes: "Now they have all circulated among the multitude, and are being read more willingly than what I later wrote seriously for sounder minds. How could I deny you, as great a man and so kind to me and pressing for them with such eagerness, what the multitude has and mangles against my wishes?" (*Seniles* XIII, 11. Transl. Bernardo).

3 For a critique of the still widely accepted theory of the "forms" of the *canzoniere* see, among others, Dario Del Puppo and H. Wayne Storey's "Wilkins nella formazione del Rvf di Petrarca." (2003); Teodolonda Barolini's "Petrarch at the Crossroads of Hermeneutics and Philology. Editorial Lapses, Narrative Impositions, and Wilkins' Doctrine of the Nine Forms of the *Rerum Vulgarium Fragmenta*." (2007); and Carlo Punsoni's "Il metodo di lavoro di Wilkins e la tradizione manoscritta dei *Rerum vulgarium fragmenta*." (2009).

4 See among others H. Wayne Storey's *Transcription and Visual Poetics in the Early Italian Lyric* (1993) and Furio Brugnolo's *Libro d'autore e forma-canzoniere: Implicazioni grafico-visive nell'originale dei* Rerum vulgarium fragmenta (1991).

5 The tag <pb> indicates a page break including the facsimile image (facs="../images/vat-lat3195-f/vat-lat3195-f-001v.jpg") of charta 1v (n="charta 1 verso") and is followed by the markup of the first line group (<lg>): sonnet *Rvf* 5 (type="sonnet" n="5"). Every tag of the alphanumeric strip of encoding refers to one specific textual, prosodic or visual component of the manuscript. This fourteen-verses line group is then subdivided into two subsequent <lg>: octave (lg type="octave"), the first four verses organized over four canvas lines (lg type="dblvrs" correps="#canvasline"); and sestet (lg type="sestet"), the remaining six verses transcribed over two canvas lines (lg type="dblvrs" correps="#canvasline"). For more on the Petrarchive encoding see my essay *I codici paralleli dei Fragmenta* (2015).

representation of Petrarch's visual poetics and editorial principles for which he worked restlessly for over a decade:

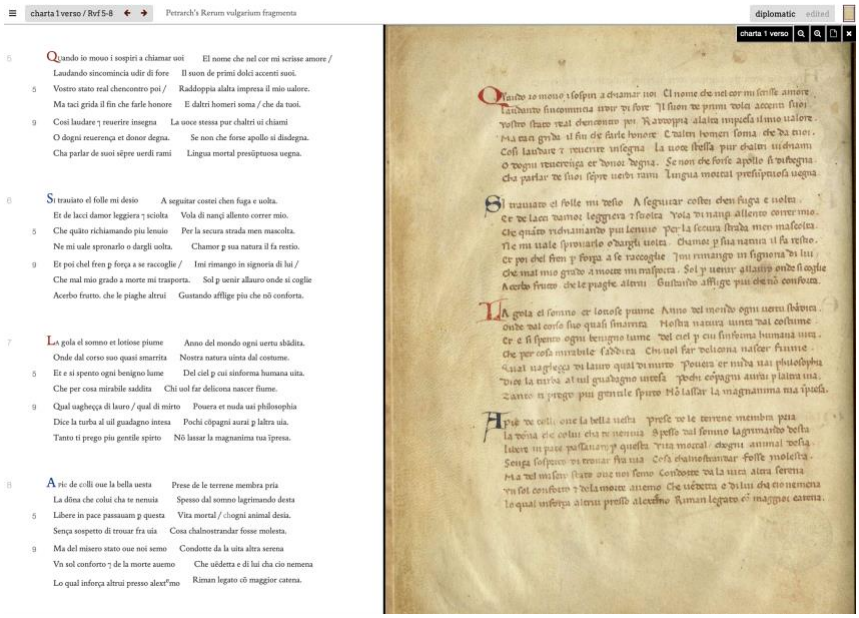


Figure 1. The Petrarchive, c.1v: diplomatic transcription and manuscript image, displayed side by side. URL: http://dcl.slis.indiana.edu/petrarchive/content/c001v.xml#c001v?facs=active

Other than being a research tool[6], text encoding also represents a close reading of the texts: it 'forces' encoders to 'break down' each poem and to 'label' its single components using specific tags. It can therefore also serve as an alternative and highly stimulating teaching tool: while being asked to encode Petrarch's texts, students necessarily need to re-think the very basic words, linguistic and prosodic structures of each poem to be able to digitally translate them into tags.

Advanced uses of text encoding also offer new ways of representing and analyzing erasures, renumbering, palimpsests, while maintaining a clean and simple digital interface. The most notable example is the palimpsest *Donna mi vene spesso ne la mente* on c.26r, erased and overwritten by *Or vedi amor che giovenetta donna* (*Rvf* 121) by the poet himself:

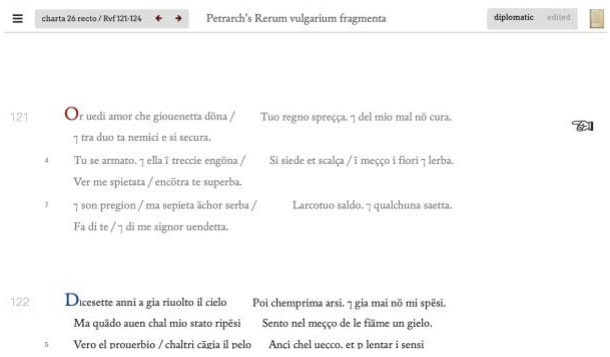


Figure 2 and 3. The Petr*archive*, c.26r: diplomatic transcription of *Rvf*.121 *Or uedi amor* (left); diplomatic transcription of palimpsest *Donna mi uene*. URL: http://petrarchive.org/content/c026r.xml

6 For more on digital editing and TEI encoding see, among others, the MLA White paper *Considering the scholarly edition in the digital age: a white paper of the modern language association's committee on scholarly editions* (2015 and 2016); Elena Pierazzo. *Digital scholarly editing: Theories, models and methods* (2015); Elena Pierazzo and Driscoll (eds.). *Digital Scholarly Editing: Theories and*. Cambridge (2016); Kenneth Price, and Ray Siemens. *Literary Studies in the Digital Age: An Evolving Anthology*. Modern Language Association, 2013; and Siemens, Ray and Susan Schreibman, *A Companion to Digital Literary Studies* (2008).

Through a combination of common Web design techniques and text encoding, by clicking on the *manicula*[7] in the right margin, the user can easily move from one version of *Rvf* 121 to the other as diplomatic and edited versions of both poems are available. To encode the simultaneous presence of erased and overwritten poems, we employ the following TEI elements: <del> (deletion): to contain the erased "Donna mi vene"; <add> (addition): to contain the added "Or vedi amor"; <subst> (substitution): to wrap the related <del> and <add> and assert that the <add> is *substituted* for the <del>; and <ab type="blockSubst"> (anonymous block)[8].

Petrarch's visual poetics is so authorially designed and so deeply part of the collection that a reader can recognize the genre and sometimes function of poems even before reading them. The SGV images created to digitally reconstruct the façade of the manuscript page are simple graphic representations of the *Fragmenta*'s material structures. On *charta* 1 verso and *charta* 3 verso, for example, users can easily identify the most frequent four-sonnets 'canvas' (c.1v) and distinguish the shift in directionality between the sonnet, horizontal, and the sestina, vertical (c.3v):

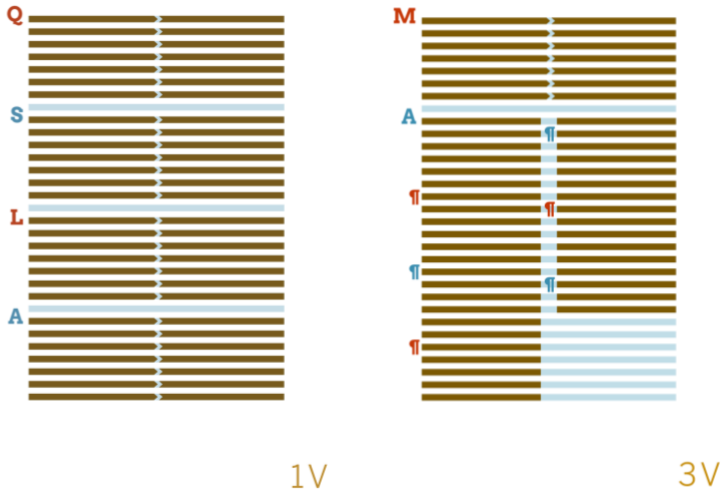


Figure 4 and 5. A visualization of c.1v and c.3v created for the Petr*archive* project.

The description of Petrarch's visual poetics embedded in the text encoding is therefore also represented in the Petr*archive* visual indexes[9]: the graphic image files are in fact XML files in the Scalable Vector Graphics format (SGV) which contain the information necessary for the Web browser to reproduce the image. The graphic information in the SVG files may also be derived from the codes embedded in the TEI/XML file, proving the visual and representational capabilities of the encoded document. From a brief look at the visual index of the entire songbook, also developed for the Petr*archive*, the nine pairings sonnet-sestina[10] present in the collection are immediately recognizable: once again, even before accessing the textual contents of the poems, the Medieval reader – and now the digital user – can collect a series of information regarding the poetic genres, their material and visual treatments:

[7] The *manicula* is not present in Vaticano Latino 3195: it is an interface element introduced by the Petrarchive, mimicking those Petrarch and other medieval and early modern readers used to draw attention to specific passages in manuscripts.
[8] For more information about the encoding developed to digitally reconstruct the palimpsest see Isabella Magni and John A. Walsh's *Digital Representations and the Pivotal Instability of* Donna mi vene spesso ne la mente *in the Study of the* Fragmenta (2016).
[9] In the examples in Figure and 5: arrows signaling the directionality of the text distribution (on c.3v the user can distinguish the shift in directionality between the sonnet, horizontal, and the sestina, vertical); paragraph markers in the sestina as internal visual indexicality indicating the vertical disposition of the text over the two columns; initials on the right of the text indentation signaling to the medieval reader, and now to the contemporary user, the beginning of new poems, marking the passage from fair- to work- and service copy (red and blue the rubricated fair-copy initials, in blank ink the remaining ones) and functioning as a textual index; blank space serving as additional punctuation device (visible in light blue color in the graphic SGV images).
[10] All of the *Fragmenta*'s sestine — except for the double sestina of part II (*Rvf* 332) — are always presented in contrast to a sonnet on the same charta (see Petr*archive Glossary* "Sestina". URL: http://dcl.slis.indiana.edu/petrarchive/content/glossary.xml#sestina).

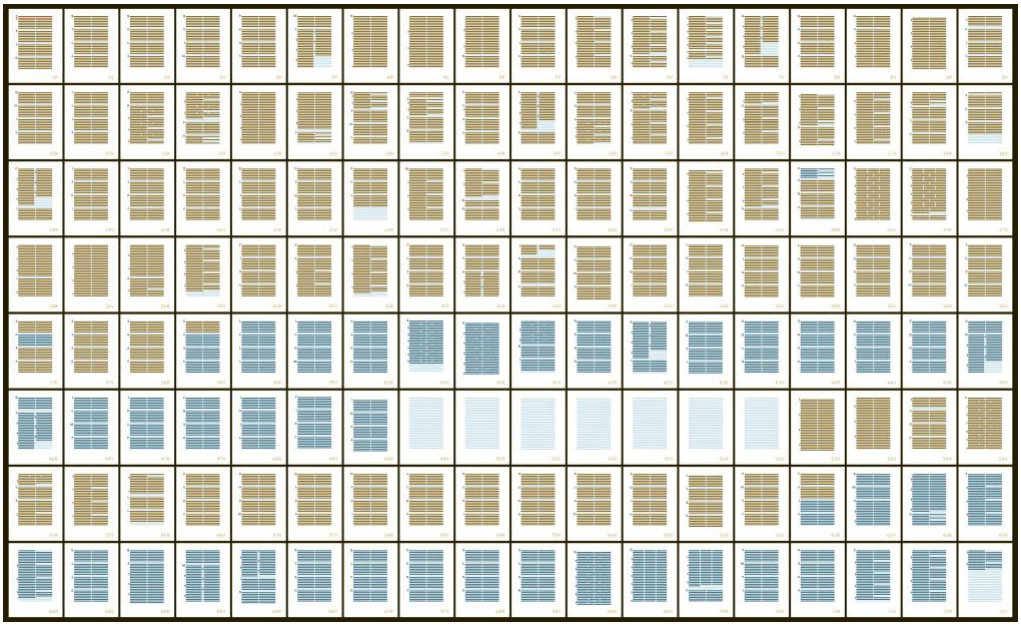


Figure 6. The Petr*archive*: visual index of MS Vaticano Latino 3195. URL: http://dcl.slis.indiana.edu/petrarchive/images/Petrarchive__Visual_Index_to_Vat__lat__3195.jpg

The representational values established by the Petr*archive* visual indices also offer a unique insight into the preparation of the manuscript, still in the form of loose gathering at the time of Petrarch's death: from the original project revealed by the rubricated *chartae* transcribed by Malpaghini (in brown) and set aside in 1368[11], to Petrarch's addenda in his own hand[12], and to the last service-copy transcriptions for the poet only[13] (both in dark blue). A newly developed visual index arranged by fascicles, also allows users to navigate into Petrarch's material construction of his fascicles, including the two final binions (cc. 63-66 and 67-70) that the poet inserted last into an already existing binion (cc.61-62, 71-72):

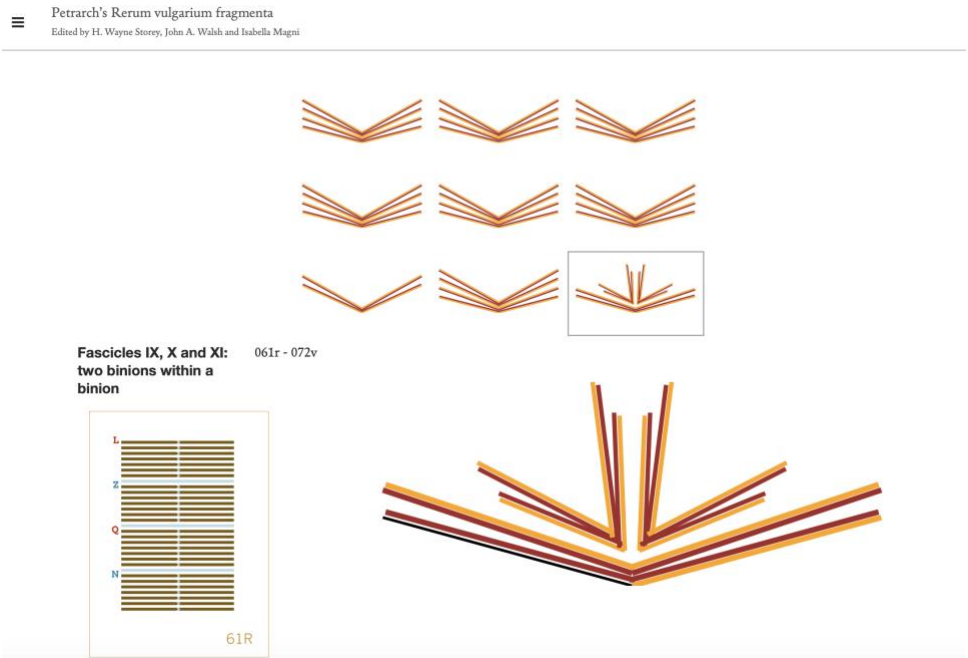


Figure 7. The Petr*archive*: visual index by fascicles. URL: http://dcl.slis.indiana.edu/petrarchive/visindex_fascicles.php

11 Five quaternions (cc.1r-40v) in Part I and two fascicles (cc.53r-60v) (cc.61r-62v and 71r-72v) in Part II.
12 Another quaternion (cc.41r-48) in Part I and four more *chartae* (c.59r-62v) in Part II.
13 This last section includes four *chartae* at the end of Part I (cc.49r-52v) and the last binion added towards the end (cc.69r-70v) with the transcription of canzone *Quel' antiquo mio dolce empio signore.*

Digital tools allow us to start from what we possess, the material evidence, and to dig below the surface to re-discover the original contexts in which the text was produced. Through carefully studied TEI encoding and the virtual representation on the web page of the different aspects of Petrarch's *Fragmenta* - its textual and graphic, temporal and spatial components – the Petr*archive* aims at re-building the structural and visual principles implemented by the poet himself at the level of single *charta*, fascicles and macro-structures and therefore to re-propose Petrarch's editorial choices, diminishing the distances between the experience of contemporary users and that of manuscript readers in the medieval context and providing innovative ways of teaching and conducting philological and literary research.

References

Aldo S. Bernardo. 1975. *Rerum Familiarium Libri.* State University of New York Press.

Aldo S. Bernardo. 1992. *Letters of Old Age.* Johns Hopkins University Press, 1992.

Carlo Pulsoni. 2009. "Il metodo di lavoro di Wilkins e la tradizione manoscritta dei *Rerum vulgarium fragmenta*." *Giornale italiano di Filologia*, 61: 257-269.

Elena Pierazzo. 2015. *Digital scholarly editing: Theories, models and methods.* Ashgate, Aldershot.

Elena Pierazzo and M. Driscoll. 2016. *Digital Scholarly Editing: Theories and Practices*. Cambridge: Open Book Publisher. https://www.openbookpublishers.com/htmlreader/978-1-78374-238-7/contents.xhtml

Furio Brugnolo. 1991. "Libro d'autore e forma-canzoniere: Implicazioni grafico-visive nell'originale dei Rerum vulgarium fragmenta." Lectura Petrarce 11: 259-90.

Gino Belloni, Furio Brugnolo, H. Wayne Storey, and Stefano Zamponi. 2004. *Rerum vulgarium fragmenta*. *Codice Vat. Lat. 3195*. *Commentario all'edizione in fac-simile*. Antenore, Roma.

H. Wayne Storey. 1993. *Transcription and Visual Poetics in the Early Italian Lyric*. Garland, New York.

H. Wayne Storey, Isabella Magni and John A. Walsh. 2013-. Petr*archive: An edition of Petrarch's songbook* Rerum vulgarium fragmenta. http://petrarchive.org

Isabella Magni. 2015. "I codici paralleli dei Fragmenta." *Medioevo letterario d'Italia* 12.

Isabella Magni and John A. Walsh. 2016. "Digital Representations and the Pivotal Instability of "Donna mi vene spesso ne la mente" in the Study of the Fragmenta" *Digital Philology* 5.2, Johns Hopkins University Press. https://doi.org/10.1353/dph.2016.0011

Julia Flanders, Ray Siemens and the MLA Committee on Scholarly Editions. 2019. *Considering the scholarly edition in the digital age: an engagement by the modern language association's committee on scholarly editions*. https://doi.org/10.1007/s42803-019-00026-4

Kenneth M. Price, and Ray Siemens. 2013. *Literary Studies in the Digital Age: An Evolving Anthology*. Modern Language Association. https://dlsanthology.mla.hcommons.org

Monica Berté. 2015. "Malpaghini copista di Petrarca?" *Cultura Neolatina* 75: 205-216.

Ray Siemens, and Susan Schreibman. 2008. *A Companion to Digital Literary Studies*. Blackwell, Oxford UK. http://www.digitalhumanities.org/companionDLS

TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/

Teodolinda Barolini. 2007. "Petrarch at the Crossroads of Hermeneutics and Philology. Editorial Lapses, Narrative Impositions, and Wilkins' Doctrine of the Nine Forms of the *Rerum Vulgarium Fragmenta*." *Petrarch and the Textual Origins of Interpretation*: 21-44. Eds. T. Barolini and H.W.Storey. Brill, Leiden–Boston.

Ugo Dotti. 1987. *Vita di Petrarca*. Laterza, Roma.

Extending the DSE: LOD Support and TEI/IIIF Integration in EVT

Paolo Monella
Venice Centre for Digital and Public Humanities, Università Ca' Foscari di Venezia
Università degli Studi di Palermo
paolo.monella@unipa.it[1]

Roberto Rosselli Del Turco
Dipartimento di Studi Umanistici
Università di Torino
LabCD, Università di Pisa
roberto.rosselidelturco@unito.it

Abstract

English. Current digital scholarly editions (DSEs) have the opportunity of evolving to dynamic objects interacting with other Internet-based resources thanks to open frameworks such as IIIF and LOD. This paper showcases and discusses two new functionalities of EVT (Edition Visualization Technology), version 2: one improving the management of named entities (f.i. personal names) through the use of LOD resources such as FOAF and DBpedia; the other, providing integration of the published text with digital images of the textual primary sources accessed from online repositories (e.g. e-codices or digital libraries such as the Vaticana or the Ambrosiana) via the IIIF protocol.

Italiano. Le edizioni critiche digitali oggi hanno l'opportunità di diventare sistemi dinamici che interagiscono con altre risorse su Internet grazie a *framework* aperti come IIIF e LOD. Questo saggio mostra e discute due nuove funzionalità della versione 2 di EVT (Edition Visualization Technology): il primo migliora la gestione delle *named entities* (ad es. i nomi di persona) attraverso l'uso di risorse LOD come FOAF e DBpedia; il secondo integra il testo pubblicato con le immagini digitali delle fonti testuali, recuperate da server online (ad es. e-codices o le biblioteche digitali Vaticana e Ambrosiana) tramite il protocollo IIIF.

1 Introduction

In the currently available DSEs there is a considerable lack of homogeneity since, besides a small number of (necessarily) common features, some are still rather traditional in the way of modelling the data, of conceiving the visualization of the edited text and in the kind of tools which they make available to the scholar, while others explore the enrichment of the edition's contents and the design of research tools in a highly innovative way. As W. Gibson (1990) said "The future is already here – it's just not very evenly distributed".

As a matter of fact, the first, pioneering DSEs were quite conservative in their approach to User Interface (UI) layout. This is understandable, since the first electronic editions were considered an equivalent of traditional editions in a different medium: the "printed page paradigm" (Sahle, 2016) inspired a mimetic design which would take into account very few advantages of the new publication framework besides its ubiquity and the apparently endless space it grants. The "remediation" (Bolter & Grusin, 2000) of the scholarly edition in the new media had not yet taken place.

When the methodology advances possible thanks to a Web-based digital edition started to be evident, a new research field was born: digital philology can be traced back to the desire to explore the potential of a truly *Digital* Scholarly Edition. This impulse has led to a lively but somewhat chaotic research activity, with an apparent paradox:

- only projects which can rely on generous resources may afford to explore new approaches to a DSE design, but usually they are focused on the specific task at hand, while the decision makers aren't interested in broader reflections on the general methodology;

[1] For the purposes of the Italian academy, R. Rosselli Del Turco is responsible for sections 1-4 and P. Monella for sections 5 and 6. The initial abstract was jointly written by the two authors, who also planned and revised the article together.

- interested scholars, on the other hand, may be hampered not only by the lack of resources to experiment, but also by the fact that the DH-related IT world is moving very fast, so fast that sometimes new technologies are introduced, enhanced, exploited and set aside within a few years.

This article aims at presenting some of the latest methods which can be applied to a DSE with the purpose of making it an even better (more flexible, modular, distributed, interconnected) research tool, and will also consider the software design and implementation issues that they imply.

2 No DSE is an island

Of the many limits artificially imposed to the DSE by the printed page paradigm, the first to fall was that of the book as a monolithic product, isolated from other books and unalterable, if not with rather high costs, after its publication. A DSE, in fact, can be changed both occasionally, f.i. to correct errors, and systematically, to add new texts, commentaries, bibliographic items, etc., so that a publication date by no means implies the end of the editing process, rather just the beginning of a new phase. Furthermore, a DSE is a dynamic, not a static object, a research tool which assists the scholar in data interpretation and analysis; and it is not a "closed box", but it can engage in dialogue and interaction with other Internet-based resources thanks to the global linking framework upon which the Web itself is built (Bodard & Garcés, 2009; van Zundert & Boot, 2011).

As a consequence, the greatest advantage of Web-based publication is not only that it makes scholarly content dissemination much easier and cheaper, but that the DSE can access other resources (and be accessed), and it can rely on external assets and services for specific functionalities. Examples include:

- taking advantage of semantic web technologies and Linked Open Data to enrich the edition content;
- text/image linking, pointing to digital collections of images of manuscripts maintained by external repositories (IIIF framework);
- modelling intertextual relationships through canonical text services (CTS and DTS protocols).

While this deep interconnecting and sharing of resources may look like a "quantitative" only advantage when compared to traditional scholarly editions (i.e., you can modify your edition when you want and integrate all the available content that you may see fit), there are important methodological consequences:

- the definitive rejection of the concept of an edition as an isolated and immutable entity;
- the acknowledgement of the fact that the use of external materials to integrate a DSE is highly desirable as it results in an enrichment of the DSE itself;
- an impulse to the collaboration between different edition projects, possibly adopting principles typical of the social edition (Siemens *et al*. 2012);
- as a consequence, an incentive to adopt a modular and distributed approach in the design of digital editions, making them more flexible;
- the possibility of virtually re-assemble dismembered manuscripts scattered across several preserving institutions (see f.i. the Fragmentarium project);
- a simplification of the problem of copyright management for digital reproductions of MS images by libraries, since images are functionally integrated in the DSE, but remain resident in their repository;
- the DSE itself may become a resource for other editions if the data on which it is based is published in such a way as to make it available to third parties.

This approach, however, poses a number of problems:

- greater technical complexity: so far only the editions defined as "haute couture" by E. Pierazzo (2019) can afford to adopt this approach because its digital implementation is certainly more complex than a simple reproduction of static images and texts in a Web-based edition;
- qualitative homogeneity of the different components of a DSE: if part of the content is entrusted to external resources, it is of critical importance that these resources meet the same academic standards as the original materials;
- long term sustainability: the interdependence between the "internal" and "external" components of a

DSE makes it possible to modify the existing connections and to add new materials as soon as they become available, but it also implies a strict and continued control on the actual availability and compatibility of these materials in the long run.

What is needed on the methodological level:

- open protocols, for data-sharing infrastructures, and open licenses, for resources to be shared;
- an open and ongoing scholarly discussion on the systematization of fundamental concepts to create a shared conceptual framework: we need shared terminologies and open ontologies as a necessary methodological condition for the creation of such shared resources;
- an integration of those methodological experiments in common editorial practice.

3 EVT 2: the linked DSE

EVT 2 is the second version, currently based on the AngularJS framework, of EVT (Edition Visualization Technology), a software tool for publishing editions based on the TEI/XML format, which has been developed in such a way as to go beyond the printed page paradigm – especially with regard to the User Interface design (see Di Pietro & Rosselli Del Turco 2018). The way in which it manages data, however, makes it mostly suitable for self-contained editions, created on the basis of local resources. Furthermore, at the present moment EVT is based on the client-only architecture, which presents many advantages (it is easy to install, little to no maintenance is required, it has indefinite durability), but also quite a few limits concerning important functionality (such as server functionality, f.i. for textual searches or to serve images).

For these reasons, it is an important goal of the development team to add support for protocols such as LOD for semantic Web resources, IIIF for images and CTS/DTS for intertextual relationships, so that scholars can count on a *prêt-à-porter* tool for their work. In the long term this goal may be achieved by adding RESTFul services, to add server functionality without encumbering the software too much, but it is already possible for projects based on EVT 2 to include LOD and IIIF resources. This will also have the beneficial effect of allowing more widespread knowledge of these protocols.

4 EVT and LOD

Although LOD resources are extremely diversified in terms of type of content, semantic-based operation, etc., it is undeniable that this is a very promising area which has seen constant growth in recent years.

EVT is already able to access resources on the Web thanks to URIs specified in the TEI markup, in fact some of the existing EVT-based editions (e.g. the Codice Pelavicino Digitale) use resources such as external repositories to provide additional information about the named entities identified in the text. The management of named entities can be significantly improved through the use of LOD resources such as FOAF and DBpedia.

In the Codice Pelavicino Digitale, person names are already linked to the *Dizionario Biografico degli Italiani* when possible, for instance:

```
<!--Code snippet 1-->
<note>Guido da Velate, arcivescovo di Milano dal 1045 al 1068, morto nel 1071 (si
veda <ref target="http://www.treccani.it/enciclopedia/guido-da-velate_
%28Dizionario-Biografico%29" type="biblio">Guido da Velate nel Dizionario
Biografico degli Italiani</ref>).</note>
```

This information can be supplemented or replaced by linking to a DBpedia entry:

```
<!--Code snippet 2-->
<ref target="http://dbpedia.org/page/Guido_da_Velate">Guido da Velate</ref>
```

so that it is possible to take advantage of the wealth of connections and of the sophisticated ontologies that LOD resources allow. Furthermore, as hinted above the DSE itself can make (part of) its material available as LOD, so that other editions can build upon it. The concept of "distributed edition", therefore, is coming closer to reality, in fact this is the goal of a new research project aiming at disseminating a DSE on sustainable and public resources such as Zenodo and GitHub (see O'Donnell *et al.*, 2018; note that EVT already runs on GitHub).

5 EVT and IIIF

5.1 The rise of IIIF

One notable example of an open protocol for online resource integration is IIIF – International Image Interoperability Framework. IIIF is rapidly emerging as a technology to exchange and integrate image-based resources in Web-based systems. One interesting use-case is a DSE in which portions of a TEI-encoded text based on a primary source such as a manuscript or a printed book are linked to the images of that source, stored and described via the IIIF framework. For instance, the transcription of a page of a manuscript can be linked to its facsimile, and the transcription of a line can be linked to the region of that facsimile corresponding to the line.

5.2 IIIF Image and Presentation APIs

The IIIF protocol defines different APIs, two of which will be briefly discussed here:

1. The *IIIF Image API* "specifies a web service that returns an image in response to a standard HTTP or HTTPS request". Simply put, this API returns *one* image or a portion of it. The description on the image are stored in a JSON file named `info.json`.
2. The *IIIF Presentation API* instead "describes how the structure and layout of a complex image-based object can be made available in a standard manner". Such a complex object can be a *collection* of digital images of a manuscript, accompanied by the relevant metadata, stored in a JSON file named `manifest.json`.

5.3 TEI and IIIF: a marriage made in heaven?

The TEI approach has always been text-centric, and only more recently the TEI editors have included a document-based approach in which the digital images of a textual source have equal dignity as its textual representation, via the <facsimile> / <surface> / <zone> encoding approach. On the other hand, IIIF is overtly and intentionally image-based. The TEI/IIIF integration thus looks very promising and productive for DSEs aiming to combine textual representation and digital images. However, at this point this is very much an open field of experimentation.

5.4 Two directions for a TEI/IIIF integration

Two approaches are theoretically possible for this integration:

1. Linking from IIIF to TEI

- *How*: according to the IIIF Presentation API, the node of a IIIF manifest identifying a specific (portion of an) image can point to an *annotationList* (a separate JSON file) including an annotation pointing to an external TEI XML file with the relevant textual representation (transcription).
- *Why this might not be a good idea*: the institution curating the IIIF collection (for textual sources, most probably a library or an archive) should create, curate and update the annotations linking to the TEI XML files. If those files belong to external DSEs incorporating the IIIF images, the DSE URIs might change and require constant update from the library's side – which is clearly not sustainable. On the other side, the library could create those TEI XML files itself to store and expose transcriptions of its own manuscripts, but in the current division of labour, libraries focus on digital imaging and metadata rather than on full transcriptions and textual criticism.

2. Linking from TEI to IIIF

- *How*: within a TEI XML file, f.i. in a <pb/> (page beginning) or <lb/> (line beginning) element, a @facs attribute points to a IIIF URI, either directly or indirectly.
    - *2A - Directly*: @facs takes the relevant IIIF URI as value (identifying, f.i., a whole manuscript page or a rectangle of that page including a line);
    - *2B - Indirectly*: @facs points to a <surface> or <zone> element within the <facsimile> section of the TEI file, and the <surface> / <zone> element points to the relevant IIIF URI.

The indirect strategy adds a layer of complexity, but also increases flexibility.

- *Why this might be a good idea*: this approach keeps resources separated (TEI XML files for text-centered digital philology and a IIIF infrastructure for digital image collections), with modularity and interoperability in mind. Many DSEs can link to the same IIIF image collections. Shortly said, philologists work on text with TEI, librarians work on document digitization with IIIF.

5.5 IIIF implementation in EVT

EVT 2 now features IIIF integration thanks to OpenSeadragon, its embedded image viewer. It implements the second approach described in the previous paragraph ("Linking from TEI to IIIF") and uses the IIIF Image API. Digital philologists can thus integrate external images of a textual source, hosted by a third-party IIIF image server, in the DSE, with an arbitrary level of alignment granularity. IIIF-compliant servers include e-codices, the Veneranda Biblioteca Ambrosiana in Milan (Cusimano, 2019) or the Biblioteca Apostolica Vaticana in Rome.

More precisely, EVT currently implements encoding strategy 2A described in paragraph 5.4 above:

1. The TEI XML source code has an element pointing to an image exposed by a IIIF server (typically a facsimile of a page) or to a portion of that image (typically a rectangle including a line or an other textual division). In the following code sample, we are pointing to a *whole image* (representing a manuscript page) from the IIIF server *e-codices* - *Virtual Manuscript Library of Switzerland*. The value of @facs is a URI following the IIIF Image API:

```
<!--Code snippet 3-->
<pb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
         full/full/0/default/jpg"/>
```

2. The image viewer integrated in EVT, OpenSeadragon, dereferences the URI from the @facs attribute, fetches the image from the external IIIF server and shows the whole image of the manuscript page alongside its TEI-based transcription.

Please note that the version of OpenSeadragon currently embedded in EVT (2.4.1) also allows to align a <pb/> element with a *specific portion of an image*, defined as a rectangle as per the IIIF Image API. Thus code snippet 3 above can be edited to:

```
<!--Code snippet 4-->
<pb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
         1800,600,3000,5000/full/0/default/jpg"/>
```

to crop out the manuscript page margins. Coordinates "1800,600" (in pixels) define the top left corner of the rectangle, "3000" defines the rectangle's base, "5000" its height. The encoding strategies described so far fulfill the common need of editors to pair <pb/> elements with a manuscript or book page (as well with the surface of an inscription, a tablet or any other support) and to display it aside the transcription.

The following code sample, instead, exemplifies the TEI XML encoding strategy currently supported by EVT to link elements such as <lb/>, <p> or to smaller portions of the surface image:

```
<!--Code snippet 5-->
<surface>
  <zone lrx="1052" lry="211" rend="visible" rendition="Line" ulx="261" uly="156"
           xml:id="zone-line-2-1"/>
</surface>
[...]
  <pb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
           full/full/0/default/jpg"/>
  <lb facs="#zone-line-2-1"/>
```

In code snippet 5 (which incorporates snippet 3 above), the <pb/> element references the whole manuscript page via a IIIF URI; <lb/> points to a <zone> element in the <facsimile> section that defines a rectangle within the IIIF image through the internal TEI XML encoding strategy: @lrx and @lry define the coordinates of the lower right corner of the rectangle, @ulx and @uly those of the and upper left corner (see attribute class att.coordinated in the P5 TEI *Guidelines*). Please note that this is different than strategy 2B from paragraph 5.4 because it still is the <pb/> element, not <zone>, that is retrieving the IIIF image.

5.6 Future development

Support for encoding strategy 2B (<pb/> or <lb/>'s attribute @facs points to <surface> or <zone>, and the latter points to an image in an IIIF server) is not yet available in EVT, but the development team aims at including it in future releases. These are examples of TEI XML code that will be managed by EVT:

```
<!--Code snippet 6-->
<facsimile>
  <surface xml:id="image-p2">
  <graphic url="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-
0730_002.jp2/
               full/full/0/default/jpg"/>
  <!--'zone' elements may be included here-->
  </surface>
</facsimile>
[...]
<pb facs="#image-p2"/>
```

Or, with a more compact encoding:

```
<!--Code snippet 7-->
<facsimile>
  <surface xml:id="image-p2" facs="https://www.e-codices.unifr.ch/loris/csg/
                  csg-0730/csg-0730_002.jp2/full/full/0/default/jpg">
  <!--'zone' elements may be included here-->
  </surface>
</facsimile>
[...]
<pb facs="#image-p2"/>
```

Another feature that may be implemented in the future, should any project collaborating with the EVT team express this need, is the definition of a rectangle within an image (e.g. for a manuscript line encoded with <lb/>) not through the TEI XML strategy (attributes @lrx, @lry, @ulx and @uly in <zone>), but directly through the IIIF image API, such as in the following code:

```
<!--Code snippet 8-->
<lb facs="https://www.e-codices.unifr.ch/loris/csg/csg-0730/csg-0730_002.jp2/
         1900,930,2580,140/full/0/default/jpg">
```

Finally, an experimental version of EVT based on the Angular 8 framework, not yet available for download, already supports the IIIF Presentation API and loads the full `manifest.json` document description including metadata on all manuscript images, while the image API currently supported by EVT provides access to one manuscript page image at a time. To load a whole `manifest.json` file, it will be enough to specify its URL in the EVT `config.json` configuration file, e.g.

```
<!--Code snippet 9-->
{ "title": "My Digital Edition",
  "manifestURL": "https://www.e-codices.unifr.ch/metadata/iiif/csg-
0730/manifest.json" }
```

Besides allowing a quick publication of all pages of a manuscript, this is a first step towards exploiting the full potential and flexibility of the IIIF framework.

6. Conclusions

The EVT development team is committed to supporting the current trend towards the distributed DSE, integrating resources such as entity or relationship definitions (LOD) and images (IIIF), as well as text fragments (CTS/DTS) in future versions. Internal and external objects alike can be better modelled by the new modular and object-oriented implementation adopted in EVT 2. Further development aims at supporting alternative encoding strategies for linking to external resources while keeping the whole edition (XML, other internal and external data, software) sustainable, durable and compliant with the FAIR principles, according to which "all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people" (Wilkinson *et al.*, 2016).

References

Gabriel Bodard and Juan Garcés. 2009. Open Source Critical Editions: A Rationale. In M. Deegan & K. Sutherland (eds.), *Text Editing, Print, and the Digital World* 83–98. Ashgate Publishing, Aldershot.

Jay D. Bolter and Richard A. Grusin. 2000. *Remediation: understanding new media*. MIT Press, Cambridge (Mass.) and London.

Fabio Cusimano. 2019. Biblioteche di conservazione & Data Curation: dal Custos catalogi al Digital Librarian. Il caso della Veneranda Biblioteca Ambrosiana. *JLIS* 10,1:125-139. DOI: http://dx.doi.org/10.4403/jlis.it-12513

Chiara Di Pietro and Roberto Rosselli Del Turco. 2018. Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. In *Digital Scholarly Editions as Interfaces*, edited by Roman Bleier, Martina Bürgemeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider. *Schriften Des Instituts Für Dokumentologie Und Editorik* 12: 133–63. BoD, Norderstedt. https://kups.ub.uni-koeln.de/9085/

William Gibson. 1990. *Cyberpunk* (Documentary). Directed by Marianne Trench, produced by Peter von Brandenburg, An Intercon Production. [Excerpt occurs in ~~Part 3 of~~ 5 parts; Timecode 12:20 of 14:59] (Video available in 5 parts on Youtube: https://www.youtube.com/watch?v=xxTuEGEl9EQ).

Lauren Magnuson. 2016. Store and display high resolution images with the International Image Interoperability Framework (IIIF). *ACRL TechConnect Blog*. URL: https://acrl.ala.org/techconnect/post/store-and-display-high-resolution-images-with-the-international-image-interoperability-framework-iiif/

Daniel Paul O'Donnell, Gurpreet Singh, Dot Porter, Roberto Rosselli Del Turco, Marco Callieri, Matteo Dellepiane, and Roberto Scopigno. 2018. Publishing (and Forgetting) the Small or Medium-sized Scholarly Edition or Cultural Heritage Collection as Linked Open Data: Using Zenodo and Github to Publish the Visionary Cross Project (Abstract).
URL: https://zenodo.org/record/3338482#.XeDjqdF7mV4. DOI: http://doi.org/10.5281/zenodo.3338482

Elena Pierazzo. 2019. What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter. *International Journal of Digital Humanities*. https://doi.org/10.1007/s42803-019-00019-3

Roberto Rosselli Del Turco, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti and Jacopo Pugliese. 2015. Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions. *Journal of the Text Encoding Initiative*. Issue 8. URL: http://jtei.revues.org/1077; DOI: https://doi.org/10.4000/jtei.1077

Roberto Rosselli Del Turco (ed.). 2017. *The Digital Vercelli Book. A facsimile edition of Vercelli, Biblioteca Capitolare, CXVII*. Transcription and encoding by Roberto Rosselli Del Turco, Raffaele Cioffi, Federica Goria. EVT software created by Chiara Di Pietro, Julia Kenny, Raffaele Masotti, Roberto Rosselli Del Turco. *Collane@unito.it*. URL: https://www.collane.unito.it/oa/items/show/11. ISBN 9788875901073.

Roberto Rosselli Del Turco. 2019. Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). *Textual Cultures* 12.2: 91–111. URL: https://scholarworks.iu.edu/journals/index.php/textual/article/view/27690; DOI: 10.14434/textual.v12i2.27690.

Patrick Sahle. 2016. What Is a Scholarly Digital Edition? In *Digital Scholarly Editing: Theories and Practices*, edited by Elena Pierazzo and Matthew J. Driscoll. Open Book Publishers https://www.openbookpublishers.com/product/483/digital-scholarly-editing--theories-and-practices

Alberto Salarelli. 2017. International Image Interoperability Framework (IIIF): A panoramic view. *Italian Journal of Library, Archives and Information Science = Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione*, 8.

Enrica Salvatori, Edilio Riccardini, Laura Balletto, et al. (eds.). 2014. Codice Pelavicino. Edizione digitale. URL: http://pelavicino.labcd.unipi.it. ISBN 978-88-902289-0-2; DOI: https://10.13131/978-88-902289-0-2.

Ray Siemens, Meagan Timney, Cara Leitch, Corina Koolen, and Alex Garnett, with the ETCL, INKE, and PKP Research Groups. 2012. Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media. *Literary and Linguistic Computing*, 27(4): 445–461. https://doi.org/10.1093/llc/fqs013

TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 3.6.0. Last updated on 16th July 2019. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Mark D. Wilkinson *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3. https://www.nature.com/articles/sdata201618

Joris van Zundert, J. and Peter Boot. 2011. The digital edition 2.0 and the digital library: Services, not resources. *Digitale Edition Und Forschungsbibliothek (Bibliothek Und Wissenschaft)* 44: 141–152.

Mapping as a Contemporary Instrument for Orientation in Conferences

Chloe Ye Eun Moon
Columbia University
ym2761@columbia.edu

Dario Rodighiero
Massachusetts Institute of Technology
rodighie@mit.edu

Abstract

English. This article presents a case study analyzing submissions from the Digital Humanities 2019 conference by visualizing a network of authors situated according to their shared lexicon. This new form of summarizing a conference is an effective way to grasp the whole conference at once. The hope is that this method of visualization will not be employed merely as a retroactive way to reflect on past events, but rather as an instrument to prepare the visit and orientate the attendees during the conference.

Italiano. Questo articolo presenta una mappa lessicale creata per studiare la distanza linguistica tra gli autori che hanno partecipato alla conferenza Digital Humanities 2019. I testi degli articoli sono raggruppati per autore e analizzati secondo la loro frequenza, per poi essere tradotti in connessioni. Il risultato è una mappa topologica della conferenza, composta degli stessi autori situati in uno spazio di linguistico. Questa nuova forma di visualizzazione è un efficace strumento per riassumere una comunità scientifica in un'immagine. L'augurio è che metodi come questo non siano soltanto degli strumenti usati a posteriori, ma piuttosto messi a disposizione in anticipo per preparare la propria visita e permettere ai partecipanti di orientarsi durante una conferenza.

1 Introduction

Maps are formidable instruments for abstracting territory and travels. Centuries of cartographic mapping marked the evolution of world history, and today's technological innovation has opened a new paradigm of mapmaking (Dodge et al., 2011; Lévy, 2016). Maps are no longer static; instead, they can change dynamically given external inputs. For instance, users can zoom into the map to obtain detailed information, and they can choose to filter selected information. Due to the big data technology, maps are now able to represent larger and larger datasets (Kitchin, 2014). Moreover, maps now have a new level of abstraction not only for the territory but also for individuals; unlike before, visualizing social relationships became a common practice since Jacob Moreno's work in social relationships (1934).

This article presents a case study of human mapping, specifically of scholars and their scientific productions, focusing on the Digital Humanities Conference 2019 that took place in Utrecht, the Netherlands in July 2019. Such a case study develops the previous work of Rodighiero (Rodighiero, 2015, 2018; Rodighiero et al., 2018), demonstrating how a map can be a powerful instrument of reduction not only for territory but also for individuals. The cartography of DH2019[1] is intended to be a generic tool for mapping scientific communities, in the form of an open-source project[2]. The article will present the result of such cartography by discussing the following four sections: 1) *Documentality* as a way human activity is regulated through textual inscriptions 2) *Lexical Analysis* as the way documents can be automatically analyzed under human supervision 3) *Graphic Design* as a visual translation that makes a conference

[1]The cartography of DH2019 is accessible at https://rodighiero.github.io/DH2019/
[2]The open-source project is available at https://github.com/rodighiero/DH2019/

attended by a thousand authors wholly graspable 4) *Reading* as a form of user interaction through which the map reader acquires information from the visual media.

2 Documentality

A series of social rules govern humans behaviors (Kaplan, 2012), and scholars are no exception. Their activities follow specific rules when they attend a conference; submitting articles, waiting for reviews, attending the conference, and presenting their work are the steps they are required to complete in order to be a part of an academic field. Such behaviors are auto-regulated by the scientific community itself, which gives a structure to the research domain.

Articles play a significant role in this process as they are an extension of their authors. They convey different types of information, such as collaboration, affiliation, scientific interests, and the writing proficiency of the authors. Therefore, articles are valuable as they portray the authors' values and social relevance. We hypothesize that documents embody textual information, which can be used to measure the proximity between scholars most effectively. As individuals express themselves through their language, authors can be described through their writing. Terms are not a barrier, nor they are private. Everyone can choose and use preferred words and speech styles according to their taste, and the choice is profoundly affected by social and cultural environments, such as their education level and location. Nonetheless, there is complete freedom in the selection of language, and this freedom goes beyond any collaboration or citation. Just like how Pierre Bourdieu used personal interviews to classify individuals (Blasius and Schmitz, 2014; Romele and Rodighiero, 2019), the goal here is to map scholars using scientific articles.

3 Lexical Analysis

Natural Language Processing (NLP) is a branch of artificial intelligence that aims to process and analyze large amounts of text (Manning and Schuetze, 1999). Since humans' natural language has no structured rules, computers can understand, NLP is especially challenging; therefore, the techniques in NLP are significant as they derive meaningful insights from texts written in such a language. In order to profile the authors who attended the Digital Humanities 2019 conference, we performed a lexical analysis to map the distance from one author to another.

First, the XML data of DH2019 were cleaned for a more accurate analysis utilizing JavaScript programming language and the Cheerio library[3]. From each article are extracted authors, title, and text body. Since multiple authors can co-author a paper, the text body is grouped by authors, allowing multiple occurrences of the same publication if a paper is co-authored. Each text is then tokenized using a lexical analyzer provided by the Natural library[4] for NLP. Successively, tokens are singularized and filtered by a list of stopwords in various languages, including Brasilian, English, French, German, Italian, and Portuguese.

The arrays of tokens associated with authors are then computed via the Term Frequency - Inverse Document Frequency algorithm, also known as TF-IDF (Luhn, 1957; Sparck Jones, 1972). TF-IDF extracts the most relevant terms for each author by counting the frequency of each term with respect to the inverse frequency of the entire collection of words. The list of terms for each other is then shortened to the fifteen most relevant terms in order to simplify the visual computation. Table 1 shows a sample concerning the scholar Frédéric Kaplan.

[3]Cheerio is an open-source library available at https://github.com/cheeriojs/cheerio/
[4]Natural is an open-source library available on GitHub at https://github.com/NaturalNode/

Token	Value
Parcel	156.29240966203554
Amsterdam	147.92161777735498
Street	97.3716554233416
City	78.98538009996346
Cadastre	68.97376382900369
Rue	64.52753719852456
Neighbourhood	62.45323622544596
Urban	61.733372168535126
Century	52.240930125777
Bottin	51.62202975881964
Transcription	51.58659792981266
Geometrical	44.74034261035416
Extraction	43.895897928544414
Geographical	41.91500209834842
Geometry	39.821449253041536
Wine	38.71652231911473
Activity	37.66779857289015
Dialect	37.53334200758336
Cinema	34.18896461357028
Time	33.67020614879688

Table 1: An excerpt from the JSON file that describes the profile of Prof. Frédéric Kaplan. Among the metadata are his name, the number of articles, and a list of fifteen tokens weighted through the TF-IDF algorithm. As his research is mainly focused on the European Time Machine, which is focused on the computational analysis of ancient maps, the result can be considered adequate.

4 Graphic Design

Data analysis is followed by the creation of a network, in which each author forms a node, and the shared tokens are transformed into weighted edges. The resulting visual rendering does not recall a classic network visualization, such as Gephi's (Bastian et al., 2009), but rather a cartographic projection. It is a hybrid form that combines the characteristics of networks and maps.

Authors are placed on the map using the Simulation function[5] from the d3.js library (Bostock et al., 2011). Then, between each pair of authors, is displayed the most relevant token whose size corresponds to the TF-IDF value; a high TF-IDF value corresponds to a high degree of relevance in the whole collection. The elevation contours (Monmonier, 1991) displayed at the end of the simulation due to computation limits make the density of documents visible; as a result, an author who authored many articles is placed at a peak. The result is an elevation map that shows the most relevant tokens, such as languages, music, newspapers, and films. When zoomed in, the map can be enlarged to display the details, and the user will notice the tokens changing. The reason is that the tokens are selected according to the zoom level; when the map is utterly visible, the user can see the most relevant tokens with high-frequency values, and by zooming he will see the most generic ones. This choice makes the view more specific and less generic, while a zoom-in allows viewing the complete gradient of tokens (See Figures 1, 2, and 3).

[5]Simulation runs to place the nodes in a proper way, more information is available at https://github.com/d3/d3-force/

5 Reading

In cartography, the reader is the individual who interacts with the map (Dodge et al., 2011; Lévy, 2016). Readers interpret the cartographic visualization differently depending on their knowledge and culture. In front of the DH2019 visualization, the reader identifies a configuration of scholars who attended the conference with an article. When authors share the same text because they co-authored a paper, they appear to be close on the map, and the most relevant token between them appears (Figure 1). When there is a continuity of tokens, it indicates an area of particular interest (Figure 2). When the density of individuals is visible, there is a high chance that it represents a collective work (Figure 3).

The map offers a novel point of view on the conference, which is a different approach from the proceedings or the website. It makes all the authors and their work graspable at the same time, while also allowing the readers to navigate through individual authors.

When a reader starts to explore a specific part of the map by zooming in, the interaction becomes personal and unique. Therefore, readers play an active role while putting an interpretation on the map; their pathway influences what they see. Furthermore, if a reader who participated in the conference recognizes her identity in the map, the reading validates the reader's representation by evaluating the correctness of the map and the neighborhood where the reader is placed (Rodighiero and Cellard, 2019).

6 Conclusion

Language is not only a means to convey one's ideas, but also to express interest and background. If a conference forms a scientific community by attendance, the articles presented at the conference shape the specific language of such a community with a delicate balance of different voices (Von Glasersfeld, 1992).

Lexical analysis of terminologies is an effective means to study the community. Thanks to the current technology and visualization techniques, we are now able to create a dynamic, interactive map of the community, which is dense and rich in information. From this new form of data visualization, readers can interpret the lexical proximity of all the authors at a glance, both the distance and placement.

Now the question is, why don't we use the map as an instrument during the conference? It would undoubtedly be a much more contextually rich and visually intriguing way of understanding the conference, instead of merely using statistics to summarize the event.

Acknowledgement

We want to thank Kurt Fendt for his constant support and supervision, and the MIT Literature section for hosting us, in particular Shankar Raman, Diana Henderson, and Alicia Mackin. Thanks also to Jeffrey Schnapp and the laboratory members of Harvard MetaLab.

Acknowledgement also goes to Stephan Risi for developing the search function and Philippe Rivière for his priceless collaboration, which was fundamental for recent projects. A special mention to Daniele Guido, an inseparable friend and colleague whose design capacities are stimulus for doing better; the grapefruit color palette is his merit.

This article is part of the grant Early Postdoc.Mobility P2ELP1_181930 *Worldwide Map of Research* funded by the *Swiss National Science Foundation*.

Figure 1: The cartography of DH2019. At this level of zoom, the most specific terms suggest some entry points to explore. For instance, in the middle, the terms “news” and “newspaper” invite to zoom in these areas. The reader is also free to explore empty spaces or search a scholar by using the search box at the top right of the interface.

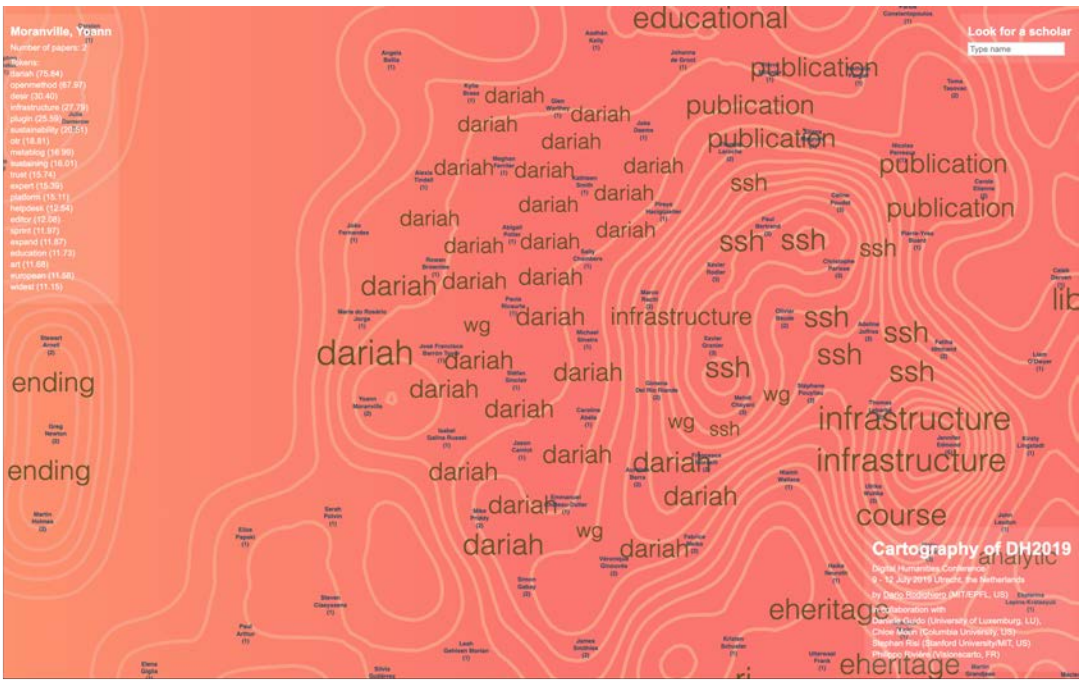


Figure 2: This image illustrates a specific area of the map in which the term 'dariah' is recurrent. The term reassembles the people working within the Dariah community. By zooming, the terms change according to the scale; the more the reader zooms in, the more generic the terms are.

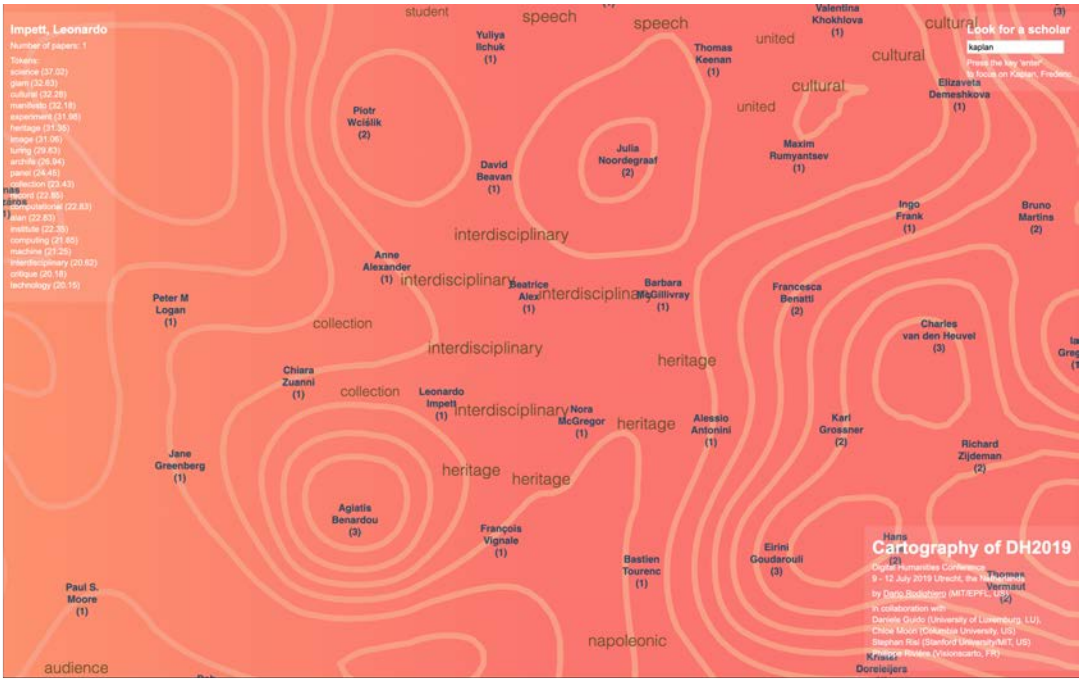


Figure 3: Co-authoring is an easily recognizable phenomenon, especially in areas with low density. At the center, it is visible a group of scholars that share the term 'interdisciplinarity.' Terms, like in this case, can be used to spot small communities within the conference.

References

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International ICWSM Conference* .

Jörg Blasius and Andreas Schmitz. 2014. Empirical Construction of Bourdieu's Social Space. In *Visualization and Verbalization of Data*, CRC Press, pages 205–222.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* 17(12):2301–2309. https://doi.org/10.1109/TVCG.2011.185

Martin Dodge, Rob Kitchin, and Chris Perkins, editors. 2011. *The Map Reader: Theories of Mapping Practice and Cartographic Representation*. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9780470979587

Frédéric Kaplan. 2012. How Books Will Become Machines. In *Lire demain : des manuscrits antiques à l'ère digitale*, PPUR, pages 27–44.

Rob Kitchin. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE Publications.

Jacques Lévy, editor. 2016. *A Cartographic Turn*. EPFL Press, Lausanne.

Hans Peter Luhn. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development* 1(4):309–317. https://doi.org/10.1147/rd.14.0309

Christopher D. Manning and Hinrich Schuetze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mark Monmonier. 1991. *How to Lie with Maps*. University of Chicago Press.

Jacob L Moreno. 1934. *Who Shall Survive?*. A New Approach to the Problem of Human Interrelations. Nervous and Mental Disease Publishing Co., Washington, DC.

Dario Rodighiero. 2015. Representing the Digital Humanities Community: Unveiling the Social Network Visualization of an International Conference. Parsons Journal of Information Mapping VII(2). https://doi.org/10.5281/zenodo.3464433

Dario Rodighiero. 2018. Printing Walkable Visualizations. In *Transimage Conference 2018*. University of Edinburgh, Edinburgh, pages 58–73. https://doi.org/10.6084/m9.figshare.6104693

Dario Rodighiero and Loup Cellard. 2019. Self-Recognition in Data Visualization. *EspacesTemps.net Electronic Journal of Humanities and Social Sciences* https://doi.org/10.26151/espacestemps.net-wztp-cc46

Dario Rodighiero, Frédéric Kaplan, and Boris Beaude. 2018. Mapping Affinities in Academic Organizations. *Frontiers in Research Metrics and Analytics* 3(4). https://doi.org/10.3389/frma.2018.00004

Alberto Romele and Dario Rodighiero. 2019. Digital Habitus, or Personalization without Personality (Forthcoming).

Karen Sparck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28(1):11–21. https://doi.org/10.1108/eb026526

Ernst Von Glasersfeld. 1992. Why I Consider Myself a Cybernetician. *Cybernetics and Human Knowing* 1(1):21–25

Argumentation Mapping for the History of Philosophical and Scientific Ideas: The *TheSu* Annotation Scheme and Its Application to Plutarch's *Aquane an ignis*

Daniele Morrone
Università di Bologna
daniele.morrone2@unibo.it

Abstract

English. This paper presents the *TheSu* XML annotation scheme, which is intended to be an indexing and mapping tool for intellectual historians. Its sheets contain "theses" extracted from written works, representing the stance of their authors or of the individuals quoted in the text, classified by themes and other peculiarities. These theses, linked between them in argumentative and expository "supports", compose a network identifiable with the "scientific discourse" that the work they are included in means to convey. Being it representative of an author's scientific or philosophical thought, it is always important for the historian researching on that author's ideas to give proper and articulate consideration to all its elements and their relations. *TheSu* is designed to aid in this operation, by providing the possibility of generating organized lists and maps of the "Argumentative-Expository Systems" of interest to the historian. In this presentation, examples are provided from an exhaustive case annotation of Plutarch's *Aquane an ignis utilior sit*. *TheSu* is also briefly compared to apparently similar annotation schemes in Argumentation Mining to better show its individual features and aims.

Italiano. Questo articolo presenta lo schema di annotazione XML *TheSu*, pensato come uno strumento di indicizzazione e mappatura per storici delle idee. Un foglio *TheSu* contiene "tesi" estratte da un testo scritto, rappresentanti il punto di vista del suo autore o degli individui da esso citati, classificate secondo temi e altre caratteristiche. Queste tesi, collegate tra loro all'interno di "supporti" argomentativi ed espositivi, compongono una rete identificabile con il "discorso scientifico" trasmesso dal testo in cui sono inserite. Poiché esso può essere rappresentativo del pensiero scientifico o filosofico di un autore, è sempre importante che gli storici che ne studiano le idee prestino la giusta attenzione all'intera articolazione di tale discorso, ai suoi elementi e alle loro relazioni. *TheSu* serve a semplificare quest'operazione, dando la possibilità di generare liste organizzate e mappe dei "Sistemi Argomentativo-Espositivi" d'interesse per gli storici. In questa presentazione sono mostrati esempi tratti da un'annotazione esaustiva dell'opera di Plutarco *Aquane an ignis utilior sit*. *TheSu* viene inoltre confrontato brevemente con altri schemi d'annotazione apparentemente simili nel campo dell'*Argumentation Mining*, per mostrare al meglio i suoi scopi e le sue caratteristiche individuali.

1 Introduction

The field of "computational history of philosophy" (Betti et al., 2019) is rather new but promising, as it can provide historians with powerful research tools to work with large amounts of data in an organized fashion, giving them the possibility of finding patterns, similarities and links. History of philosophy and History of science can be regarded as subfields of History of ideas – meant in the broadest possible sense – and although digital methods seem to have only recently been introduced in this latter (Betti and van den Berg, 2016)[1], History of science has been benefiting from them for a long time already, under the influence of Computational linguistics (Dibattista, 2009). By presenting the novel XML annotation scheme *TheSu*, this paper aims to contribute to the general trend of digitalizing the research methods in these fields, focusing on "ideas" in the sense of judgements about states of things and giving relevance to the way these judgements are presented and promoted by their authors.

Plutarch's short conference (D'Ippolito and Nuzzo, 2012, pp. 180–191) *Aquane an ignis utilior sit* (*Aq.*) — "Whether fire or water is more useful"— has been annotated according to the *TheSu* scheme to give some examples of this latter's possible applications and capabilities. The digital XML/TEI edition (TEI Consortium, 2019) of the original Greek text chosen as a base for the annotation has been downloaded from PerseusDL/canonical-greekLit (Cerrato et al., 2019), and corresponds to Bernardakis's critical edition of the work (1895, pp. 1–10).

[1] Betti and Van den Berg do not seem to consider the activity of the ILIESI (Istituto per il Lessico Intellettuale Europeo e Storia delle Idee) in Rome, which has long been working on History of ideas in the frame of Digital Humanities. See http://www.iliesi.cnr.it/.

2 *TheSu* and related work in Argumentation Mining

The aim of the *TheSu* (*Thesis-Support*) annotation scheme is to provide the possibility of easily navigating through enunciates (*Theses*) contained in written texts and all their linked explanations, justifications and refutations (*Supports*), each indexed as a node in an abstract network defined as "Argumentative-Expository System" (AE System), which is stored in a database. Focusing on argumentative relations of whatever rhetorical nature, *TheSu* can be likened to the various annotation schemes that are being proposed in the field of Argumentation Mining (Lippi and Torroni, 2016; Stede and Schneider, 2019), even if it doesn't share their common objective of digitally automatizing argument extraction from texts. *TheSu*, although similar to these approaches, is different from them for two main reasons:

(1) It builds its system on theses abstracted from the texts by human interpreters, which can then be linked to their possible textual supports (if there are any). Argumentation mining approaches influenced by Toulmin (2003 [1958]) and Walton (1998; Id. et al., 2008) tend to directly search the texts for premise-conclusion enunciative pairs to tag them under schemes such as Walton's "argumentation schemes" (see *e.g.* Lauscher et al., 2018; Mochales Palau and Moens, 2009; Rocha et al., 2016; Green, 2018a); approaches based on Rhetorical Structure Theory (RST) instead (see Mann and Thompson, 1987; Taboada and Mann, 2006, secs. 2.4, A.2) select their elements through objective textual markers (see the definitions of EDUs —Elementary Discourse Units— in *e.g.* Carlson et al., 2001; Marcu et al., 1999), and as a consequence segment the text into discrete —albeit interconnected— non-overlapping units (on the undesirable aspects of these approaches see Green, 2018b; Peldszus and Stede, 2013, pp. 15–19). In contrast, *TheSu* focuses first on the indexing of individual theses, *i.e.* treating every single declarative sentence as a "claim", and then on their connection with supportive spans of text: the latter can be contiguous to their targeted theses or very far away in the text, as well as in other works from the same author or from different authors too (as will become clearer below).

(2) While Argumentation Mining methods are generally concerned with *textual* cohesion and natural argumentation patterns, *TheSu* is interested in the coherence and justification of an author's ideas in her *thought*, inasmuch as it is exhibited in her textual production. This also differentiates *TheSu* from annotation

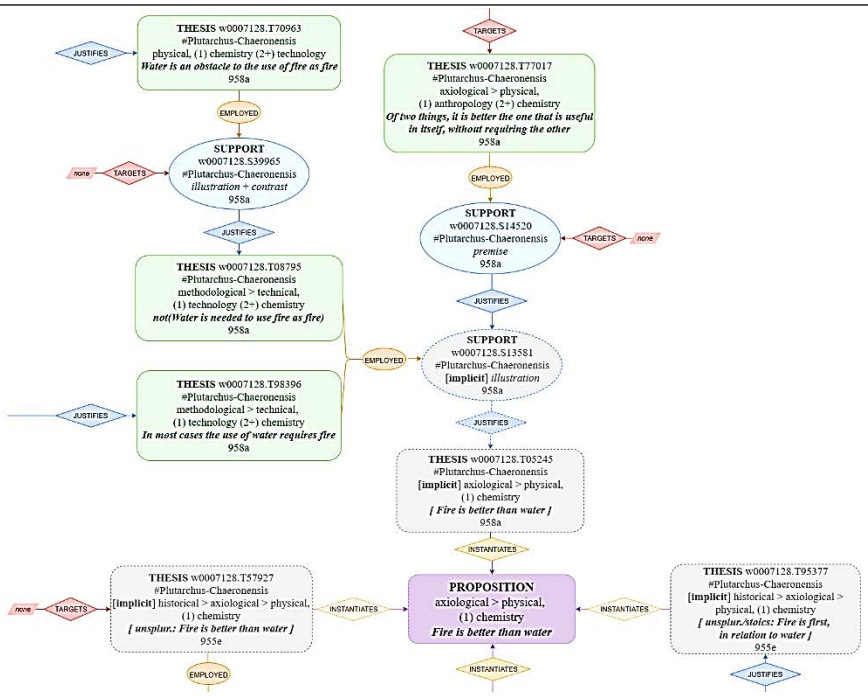


Figure 1. Fragment of a concept visualization of a *TheSu* map: Argumentative-Expository contexts linked to the theses in *Aquane an ignis utilior sit* instantiating the proposition 'Fire is better than water'. Every THESIS and SUPPORT that doesn't receive justifications or explanations is highlighted by "*none* –targets→": it's desirable to be able to notice them at a glance because it can be proof that their speaker considered them clear and non-controversial enough not to spend more (supportive) words on their presentation, thus being the ideological 'building-blocks' of the whole argumentative discourse.

schemes in Argumentation Mining that seem to be more independent from Walton's and RST's influence (*e.g.* Peldszus and Stede, 2013). An intellectual historian, while researching on an author's thought, usually tries to reach a comprehensive view of it in order to identify trends and elements of cohesion, incompatibility, and evolution. When the historian extends the scope of her research to include texts from different authors, her

aim is usually to be able to discover traces of historical influences or innovations based on independent reasoning. Sometimes she tries to elucidate the author's texts by putting them in relation to others pertaining to the same culture or current of thought: when certain ideas are presented synthetically and without explanation, she can always look at works from different authors —culturally and philosophically close to the first— to find their plausible sense and justifications (on current research practices in History of ideas cf. *e.g.* van den Berg et al., 2014, sec. 3). *TheSu* is intended as a tool to help the historian reach these aims, by providing databases for generating maps of the networks of ideas conveyed by texts, and arrange and filter them according to her interests (see Figure 1).

TheSu is thus distinguished from the other annotation schemes in a way that can be summarized as follows: although it always starts from a text containing natural argumentation, it only uses it as a proof for the existence of a *scientific discourse* that the text's author intends to convey. The "discourse" is composed of both explicitly stated enunciates and their implicit assumptions and alluded consequences, as well as all the explicit and implicit argumentative links between them. These are only "scientific" in the sense that they are to be 'taken seriously' by the interpreter, who must always start by assuming the hypothesis that the author has legitimate reasons to believe in and present all of them: to test her hypothesis, the interpreter must thus do her best to find in the text all the supports that might qualify the claims as *well founded* and *adopted critically* by the author, and so "scientifically" legit (in the context of their existence). In so doing, the interpreter cannot but be guided by a strong principle of charity[2], and in this way detach the scientific discourse from the text up above a certain degree of 'charitable' arbitrariness. The structure of the scientific discourse, then, can not always correspond to the structure of the text, and the latter is only used as grounding for the reconstruction of the former.

TheSu annotations, in addition, can serve the purpose of gathering organized data as a basis for logical and epistemological evaluations of an author's style of reasoning. To make these further analyses possible, the interpreter must be as non-judgemental as possible in the annotation phase: weird and weak as they may seem, every extra-logical "argumentation" practice deserves the same space as the actual "demonstrations"— adopting Perleman and Olbrecht-Tyteca's distinction (2013)— in the network of ideas. This also distinguishes *TheSu* from more 'normative', logically rigid, approaches in Argumentation Mining (*e.g.* Green, 2018a), and from the CRM*inf* Argumentation Model, an extension to the CIDOC CRM that complements CRM*sci*, a model for the structuring of metadata about contents and practices of current empirical sciences (Stead et al., 2019). In CRM*inf*, the epistemological evaluation of the arguments is embedded in the annotation itself (*e.g.* its class "I3 Inference Logic" can only include «anything that is scientifically or academically acceptable as a method for drawing conclusions», *ib.* p. 11), and much of the discourses' rhetorical contexts is thus ignored.

Plutarch's *Aq.* has been chosen as a case study because of its short length and its elaborate, though very clear, argumentative structure. It is a rhetorical exercise where both the superiority of water and the superiority of fire are argued for in persuasive speeches that are symmetrical in extension as well as in cogency, and wherein no final solution is provided to the controversy. It contains way more "argumentation" than "demonstration", and its interesting rhetorical features have already been analysed by Milazzo (1991), although with a different approach. In this paper, its theses will only be quoted by their annotated paraphrases in English, which is the standard language for the *TheSu* sheets: considering that all the theses have been extracted from the original Greek text, in this case every paraphrasis is also a translation, original to this annotation and sometimes diverging from the previous ones —including Helmbold's in Cherniss and Helmbold (1957)— to improve on clarity and faithfulness. The original (pre-annotated) text will be quoted in translation as well.

3 Encoding the Argumentative-Expository Systems

Every *TheSu* XML sheet corresponds to at least one work to be annotated. Considering the general need for historians to keep track of the textual *locus* of every passage that they analyse and quote, it's better for the annotator to work on already-existing XML/TEI editions of the texts, if suitably provided with `milestone` elements with IDs corresponding to the desired reference system. This has been the case with the adopted digital edition of *Aq.* Often *TheSu* elements need to include non-contiguous spans of text. These, in turn, can often be interpreted as composing multiple theses or supports (explicit or implicit) cumulatively, sometimes leading to the problem of overlapping hierarchies. For these two reasons stand-off markup has been chosen as the annotation method for *TheSu*: each of its elements has to refer to a span of text in another document, linked through xLink and xPointer.

[2] See Davidson's "Principle of Coherence" (Davidson, 1991).

Every *TheSu* sheet contains an Argumentative-Espository System (AE System), that is theoretically defined as a set containing theses, their argumentative and expository supports, and the functional relations between the two. As will be shown below, this also needs to include a few more elements in its digital implementation.

A "**thesis**" is an instantiation of a declarative proposition at a certain point of the text representing the stance of its speaker. It can be explicit in the form of an enunciate (*e.g.* 'Putrefaction is the decay of liquids in the flesh', *Aq.* 957[e]) or implicit, *e.g.* in the form of a rhetorical question (*e.g.* '[Water is more useful to humans than fire]' in «how, then, should water not be more useful… ?», 957[b]).

A "**support**" is a segment of text that is presented by its speaker *in function of* a part of the scientific discourse conveyed by the same text. A "support" can:

[1] provide justifications for the acceptance or refusal of a thesis or of another support (*argumentative support*): *e.g.* «In most cases, it's not possible to use water without fire: in fact, it's more useful when it's heated, otherwise it's harmful», 958[a].

[2] explain more clearly, stylistically, or in depth the meaning of another segment of text containing theses and/or supports (*expository support*): *e.g.* «Isn't it more helpful what we always and continuously stand in need of, like a tool and an instrument, …?», 955[f];

[3] expand on an information conveyed by a thesis, favouring a more complete knowledge and understanding of it (*expansive support* or *excursus*): «… and (don't you see) that every sense partakes of fire, as it fabricates the vital principle, and especially sight, which is the keenest of the bodily senses, being an ignition of fire… ?», 958[e];

[4] contextualize the interpretation and reception of another segment of text containing theses and/or supports (*contextualizing support*): «In fact, (about) the saying that sometimes humans exist without fire: humans can't at all exist (without it)», 958[b].

The reader here may notice that in *TheSu*'s annotation scheme the "support" elements, having four distinct functions, include rhetorical uses that do not correspond directly to argumentative and expository aims. One can still speak of "Argumentative-Expository Systems", though, because careful consideration of both the expansive and contextualizing supports is needed for a complete understanding of the argumentative and expository roles of the theses surrounding them, and of their linked segments of text.

"Theses" and "supports" are encoded as `THESIS` and `SUPPORT` XML elements, both children of an `AEsystem`, which is in turn child of a `work`. *Aq.*'s AE System, in its current version, contains 259 manually annotated `THESIS` elements (corresponding to 334 theses, 56 of which are implicit) and 216 `SUPPORT` elements (121 implicit). These numbers are striking if the very short nature of the text is considered (1627 words in total). It's clear that a high amount of information on an author's thought and on her cultural context can always be extracted from even relatively small bits of text: mapping it in detail can be crucial to avoiding misinterpretations and misattributions.

Every `THESIS` and `SUPPORT` must have its own ID, so that each can be targeted by `SUPPORT` elements through xPointer. `THESIS` elements' IDs are also necessary for the most original feature of the *TheSu* annotation scheme. Absent, to the best of my knowledge, from current Argumentation Mining techniques is the possibility of linking together unrelated argumentative-expository chains when converging towards the same idea. It is a need for the historian, when studying the thought of a certain author, to have a clear view of how the same theses are presented and argued for in different contexts, even when unrelated. For example, if the author does not provide supports for a judgement in a certain work or paragraph, it does not necessarily mean that she does not argue for it, or better explains it, elsewhere. To have a map where all its occurrences in different *loci*, with all their corresponding argumentative-expository apparatuses, are linked together, would naturally be helpful to the researcher. This is made possible, in *TheSu*, through the creation of a "propositions" sheet containing only `PROPOSITION` elements (a modified version of `THESIS` for the annotation of non-textual declarative sentences), and by linking to their IDs all the textual `THESIS` elements instantiating them. In *Aq.*, the proposition *e.g.* '{ Water is more useful than fire }' is repeatedly argued for in different manners, and implicitly conveyed by the words in [a] 955[f]-956[a], [b] 956[c] and [c] 957[b]. The thesis at [a] is the target of 5 supports, the one at [b] of 5 more, and the one at [c] of only 2. It is undesirable to keep these 12 supports fragmented in their respective rhetorical chains, as they all converge towards the same idea. Indeed, it is interesting to see how this proposition is argued for in *all* of its enunciative occurrences. Accordingly, it is preferable to connect each of the textual theses to their common abstract proposition within the same network. The usefulness of such a connection becomes even clearer if one imagines its extension to the whole textual production of an author, as well as to works from different authors.

What follows is a non-exhaustive presentation of some of the required or optional attributes and sub-elements of the [i] `THESIS` and [ii] `SUPPORT` elements.

[i] Every **THESIS** has an @id, a @value (affirmative or negative) and a @quantity. It can sometimes be @implicit (boolean), as has been explained above. Every non-propositional THESIS can have one or more child elements instanceOf, each with a @propRef pointing to the corresponding PROPOSITION. A required child element is the speakersGroup, containing at least one speaker, corresponding to the person, group or entity the thesis is interpreted to be 'pronounced' by, with a @ref pointing to its name in an authority sheet. The THESIS's child element assent is used to specify whether the thesis is shared, unaccepted or actively attacked by its speaker (sub-element assentSpeaker with its @assentValue), or by the author of the work (assentAuthor). The child element thesisType mainly serves indexing purposes, as it classifies the THESIS through its sub-elements: value (epistemic — to specify with @valueTag whether the thesis is offered as the speaker's real stance, as a hypothesis, or fictitiously), macroThemesGroup (to specify the 'macroscopic' theme(s) of the thesis, *e.g.* "physical", "historical", "axiological"), microThemesGroup (for the 'microscopic' theme(s) of the thesis, *e.g.* "physiology", "cosmology", "dialectic"), and keywordsGroup (to point through keywordRef elements to the textual or implicit keyword(s) corresponding to the object(s) of the thesis).

Note that each keywordRef's @ref links to the ID of a **keyword** that is a child of AEsystem. Separating the keywords from the theses becomes necessary due to the possibility of different theses including the same keywords: in 957^{c} («but, in general, water (τὸ ὕδωρ) is so far away from being self-sufficient for self-preservation or the bringing-forth of other things that lack of fire, for it, is even destruction») the theses 'not(Water is self-sufficient for self-preservation)', 'not(Water is self-sufficient for the bringing-forth of other things)' and 'Without fire, water is destroyed' all share the textual keyword τὸ ὕδωρ. Each keyword can point to a segment of the annotated text or be 'implicit', and must always be tagged semantically through an attribute @namely, pointing to a class in a vocabulary sheet (*e.g.* "water"). Although the choice of the controlled vocabulary can be left to the interpreter, all new exhaustive *TheSu* annotations should consider the keyword classes already used in the previous ones, to facilitate the linking of the novel theses to all the corresponding previous propositions. It is better not to refer to an ontology of real-world entities, both to free the classification from the need of specifying vague or untranslatable terms, and to avoid projecting alien categories of thought to different cultural and scientific contexts. More freedom can be granted in the choice of the classes for the "macro-" and "micro-themes", as coherent keywords give sufficient help for the discovery and aggregation of (quasi-)equivalent theses. Each of the microTheme and keywordRef elements also has an attribute @focus to specify, by order of rank, their relative prominence in the thesis: the one just quoted, 'Without fire, water is destroyed', is about "water" and "fire" and includes both as its keywords, but it's more relevant to an understanding of Plutarch's ideas on water than those on fire. The keywordRef linked to it has thus been given @focus = 1, and the other @focus = 2. keywordRef can be used as grounding for visualizable analyses such as the one in Figure 2, where fire- and water-related keywords are assigned a score ("Epistemic relevance") based on the quantity of THESIS elements containing them at different points of the text, weighted on the basis of their @focus. One can learn from such a graph that a comparative style is maintained (almost) throughout the text, instead of it featuring two 'separate' speeches on the individual excellence of each element: such an analysis can lead to interesting findings if compared to similar analyses of other works of the same genre.

Other child elements of THESIS are recap and text. The former contains a short paraphrase in English of the thesis as interpreted and annotated: no logical formalization is required, as the annotation process must remain accessible to interpreters untrained in logic. The same goes for the PROPOSITION elements' recap: avoiding a strict logical formalization of the propositions allows the interpreter to consider as their instances theses that are not quite logically equivalent, but that can count as *synonymous enough* for the History of ideas, as is the

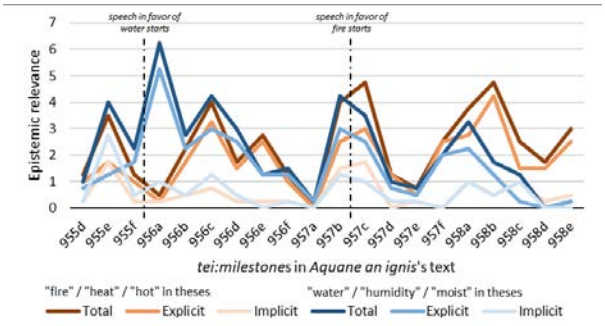


Figure 2. Relevance of fire- and water-related keywords to the theses conveyed by different contiguous spans of *Aquane an ignis*'s text.

	Use in SUPPORTS	Form of SUPPORT	Justified?	THESIS quantity	Total	Justified / Total
THESES	Employed in justifications	as premises	yes	36	74	49%
			no	38		
		as illustrations	yes	65	140	46%
			no	75		
		in other forms	yes	9	50	18%
			no	41		
	Employed in explanations		yes	9	30	30%
			no	21		
	Employed in jstf./expl.?	as examples	yes	0	1	0%
			no	1		
	Unused		yes	19	83	23%
			no	64		
	All THESES		yes	208	334	62%
			no	126		

Table 1. Theses in *Aquane an ignis* in relation to supports: by how many and in which forms they are employed, and by how many they are targeted.

case with the thesis in the bottom-right corner of Figure 1 (quoting 'Fire is first, in relation to water') in respect to 'Fire is better than water'. Finally, the `text` points through its sub-element `textRef` (containing at least one `segment` with `@from` and `@to`) to the textual proof of the existence of the thesis at a certain point of the discourse.

[ii] `SUPPORT` elements share with `THESIS` the attributes `@id` and `@implicit`. The sub-elements `speakersGroup`, `assent`, `recap` and `text` are present here as well. The first unique child element of the `SUPPORT` is `targetsGroup`, containing at least one `target` pointing through `@ref` to the ID of a supported element. Very useful is `empolyedTheses`, including one or more `thesisRef` (with `@ref`) to link to the theses in the `SUPPORT`'s textual span that are actually presented to support the targeted element(s), discriminating between them and other non-relevant theses possibly annotated in the same text, thus solving ambiguities.

For mainly indexing purposes, as with `thesisType`, each `SUPPORT` element contains a `supportType`, also necessary for the analysis of the reasoning styles of the discourses they are part of. While their child element `value` is identical to the one in `thesisType`, they also include their own `function` and `form`. The `function`'s sub-elements are `justification`, `explanation`, `expansion` and `contextualization`, each with a `@rank` (default = 4) representing their relative centrality to the support (most central = 1). The idea is that every support, as everything else in a cohesive discourse, is always at the same time justifying, expository, expansive and contextualizing of its surroundings to a certain degree (cf. Perelman, Olbrechts-Tyteca, 2013 [1958], p. 203), and that its speaker, in order to achieve different rhetorical effects, simply choses to make one or another of these functions more prominent than the others. The possibility of ranking the functions solves the problems that would come from having to choose *only one* of them even in cases where there is enough ambiguity to make it seem impossible. For the annotation of whether the support, when "justifying", serves the purpose of arguing *for* or *against* its target(s), `justification` has been given the attribute `@for` (= "acceptance", "refutation" or "mix"). Finally, using the element `form` the interpreter can classify the support by its rhetorical type, referring through `@formTag` to any class in a typology contained in an authority sheet. The *TheSu* standard typology of supportive forms is meant to be very simple and intuitive for intellectual historians: among the "justifying" forms, the "logical premise" is a sentence from which the supported target can be inferred by deduction, the "illustration" is a particular case from which the conclusion can be derived by induction, the "authority" is an appeal to an authoritative figure that adheres to the targeted idea, etc. Table 1 illustrates a quantitative analysis strictly dependent on the elements `SUPPORT`, `function` and `form`: it is not surprising that in a rhetorical work such as *Aquane an ignis* a very high amount of theses are given argumentative support (62%), but it is not necessarily expected that "illustrative" supports are twice the deductive "premises" (140 to 74), characterizing the speech as scarcely "logical" in tone and much more "exemplary". It is also interesting that theses employed in supports tend here to attract further argumentation, especially the "premises" (49% justified) and "illustrations" (46%), in contrast with the theses not used in supports (23%). This breakdown is only a small tile of the mosaic that is Plutarch's personal argumentation style, waiting for further analyses to be combined with and compared to.

Conclusion

The previous sections have described the essential features of the *TheSu* annotation scheme, its theoretical framework, and some of the potential uses of a *TheSu* sheet. This exposition has focused on the methodological usefulness of this kind of argumentation and exposition mapping for an historian working on a text, but *TheSu* can also be helpful for an optimal, *transparent* and *reusable*, exposition of the basis and results of her research: a historian's 'secondary' interpretation of a certain text —*e.g.* its ideas' dependency from the ones in a contemporary philosophical current, or their ideological or popular nature— always depend on a 'primary' interpretation of the argumentative and expository chains it is composed of. Storing these primary interpretations in easily-accessible *TheSu* databases would help with the evaluation of the secondary interpretations proposed by the historian, and would facilitate the work of future researchers who wish to build upon her research and generate new interpretations from the argumentative-expository material. This is only possible thanks to digital interfaces and database interrogation techniques, and would otherwise be too difficult and/or time consuming using traditional, non-digital methods.

Acknowledgements

This publication is part of the research project *Alchemy in the Making: From Ancient Babylonia via Graeco-Roman Egypt into the Byzantine, Syriac, and Arabic Traditions*, acronym *AlchemEast*. The *AlchemEast*

project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (G.A. 724914)[3].

References

Bernardakis, G.N. (Ed.), 1895. Plutarchi Chaeronensis Moralia, vol. 6. Teubner, Lipsia.

Betti, A., van den Berg, H., 2016. Towards a Computational History of Ideas, in: Wieneke, L., Jones, C., Düring, M., Armaselu, F., Leboutte, R. (Eds.), Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age. CEUR Workshop Proceedings, CEUR-WS.Org, vol. 1681. Aachen.

Betti, A., van den Berg, H., Oortwijn, Y., Treijtel, C., 2019. History of Philosophy in Ones and Zeros, in: Fischer, E., Curtis, M. (Eds.), Methodological Advances in Experimental Philosophy. Bloomsbury Academic, Great Britain, pp. 295–332.

Carlson, L., Marcu, D., Okurowski, M.E., 2001. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory, in: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, SIGDIAL '01, vol. 16. Association for Computational Linguistics, PA, USA, pp. 1–10. Stroudsburg, https://doi.org/10.3115/1118078.1118083

Cerrato, L., Almas, B., TDBuck, srdee, ahanhardt, Clérice, T., … Sowell, E., 2019, June 10. PerseusDL/ canonical-greekLit 0.0.1923 (Version 0.0.1923). Zenodo. http://doi.org/10.5281/zenodo.3525369

Cherniss, H., Helmbold, W.C. (Eds.), 1957. Plutarch. Concerning the Face Which Appears in the Orb of the Moon. On the Principle of Cold. Whether Fire or Water Is More Useful. Whether Land or Sea Animals Are Cleverer. Beasts Are Rational. On the Eating of Flesh, Plutarch's Moralia Vol. 12. Loeb Classical Library No. 406. Harvard Univ. Press, Cambridge, Mass.

Davidson, D., 1991. Three Varieties of Knowledge. Royal Institute of Philosophy Supplements 30, 153–166.

Dibattista, L. (Ed.), 2009. Storia della scienza e linguistica computazionale: sconfinamenti possibili. Angeli, Milano.

D'Ippolito, G., Nuzzo, G. (Eds.), 2012. Plutarco. L'origine del freddo. Se sia più utile l'acqua o il fuoco, Corpus Plutarchi Moralium, 49. M. D'Auria, Napoli.

Green, N.L., 2018a. Towards mining scientific discourse using argumentation schemes. Argument & Computation 9, 121–135. https://doi.org/10.3233/AAC-180038

Green, N.L., 2018b. Proposed Method for Annotation of Scientific Arguments in Terms of Semantic Relations and Argument Schemes, in: Proceedings of the 5th Workshop on Argument Mining. Association for Computational Linguistics, Brussels, Belgium, pp. 105–110. https://doi.org/10.18653/v1/W18-5213

Lauscher, A., Glavaš, G., Ponzetto, S.P., 2018. An Argument-Annotated Corpus of Scientific Publications, in: Proceedings of the 5th Workshop on Argument Mining. Association for Computational Linguistics, Brussels, Belgium, pp. 40–46. https://doi.org/10.18653/v1/W18-5206

Lippi, M., Torroni, P., 2016. Argumentation Mining: State of the Art and Emerging Trends. ACM Transactions on Internet Technology 16, 10–10. https://doi.org/10.1145/2850417

Mann, W.C., Thompson, S.A., 1987. Rhetorical structure theory: Description and construction of text structures, in: Kempen, G. (Ed.), Natural Language Generation. New Results in Artificial Intelligence, Psychology and Linguistics, NSSE, 135. Martinus Nijhoff Publishers, Dordrecht, pp. 85–95.

Marcu, D., Amorrortu, E., Tajahuerce Romera, M., 1999. Experiments In Constructing A Corpus Of Discourse Trees, in: Walker, M.A. (Ed.), Towards Standards and Tools for Discourse Tagging. Presented at the workshop, 21 June 1999, University of Maryland, College Park, Maryland, USA, NJ Association for Computational Linguistics, New Brunswick.

Milazzo, A.M., 1991. Forme e funzioni retoriche dell'opuscolo «*Aqua an ignis utilior*» attribuito a Plutarco, in: Gallo, I., D'Ippolito, G. (Eds.), Strutture formali dei «Moralia» di Plutarco. Atti del III Convegno plutarcheo - Palermo, 3-5 maggio 1989. M. D'Auria, Napoli, pp. 419–434.

Mochales Palau, R., Moens, M.-F., 2009. Argumentation mining: the detection, classification and structure of arguments in text, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law. ACM, pp. 98–107. https://doi.org/10.1145/1568234.1568246

[3] I am grateful to Eduardo Escobar, who accepted to check my English and gave me valuable feedback on content and methodology.

Peldszus, A., Stede, M., 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. International Journal of Cognitive Informatics and Natural Intelligence 7, 1–31. https://doi.org/10.4018/jcini.2013010101

Perelman, C., Olbrechts-Tyteca, L., 2013. Trattato dell'argomentazione. La nuova retorica, 4th ed. Einaudi, Torino (Italian translation of 1958. *Traité de l'argumentation. La nouvelle rhétorique*, PUF, Paris).

Rocha, G., Lopes Cardoso, H., Teixeira, J., 2016. ArgMine: A framework for argumentation mining, in: Silva, J., Ribeiro, R., Quaresma, P., Adami, P., Branco, A. (Eds.), Computational Processing of the Portuguese Language : 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016: Proceedings, Lecture Notes in Computer Science. Springer.

Stead, S., Doerr, M., Ore, C.-E., Kritsotaki, A., et al., 2019, October. CRMinf: the Argumentation Model. An Extension of CIDOC-CRM to support argumentation (Version 0.10.1). http://www.cidoc-crm.org/crminf/ModelVersion/version-10.1

Stede, M., Schneider, J., 2019. Argumentation mining, Synthesis lectures on human language technologies, 40. Morgan & Claypool Publishers, San Rafael.

Taboada, M., Mann, W.C., 2006. Applications of Rhetorical Structure Theory. Discourse Studies 8, 567–588. https://doi.org/10.1177/1461445606064836

TEI Consortium (Eds.), 2019, July 16. TEI P5: Guidelines for Electronic Text Encoding and Interchange (Version 3.6.0). TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Toulmin, S.E., 2003. The Uses of Argument: Updated Edition, 2nd ed. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511840005 (First edition: 1958. Cambridge University Press, Cambridge).

van den Berg, H., Parra, G., Jentzsch, A., Drakos, A., Duval, E., 2014. Studying the History of Philosophical Ideas: Supporting Research Discovery, Navigation, and Awareness, in: Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business. ACM, New York. https://doi.org/10.1145/2637748.2638412

Walton, D., 1998. The New Dialectic: Conversational Contexts of Argument. University of Toronto Press.

Walton, D.N., Reed, C., Macagno, F., 2008. Argumentation Schemes. Cambridge University Press, Cambridge.

Leitwort Detection, Quantification and Discernment

Racheli Moskowitz, Moriyah Schick, Joshua Waxman
Stern College for Women, Yeshiva University
New York, NY
joshua.waxman@yu.edu

Abstract

English. A *Leitwort* ("leading word") is a word or word-root deliberately repeated as a literary technique, for the sake of emphasis or to establish an underlying theme. *Leitworte* occur in many different works of literature, and traditional humanists who work to interpret those texts apply their literary expertise to discern those occurrences and explain their function. There is an element of subjectivity involved, and so, working from a digital humanities perspective, we developed an algorithm to detect potential *Leitworte* and rate their significance, by counting repetitions within a moving window and calculating tf-idf scores for candidate words. This reflects a convergence of humanistic and computer science methodologies. By compiling a list of *Leitworte* identified by an expert traditional literary scholar and then comparing it to the list produced by our computational approach, we attempt to (a) evaluate the subjective work of a particular human scholar against an objective standard, (b) consider whether the algorithmic approach we chose was sophisticated enough to match honed scholarly discernment, and (c) reflect on the difference between a traditional and digital humanities approach. In particular, we examine the Hebrew Bible and the *Leitworte* identified by Umberto Cassuto, an Italian Biblical scholar.

Italiano. Una *Leitwort* ("parola guidante") è una parola o parola-radice deliberatamente ripetuta come tecnica letteraria, allo scopo di porre enfasi o stabilire un tema sottostante. Le *Leitworte* si presentano in molte diverse opere di letteratura, e gli umanisti tradizionali che lavorano per interpretare questi testi applicano la propria competenza letteraria per discernere tali occorrenze e spiegare la loro funzione. Si coinvolge un elemento di soggettività, e quindi, lavorando alla prospettiva degli studi umanisti digitali, abbiamo sviluppato un algoritmo per rilevare potenziali *Leitworte* e valutare la loro significatività, contando le ripetizioni all'interno di una finestra mobile e calcolando punteggi tf-idf per le parole candidate. Questo riflette una convergenza di metodologie umanistiche e informatiche. Compilando una lista delle *Leitworte* identificate da parte di uno studioso di letteratura tradizionale esperto e poi comparandola alla lista prodotta dal nostro approccio computazionale, tentiamo di valutare il lavoro soggettivo di un particolare studioso umano in confronto ad uno standard oggettivo; considerare se l'approccio algoritmico da noi scelto è abbastanza sofisticato da tenere testa ad un discernimento affinato dagli studi; e riflettere sulla differenza fra un approccio tradizionale e uno digitale agli studi umanistici. In particolare, esaminiamo e le *Leitworte* identificate da Umberto Cassuto, studioso biblico italiano.

1 Introduction

1.1 The *Leitwort* in Literature

In music, a *Leitmotif* is a recurring musical phrase (a "motif") which is used to "lead" or guide the listener to recall something or make a connection. In literature, the *Leitwort* ("leading word") plays a similar function. A certain word, or word root, is unusually repeated several times in a passage, in a way that jumps out at the reader, in order to establish a theme. Optionally, once established, it might then be echoed in a later passage to recall that theme. Here is one of the many *Leitworte* which Pinault (1986) identifies in his analysis of stylistic features in *The Arabian Nights*. In the story "The City of Brass", over the course of a few consecutive pages, in poetry and prose, there is repetition of words with the Arabic root موت / *mawt,* meaning death.

أبادهم الموت / abadahum **mawt**, "Death destroyed them".
كأس الموت / ka`s al-**mawt**, "the cup of doom".
إن كان موتي صلاة سجال / in kana **mawti** mahtuman cala cajal, "when my death was decreed all at once".
بإسم الحي الذي لا يموت / bi-ism al-hayy alladhi la **yamut**, "in the name of the Living, who dies not".
يُزخرفها الشيطان للإنسان إلى الممات / yuzakhrifuha al-shaytan lil-insan ila al-**mawt**, "Satan adorns it for man to lead him to death". (MacNaghten edition)

This literary technique has been discussed by scholars as it is found in sacred texts such as the Hebrew Bible (Alter, 1981) and the Koran (Wansbrough, 1978). It is similarly employed in the works of Goethe, Nietzsche, Heidegger and others. Then, one task of those who analyze these texts is to discern the use of a *Leitwort* and explain its purpose.

1.2 *Leitworte* and the Hebrew Bible

Various interpreters of the Hebrew Bible within the past century have taken note of the use of the *Leitwort*, though they differ as to its parameters and purpose. In Buber (1927), and in Buber and Rosenzweig (1936), the purpose of a thematic repetition in a passage is to reveal or clarify a meaning in the text, or to emphasize that meaning. They select *Leitworte* based on their subjective estimation of the word's significance, rarity, and the degree of repetition. These are all factors which contribute to a word capturing the attention of the reader. The repetitions can occur densely in a single passage, or can be distributed throughout the text. They draw connections between passages in which the same *Leitwort* occurs, as indicating an allusion or thematic echo. In one famous example from Buber, the seven scenes of revelation that compose the Abraham story arc are all tied together by the use of the term ראה / *ra'ah* / "see" in each passage, and unlike other translations which obscure this connection, Buber and Rosenzweig's German translation preserves the *Leitwort* by translating it consistently throughout.

At around the same time, Umberto Cassuto also took an interest in *Leitworte* in the Hebrew Bible. Cassuto was first the chief rabbi of Florence and subsequently a professor at University of Florence and at the University of Rome La Sapienza. Like Buber and Rosenzweig, Cassuto (1961, published posthumously) considered that *Leitworte* in the Bible served the purpose of establishing the theme of a passage or emphasizing a point. However, he was more selective in what types of repetitions he would consider to be bona fide *Leitworte*. In Cassuto's view, repetitions are only interpretable if they occur within a coherent passage and occur in a multiple of 3, 7, or 10. As illustrated below, he discusses the threefold repetition as a way of emphasizing a point within or between passages. The numbers seven and ten are chosen because they are particularly significant to an ancient Israelite author from a ritual and spiritual perspective. For instance, the number seven is echoed from the Creation narrative to a weekly cycle of seven days, a seven-year cycle before each Sabbatical year appears, and seven Sabbatical years leading up to the Jubilee. Similarly, the number ten brings the Ten Commandments immediately to mind.

We present an extended example to illustrate the position of *Leitworte* in Cassuto's textual analysis. In *Exodus* 2:2-6, in the background of Egyptian governmental decrees to drown all firstborn Hebrew boys, Moses is born. This story, Passage A, contains the word-root ראה */ ra'ah* / "saw", repeated three times:

Moses' mother "saw [**וַתֵּרֶא / vateire`**] him that he was a goodly child", hid him for as long as she could, and then placed him in an ark on the riverside. When Pharaoh's daughter visited the river, "she saw [**וַתֵּרֶא / vateire`**] the ark among the flags, and sent her handmaid to fetch it. (6) And she opened it, and saw it [**וַתִּרְאֵהוּ / ־vatir`eihu**], even the child; and behold a boy that wept. And she had compassion on him, and said: 'This is one of the Hebrews' children.' " (Jewish Publication Society Translation)

A few verses later (*Exodus* 2:11-15), Moses, after being raised as a prince in Pharaoh's house with his biological mother as his nursemaid, decides to check on his enslaved brethren, sees their suffering, and reacts. Once again, the thrice-repeated index root is *ra'ah*, though there are others marked. We label this Passage B.

יא וַיְהִי בַּיָּמִים הָהֵם, וַיִּגְדַּל מֹשֶׁה וַיֵּצֵא אֶל-אֶחָיו, **וַיַּרְא**, בְּסִבְלֹתָם; **וַיַּרְא** אִישׁ-מִצְרִי, **מַכֶּה** אִישׁ-עִבְרִי מֵאֶחָיו. יב וַיִּפֶן כֹּה וָכֹה, **וַיַּרְא** כִּי אֵין אִישׁ; **וַיַּךְ**, אֶת-הַמִּצְרִי, וַיִּטְמְנֵהוּ, בַּחוֹל. יג וַיֵּצֵא בַּיּוֹם הַשֵּׁנִי, וְהִנֵּה שְׁנֵי-אֲנָשִׁים עִבְרִים נִצִּים; וַיֹּאמֶר, לָרָשָׁע, לָמָּה **תַּכֶּה**, רֵעֶךָ. יד וַיֹּאמֶר מִי שָׂמְךָ לְאִישׁ שַׂר וְשֹׁפֵט, עָלֵינוּ--**הַלְהָרְגֵנִי** אַתָּה אֹמֵר, כַּאֲשֶׁר **הָרַגְתָּ** אֶת-הַמִּצְרִי; וַיִּירָא מֹשֶׁה וַיֹּאמַר, אָכֵן נוֹדַע הַדָּבָר. טו וַיִּשְׁמַע פַּרְעֹה אֶת-הַדָּבָר הַזֶּה, וַיְבַקֵּשׁ **לַהֲרֹג** אֶת-מֹשֶׁה; וַיִּבְרַח מֹשֶׁה מִפְּנֵי פַרְעֹה, וַיֵּשֶׁב בְּאֶרֶץ-מִדְיָן וַיֵּשֶׁב עַל-הַבְּאֵר.

(11) And it came to pass in those days, when Moses was grown up, that he went out unto his brethren, and looked [**vayar`**] on their burdens; and he saw [**vayar`**] an Egyptian smiting [**makeh**] a Hebrew, one of his brethren. (12) And he looked this way and that way, and when he saw [**vayar`**] that there was no man, he smote [**vayakh**] the Egyptian, and hid him in the sand. (13) And he went out the second day, and, behold, two men of the Hebrews were striving together; and he said to him that did the wrong: 'Wherefore smitest [**takeh**] thou thy fellow?' (14) And he said: 'Who made thee a ruler and a judge over us? thinkest thou to kill me [**haleharegeni**], as thou didst kill [**haragta**] the Egyptian?' And Moses feared, and said: 'Surely the thing is known.' (15) Now when Pharaoh heard this thing, he sought to slay [**laharog**] Moses. But Moses fled from the face of Pharaoh, and dwelt in the land of Midian; and he sat down by a well. (Jewish Publication Society Translation)

Before we turn to Cassuto's analysis, a brief discussion of Passage B can help us understand the general phenomenon of *Leitwort*. Note that while ראה / *ra`ah*, "saw", is repeated three times, this is evident only when reading the passage in the original Hebrew, rather than in the English translation above. It appears twice in verse 11; the first time it is rendered as "looked" and the second time as "saw". Meanwhile, in verse 12, the word "looked" appears translating a different word-root, פנה / *panah*, while the word "saw" translates ראה / *ra'ah*.

These threefold repetitions seem deliberate. Firstly, an author can include or omit details while still advancing the narrative, so many times, a repeated word didn't truly need to appear. Secondly, Hebrew has synonyms. An author can select from an inventory of terms. For instance, in the same passage, the roots *nakhah* ("smite", "strike") and *harag* ("kill") are used to refer to Moses' killing of the Egyptian taskmaster.

In analyzing Passage B, Cassuto notes the threefold repetition of the root *ra'ah*. He says it is for emphasis. It matches the threefold repetition of the same root in passage A. The parallel is not accidental, but it is to stress that just as Moses' mother, and Pharaoh's daughter saw and had mercy on him, so did Moses take pity and have mercy on his brethren. Cassuto also notes the threefold repetitions of smiting and killing.

1.3 The Problem of *Leitwort* Operationalization

Despite the value of *Leitworte* as a literary technique that unifies a text and enriches the experience of the reader, any attempt to accurately identify *Leitworte* is somewhat problematic, particularly in a large and varied corpus such as the Hebrew Bible. The problem, essentially, is that language is repetitive by nature. Zipf (1945) discusses the relative frequencies of all words in a corpus, repetitions within clusters, as well as intervals between clusters, and these phenomena are observed in the absence of deliberate stylistic repetition.

Subject matter or narrative concerns can require a certain word to appear more than once in a story. For instance, the first chapter of Carlo Collodi's *Le Avventure di Pinocchio* contains repetitions of the Italian word *legno*, "wood" and *pezzo*, "piece". This is because it is where we first encounter a talking, weeping, and laughing piece of wood. Even if Collodi had no intention of drawing the reader's attention to these words, the narrative would be senseless without them. Luhn (1958) demonstrates that this occurs in non-narrative texts as well, and establishes the content of a text based on the non-deliberate repetition of words or word roots as an author advances his arguments or elaborates on an aspect of a subject. Additionally, certain words, such as function words (e.g., the articles "a" and "the"), are extremely common in language because they assist communication. These words will necessarily occur numerous times throughout a text, and will be totally unrelated to any theme that the writer wishes to emphasize. In fact, Luhn suggests that words that occur above a certain frequency threshold can generally be considered insignificant.

In the midst of a sea of non-significant word repetitions, identification of meaningful *Leitworte* for further study poses a challenge which can be addressed in one of two ways. The first solution is to maximize subjectivity, simply relying on a human interpreter to name which words should be considered *Leitworte*. This traditional approach has the benefit of embracing the nuances of human insight and the finely-honed skills of expert analysis, but can also be considered arbitrary and subjective by its very nature. An alternative solution, which we implement in this paper, is to introduce objective measurement tools in an attempt to quantify significance of repeated words in a given passage. Of course, our algorithm cannot truly appreciate a text, and standardized rules do not take intuitive understanding into account. However, our program has the advantage of working systematically to find every candidate *Leitwort*, in contrast to human experts who are subject to the limitations of their scanning and matching capability and who may be biased by selective interest in certain terms. This novel approach allows us to move towards greater objectivity in analysis, and is a great tool for scholars who want to consider every potential *Leitwort* in the text.

In the next section, we detail our programmatic approach. Our goals were to devise a quantitative measure of repetition significance, and to compare output of our program to the list of *Leitworte* identified by Cassuto, a traditional expert with a relatively systematic approach. To that end, our operational definitions are modeled on Cassuto's definitions but strip out the subjective components of his expert analysis.

2 Our Approach

For clear and consistent *Leitwort* identification, several practical questions must be answered. First, what constitutes a "word"? Second, what is a qualifying number of repetitions? Third, how closely spaced must the repetitions be? Fourth, how is significance of candidate words defined? Here we explain how we addressed each of these questions in designing our program.

2.1 Reduction to Lexemes or Roots

Since Semitic languages such as Hebrew are inflected, and the typical *Leitwort* is based on repetition of the word's root, we first reduce the words in the corpus to their lexemes. For the Hebrew Bible, we use the ETCBC dataset described in Roorda (2015), which contains manual marking of rich linguistic features by human experts. One such feature is the lexeme, which is a close approximation to the root. In Hebrew, most words contain a triliteral root which conveys a core meaning. For instance, אור / *`or* has the meaning of "light". In the ETCB dataset, words with this root are divided into two separate lexemes, אור / *`or* ("light") and מאור / *ma`or* ("luminary", thing which gives light). The lexemes are stemmed versions of the full word, stripping out definiteness, gender, number and person. Thus, the full word לִמְאוֹרֹת / *lim'orot* / "as luminaries" in *Genesis* 1:15 is marked with the lexeme מאור / *ma`or* while the word לְהָאִיר / *leha`ir* / "to give light" in the same verse is marked with the lexeme אור / *`or*. The ETCB dataset does not have a root feature.

We differ here from Cassuto, who primarily considers repetitions of roots. However, it is noteworthy that many of these lexemes are also the simple root (such as the אור example above). Of the 788 sevenfold lexeme-based *Leitwort* candidates our algorithm discovered across the Pentateuch, 81% consisted of triliteral roots. Many of the non-root lexeme candidates are names of nations or places.

2.2 Counting Repetitions

Following Cassuto, our candidate *Leitworte* must occur a multiple of 3, 7, or 10 times in a passage. We gravitate towards Cassuto's definition of *Leitwort* for a few reasons. Many modern Biblical interpreters (such as Elchanan Samet of Yeshivat Har Etzion) employ both Buber and Cassuto-type *Leitworte* in their analyses, but consider the more rigorously defined Cassuto-type *Leitworte* as especially significant (Grossman, 2011). Further, as discussed above, Cassuto shows that these are meaningful numbers for an ancient Israelite author, and he consistently demonstrates that thematic words are repeated this precise number of times, or a multiple thereof. Indeed, we are treating this threefold and sevenfold repetition as a mark of authorial deliberateness – that the author has set out to employ the *Leitwort* style. If a word were repeated by chance, simply because it is the topic of a passage (see the *legno* and *pezzo* examples above) or because it is a commonly occurring word (such as "said"), then it mostly would not occur specifically as a sevenfold repetition.

2.3 Scanning for Repetitions

We scan for repetitions in the text, in a moving window. For face validity, we require a certain minimum density of repetition. Buber did not require close proximity; the seven Abraham scenes that he connects with the root ראה / *ra'ah* / "see" span 178 verses over 8 chapters. Cassuto only identified *Leitwort* occurring within self-contained passages, but personally determined section and paragraph boundaries based on his own close reading analysis. Neither of these approaches is appropriate to our method, as both rely on subjective expert judgement to decide whether a given set of repetitions occurs within an acceptable space.

To define objective limits, we turned to the historical Jewish segmentation scheme of the *sidra*: the entire Pentateuch is chanted by a reader in synagogues over the course of a lunar year, one portion each week, on the Jewish Sabbath, though there are modifications due to holidays. The Pentateuch was divided into 54 such portions, or *sidrot*. While the calendar influenced the number of portions, scholars segmented the text at appropriate positions, such that there is often a consistency in the narrative or legal codes within the text. To make use of this narrative consistency, we only count repetitions within a *sidra*. Another Biblical segmentation scheme, of Christian origin, is the well-known series of chapter divisions (e.g. Genesis 1, Genesis 2), which breaks up the full text into chapters of about 30 verses each. We further require our repetitions to occur within a maximum window of 60 verses (approximately 2 chapters). Thus, if a word randomly occurs 4 times at the start of a *sidra* and much later has a sevenfold repetition, for a total of 11 occurrences, the sevenfold repetition will still remain a candidate.

Once a qualifying repetition has been found, we continue scanning along two pathways: one in which the passage stops with the most recent verse (and can thus be far smaller than 60 verses in length), and one in which it continues and allows for higher multiples to be identified.

Because our sections have flexible starting and ending points, a separate method must ensure that word appearances from other sections are kept distinct. In line with Cassuto's numerical definition of *Leitworte*, an eighth appearance of a candidate word in the same passage should disqualify the word. However, we would not wish to incorrectly disqualify a candidate merely because it occurs in an unrelated passage later in the

sidra. Using his idiosyncratic paragraph divisions, Cassuto would find a sevenfold repetition within a paragraph and ignore an unassociated occurrence one or two paragraphs earlier. Lacking such boundary lines, we create a buffer zone around each of our identified *Leitwort* passages. This zone is defined as ¼ the number of sentences of the passage span, and we require a total absence of the candidate word within that zone. Thus, a word will qualify as a *Leitwort* candidate if it occurs 7 times within a 16-verse span but does not appear at all in the 4 preceding and 4 subsequent verses. By requiring the word to appear in this "island," we create de facto passage boundaries in a flexible way.

2.4 Filtering for Significance

Finally, we rate the candidate words for significance. In making estimation of significance the last step in our process, we diverge from the traditional expert-reader model. Scholars such as Cassuto and Buber would start with an impression that a word was significant, in the sense of meaningful and important. To Buber, if such a word was relatively rare (an undefined term) and also repeated within a story, it was a *Leitwort*. Cassuto required a precise number of repetitions and did not restrict based on rarity, but only examined words that he deemed especially significant rather than identifying every threefold or sevenfold repetition. Indeed, it would be simplistic to say that all of them are significant; these numbers can occur by chance just like any other.

Therefore, after systematically compiling a list of all sevenfold repetitions within our corpus, we employ a tf-idf measure to weed out the most clearly insignificant of them. The term frequency (tf) is the number of times a word appears in a given document, while the inverse document frequency (idf) is the log of the total number of documents N divided by the number of documents that contain the word. If a word is frequent in the current document and infrequent elsewhere, then the product of the tf and the idf will be high. The "documents" we use for this computation are the 54 aforementioned *sidrot*, since the text in each such division will typically be of a consistent genre (e.g. genealogy, legal code, narrative) and topic (e.g. trials in the wilderness).

We stress that the purpose of this tf-idf ranking is not to discover the emphatic and thematic words. The specific numerical repetition establishes that. Rather, our aim was to filter out common yet highly insignificant words, which will occur in sevenfold repetition (along with eightfold repetition, ninefold repletion, etc.) purely by chance. For this reason, we set our tf-idf threshold very low, at 0.07. After examining a small portion of unfiltered candidates, we chose this value because it could retain words that appeared thematically relevant, while excluding common words with high frequency throughout the corpus but no discernable relevance to the specific passage. We did not use a simple stoplist of frequent words since a common word might be extremely significant in a given context. For instance, in *Genesis* 1-2, in which God creates the Universe in a sequence of speech acts, the lexeme אמר / *amar* / "said" occurs 28 times (a multiple of 7) and has a tf-idf score of 0.13, above our threshold of significance. It also occurs seven times in *Deuteronomy* 5-7 with a non-significant tf-idf score of 0.04.

3 Results

In the five books of the Hebrew Bible, we discovered a total of 788 potential *Leitwort* candidates that appeared a multiple of seven times in an island of text. Of these, 332 (or 42%) exceeded our tf-idf threshold and were counted as significant. Passage span ranged from 5 to 60 verses; and candidate lexemes were repeated within these passages 7, 14, 21, 28, 35, 42, or 49 times. As would be expected, threefold repetitions had shorter passage spans on average, and many fewer of them were deemed significant. We compiled a comprehensive list of Cassuto's *Leitworte* and compared them against the output of the program. Cassuto wrote commentary on the first 13 chapters (out of 50) of *Genesis* and on the entire (40 chapter) book of *Exodus*, identifying 164 *Leitworte,* of which 142 were of simple word or root repetitions. Of these root repetitions, 59 represented a sevenfold recurrence. For the same group of chapters, we found 207 potential candidates appearing a multiple of seven times, of which 102 (49%) exceeded our tf-idf threshold.

Table 1 cross-tabulates our results with Cassuto's. Twenty words were deemed significant by our program and also discussed by Cassuto, 82 are marked at *Leitworte* by our program only, 39 by Cassuto only, and 105 words that do not appear in Cassuto's work were originally flagged by our program but fell below our significance threshold.

		Algorithm	
		Yes	**No**
Cassuto	**Yes**	***Total: 20*** *Genesis: 8* *Exodus: 12*	**Total: 39** Genesis: 18 Exodus: 20
	No	**Total: 82** Genesis: 17 Exodus: 65	***Total: 105*** *Genesis: 23* *Exodus: 82*

Table 1: Cross-tabulation of the results of Cassuto and our algorithm. "Yes" means that it appears in the list (and, for the algorithm, deemed significant). Cells representing agreement of the two sources are italicized.

A few facts are apparent from these results. We see that our algorithm identified many more potential *Leitworte* overall than Cassuto. Also, Cassuto and the algorithm agreed about 50% of the time, and were much more likely to agree that a word was non-significant than that it was significant. Cassuto discussed many *Leitworte* that were not accepted by the algorithm, and vice versa. Finally, it is noteworthy that the results differ substantially based on specific text. Cassuto described almost as many *Leitworte* in the first 13 chapters of Genesis as in the 40-chapter Exodus. Meanwhile, the algorithm flagged repetitions with similar density across the two books and consistently identified about half of them (52% in Genesis, 48% in Exodus) as potentially significant. Cassuto and the algorithm therefore find about the same number of Genesis *Leitworte*, with few of Cassuto's appearing in the computer-generated list, whereas the algorithm finds more than twice as many Exodus *Leitworte* as Cassuto does.

If Cassuto's work is held up as the gold standard, one can say that the algorithm achieved 19.6% precision (32.0% in Genesis, 15.6% in Exodus) and 33.9% recall (30.8% in Genesis, 37.5% in Exodus). This suggests that it is able to catch about a third of the *Leitworte* discerned by an expert, and introduces a high number of spurious candidates, particularly in Exodus. Valid *Leitworte* can be missed by the program either because they are never identified or because they are rejected as insignificant. Close inspection of the data reveals that only 5 of Cassuto's *Leitworte* that were flagged by our algorithm fell below our tf-idf significance threshold. Most did not meet the algorithm's criteria for being a sevenfold repetition. This may be because we were restricted to using lexemes while Cassuto primarily used roots or was more flexible about linguistic features, or because we lacked his sharp boundaries of paragraph and story. Therefore, we may have inadvertently cut off our "passages" before the end of a scene, or disqualified a true *Leitwort* because it re-appeared in an unrelated context within our buffer zone. Further work can address some of these issues.

If, on the other hand, the objective algorithmic approach is considered the ideal, one can say that Cassuto obtained 33.9% precision (30.8% in Genesis, 37.5% in Exodus) and 19.6% recall (32.0% in Genesis, 15.6% in Exodus). This suggests he found about a fifth of possible *Leitworte* in his chosen text, and that about a third of his self-defined *Leitworte* are valid. Due to the limits of human attentional capacity, it would be practically impossible for a person to manually identify all existing *Leitworte* in such a complex text. Humans can be biased by their own interests to overlook many details, which can lead both to false positive and false negative detection errors. Notably, Cassuto's list for Genesis, the text in which he first perceived *Leitworte* and which evoked tremendous enthusiasm for the task, has the highest recall and lowest precision compared against the algorithm.

The truth probably lies somewhere between these extremes. Dismissing 80% of the algorithm's suggestions as invalid merely because Cassuto did not talk about them ascribes omniscience to the human expert, which is absurd. Similarly, it is ridiculous to say that Cassuto's analysis is only meaningful if its tf-idf score falls above our program's cut-off. The low overlap between Cassuto's results and the algorithm's is evidence that the two methodologies bring different perspectives and different strengths. Only a human being can explain the meaning of a *Leitwort* in context and weave it into a consistent tapestry with other methods of literary analysis. However, the ability to systematically evaluate every instance, and to apply objective criteria undiluted by personal bias, are core benefits of computerized *Leitwort* detection. The best use of such digital tools will be to allow merging of these two approaches by using algorithms before or after the human eye. Modern scholars of Biblical literature might use our program to systematically generate a list of repetitions to consider in their analyses, or consult its quantitative information (e.g. tf-idf scores) to consider whether their initial impressions might be distorted and in need of further scrutiny. Thus, objective methodology can become a thread woven into the subjective tapestry.

References

1917, *The Holy Scriptures According to the Masoretic Text: A New Translation*, The Jewish Publication Society of America, Philadelphia

Nevin Reda Al-Tahri, 2012. Textual Integrity and Coherence in the Qur'an: Repetition and Narrative Structure in Surat al-Baqara. Doctoral Thesis, University of Toronto.

Robert Alter, 1981. *The Art of Biblical Narrative*, Basic Books, New York, NY

Ronald D. Anderson, "Lietworter in Helaman and 3 Nephi," in The Book of Mormon: Helaman Through 3 Nephi 8, According to Thy Word, ed. Monte S. Nyman and Charles D. Tate, Jr. (Provo, Utah: Religious Studies Center, Brigham Young University, 1992) 241–250.

Martin Buber, 1927 lecture, Leitwort Style in Pentateuch Narrative.

Martin Buber and Franz Rosenzweig, 1936. *Die Schrift und ihre Verdeutschung*, Schocken, Germany.

Umberto Cassuto, 1964, *A Commentary on the Book of Genesis (Part II): from Noah to Abraham* (Israel Abrahams, Trans.). Magnes Press, Jerusalem, Israel

Umberto Cassuto, 1967, *A Commentary on the Book of Exodus* (Israel Abrahams, Trans.). Magnes Press, Jerusalem, Israel

Umberto Cassuto, *Peirush al Sefer Bereishit*, 1965. Jerusalem, Israel.

Umberto Cassuto, *Peirush al Sefer Shemot*, 1974. Jerusalem, Israel.

Umberto Cassuto, 1934, La creazione del mondo nella Genesi, *Annuario di studi ebraici*, vol. i, pp. 47–49

Umberto Cassuto, 1961, *A Commentary on the Book of Genesis (Part I): from Adam to Noah* (Israel Abrahams, Trans.). Magnes Press, Jerusalem, Israel

Yonatan Grossman, 2011. Literary Study of Biblical Narrative, Lecture 11, Leitwort, Part I, The Israel Koschitzky Virtual Beit Midrash, https://www.etzion.org.il/en/leitwort-part-i

Hans Peter Luhn, 1958. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, 159-165

William Hay MacNaghten, 1839-1842. *The Alif Laila or Book of the Thousand Nights and One Night*, 4 vols, W. Thacker and Co., Calcutta

David Pinault, 1986, Stylistic Features in Selected Tales From "The Thousand and One Nights", Doctoral Dissertation, University of Pennsylvania

David Pinault, 1992, *Story-Telling Techniques in the Arabian Nights. Studies in Arabic Literature*, volume xv. Brill, Leiden, NY

Dirk Roorda, 2015. The Hebrew Bible as Data: Laboratory - Sharing – Experiences, https://arxiv.org/pdf/1501.01866.pdf

John Wansbrough, 1978. *The Sectarian Milieu: Content and Composition of Islamic Salvation History*, Oxford University Press, London.

George Kingsley Zipf, 1945. The Repetition of Words, Time-Perspective, and Semantic Balance. The Journal of General Psychology, 32(1), 127–148.

From Copies to an Original: The Contribution of Statistical Methods

Amanda C. Murphy
Università Cattolica
amanda.murphy@unicatt.it

Felicita Mornata
Independent scholar
mornata.felicita@gmail.com

Raffaella Zardoni
Independent scholar
raffaella.zardoni@me.com

Abstract

English. Despite the almost infinite number of existing copies, the exact appearance of the medieval veronica – the sudarium kept in Rome imprinted with the face of Christ – is not known. This paper illustrates an attempt to find the 'true icon', creating a sort of identikit by means of the statistical processing of 4500 works, with analysis of the concentration of the copies and their characteristics, together with multivariate analysis tools.

Italiano. Nonostante le infinite copie esistenti non è noto quale fosse l'aspetto della veronica medievale, il sudario con impresso il volto di Cristo conservato a Roma. Il documento dimostra il tentativo di ritrovare la "vera icona" creando una sorta di identikit attraverso un'elaborazione statistica di 4500 opere con analisi della concentrazione delle copie e delle loro caratteristiche, unitamente a strumenti di analisi multivariata.

1 Introduction

1.1. The missing original veronica

According to tradition, the veronica, the medieval relic conserved in St Peter's, was the sudarium St Veronica offered Christ, on which his face was imprinted. Despite vast documentation from 1200 to 1500,[1] and the witness of pilgrims (such as Dante and Petrarch) visiting the relic in Rome, and countless extant copies, the exact aspect of the veronica is unknown.[2] There is great variance among its copies: e.g. Christ's face can be transfigured, or show signs of suffering, be with/without the crown of thorns, with open/closed eyes.

The first systematic study of copies of the relic in order to find the original was the work of Karl Pearson in 1887.[3] He compared literary and liturgical texts, and lined up images in chronological order, seeking reasons for the lack of continuity between representations. Scholars from literature, history, history of art, and theology have continued this research and the year 2000 saw a growth in interest in the topic;[4] there is, however, still no definitive answer to the question of what the original medieval veronica looked like.

1.2. Veronica Route project

The Veronica Route project[5] (VR) joined this field of research in 2010, with the creation of an open, expanding, interdisciplinary database of artistic and literary citations, ordering 'the infinite copies' of the image in an online catalogue. VR holds 4500+ tagged objects, particularly from the Middle Ages.[6] The classifications were carried out manually, following a traditional methodological approach.[7] The collected data was rendered

[1] In 1289 the Veronica was declared the most important relic in St Peter's (Coll. Bull. SS. Eccl. Vat., I, 214).
[2] The image kept in St Peter's Basilica shows indistinct markings and has not yet been studied. We can only be certain of the size of the medieval veronica (40 x 37 cm) thanks to the 14th century frame kept in the Vatican (Sturgis, 2000:75).
[3] It is a curious coincidence that Pearson worked out the statistical index of correlation, a starting point for several statistical methods.
[4] Belting (1990); G. Morello (1997); Hamburger (1998); Kessler, Wolf (1998); D'Onofrio (1999); Frugoni (1999); Morello, Wolf (2000); Di Blasio (2000); Burgio (2001); Di Fruscia (2013).
[5] Veronica Route was presented at the conference 'The European Fortune of the Roman Veronica', Magdalene College, University of Cambridge, April 2016.
[6] The works are signalled and sent in by volunteers together with the information found in loco: the sources are considered trustworthy, unless an error of attribution or dating is easily demonstrable. Many works recorded in Veronica Route do not yet have captions as complete as those in museum catalogues. Although the VR database covers all centuries, the richest and most relevant period for the present purpose is pre-1600.
[7] With time, it will be interesting to be able to adopt automatic face analysis tools, which are not yet appropriate for the recognition of iconographic characteristics, although they already work well in estimations of eye, nose and mouth positions, the degrees of different emotional expressions, etc. as shown in the Selfiecity project, classified as one of the most significant examples of "distant viewing".

available through tags marking iconographic characteristics, the dimensions of time (dating) and space (geographical positioning), with a visualisation function of the results allowing maps of veronicas to be manipulated.[8] Thus classified according to >50 features, the images were turned into information. Data mining with appropriate statistic methodologies on a statistically significant number of images has yielded the definition of significant models and new interpretative hypotheses/research paths about the relic which 'for three centuries, exerted such a great influence on the literary and artistic artefacts of our ancestors'.[9]

2 The variants of the iconographic subject and their geographical spread

To be able to identify the prototypes of the veronica, we used two different statistical tools, the index of transversality and multivariate analysis which juxtapose and aggregate the various tagged features.

2.1 Index of transversality

The recurrent iconographic characteristics, the subjects (St. Veronica, angels, sudarium, etc.), the various transversal themes (Roman relic, Strozzi, St. Spirit, etc.) and the supports (painting, miniature, sculpture), available for each veronica in the Veronica Route database, are all dichotomic variables (0-1 presence/absence).

The frequencies of the dichotomic variables to be compared through the centuries have to be normalised in order to make the data homogeneous. For this reason we identified the index of transversality in time, which was calculated as follows:

$$I_{KS} = \frac{\left(Freq_{KS}/Freq_{K\sum S}\right)}{\left(Freq_{S}/Freq_{\sum S}\right)} \qquad K = characteristic, \qquad S = century$$

Figure 1 shows that the graphic representation of some indices clearly show the strong coupling of DOUBLE-POINTED BEARD-OPEN EYES and the late appearance of the feature CLOSED EYES.

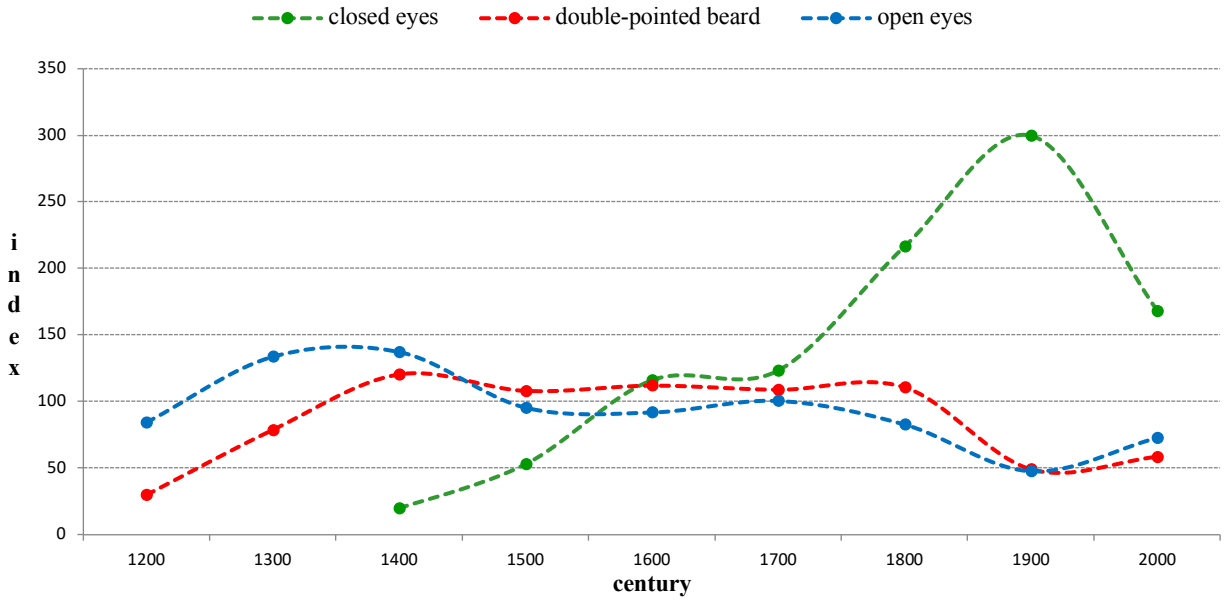


Figure 1. Index of transversality

[8] Software developed by Giacomo Aletti, professor of Probability and Statistical Mathematics, Dipartimento di Scienze e Politiche Ambientali, Università degli Studi di Milano.

[9] 'In this age when scholars attempt to trace the journey of a saga from India to Iceland, I hope that a description of the origin and development of a legend which has exercised such a huge influence on the literary and figurative works of our forebears over the centuries will not be considered superfluous.' (Karl Pearson, *Die Fronica*, p. 94, our translation)

2.2 Multivariate analysis: K-Means Cluster Analysis

From 1300 on, the considerable number of veronicas makes multivariate analyses on the available data meaningful. Another fruitful approach to investigating the correlations between the features is to aggregate veronicas from the same time frame into homogeneous groups, using the methodology K-means Cluster.[10] This algorithm considers each veronica like a point in a space of N dimensions (N = the available fields for each record). The value of each field is interpreted as distance from the origin along the corresponding spatial axis. In the K-means methods the original choice of a value for K determines the number of clusters that will be found. The analyst experiments with different values of K and each set of clusters is then evaluated: the one which shows the clearest interpretation of the data is chosen. It is an iterative process, which begins by identifying K points as seeds, and continues by aggregating all the other records, assigning each record to the closest centroid cluster. After this process the new cluster centroids are calculated, and, according to the proximity rule, the cluster to which each point belongs is recalculated. This iterative process ends when the cluster boundaries stabilize. Once the clusters have been defined, the results can be interpreted.

In order to describe the elements shared by the veronicas belonging to the same cluster, the matrix of the "final centroids" must be analysed, which, being dichotomous variables, quantifies the influence of each variable within the cluster. Therefore, Final Cluster Center ➔1 corresponds to the predominance of that variable in that particular cluster, and vice versa.

2.3 The fourteenth century

	1	2	3
without crown	0,937	0,922	0,897
open eyes	0,984	0,882	0,759
transfigured	0,921	0,745	0,655
St.Veronica	0,000	0,725	0,000
cruciform halo	0,556	0,667	0,379
double-pointed beard	0,444	0,353	0,448
head of Christ	0,032	0,196	0,000
cut out	0,492	0,039	0,138
sudarium	0,508	0,020	0,000
dark face	0,048	0,020	0,000
fleury cross	0,032	0,020	0,034
arma Christi	0,032	0,020	0,000
suffering	0,000	0,020	0,069
transparent veil	0,000	0,020	0,000
ascent to Calvary/Calvary	0,000	0,020	0,000
open mouth/visible teeth	0,063	0,000	0,069
dark veil	0,048	0,000	0,000
Sts.Peter and Paul	0,048	0,000	0,000
imago pietatis	0,048	0,000	0,000
green crown	0,016	0,000	0,000
crown of thorns	0,000	0,000	0,069
blank veil	0,000	0,000	0,000
triple veil	0,000	0,000	0,000
folds	0,000	0,000	0,000
monochrome	0,000	0,000	0,000
angel/s	0,000	0,000	0,862
Mass of Saint Gregory	0,000	0,000	0,000

[10] Implemented in the software SPSS Statistics (Analyze / Classify / K-Means Cluster procedure).

way of the Cross	0,000	0,000	0,000
Frequency of veronicas (143)	63	51	29

Figure 2. Clusters in the fourteenth century

The works in the 1300s tagged with ROMAN VERONICA[11] can be analysed in three clusters (excluding those with the tags BADGE and TEXT). In Figure 2, the predominant characteristics are in yellow, the absent ones in green. In this period all the veronicas are characterised, homogeneously, by the serene face of Christ, without a trace of suffering. Differences are found in the subjects showing the veronica: cluster 1 includes almost all the cases of the sudarium alone (with the tags CUT OUT and DARK FACE); cluster 2 aggregates the figure of St Veronica (particularly in Lombardy where we can find one of the first pictorial representations of the saint);[12] cluster 3 aggregates angels holding up the sudarium, positioned throughout eastern and central Europe.

2.4 The Fifteenth century

The 1400s are the most popular century for the image[13] (with 1122 works compared to 264 in the 1300s); in the second half of the century, veronicas appear with Christ's face bearing signs of suffering and drops of blood, and with the iconographic subject of the ascent to Calvary[14]. The considerable number of works and their variations in this century determined the choice to analyse the data in four clusters, in the interest of the best interpretation of the dominant characteristics in the variants. The 709 works tagged with ROMAN VERONICA, (excluding those tagged BADGE and TEXT) were thus aggregated: in Cluster 1 (234 veronicas, coloured in light blue in Figure 3) we find the transfigured face of Christ; in Cluster 2 (139 veronicas, dark blue) St Veronica; in Cluster 3 (166 veronicas, red) the new features of the Passion; and Cluster 4 (170 veronicas, orange) the sudarium.

Figure 3 displays the distribution of the works across Europe: we see a dominance of the figure of St Veronica with the transfigured face of Christ in France, where St Veronica is considered the evangeliser of the region,[15] and in Flanders (where she is the patron saint of linen and cloth merchants); the first veronicas with signs of suffering appear in all European countries, and it becomes slightly prevalent as a characteristic in Germany, while in England there is a prevalence of the sudarium as the iconographic subject.

[11] The tag ROMAN VERONICA is used when the veronica is the only subject or the main subject, and not when the veronica is part of another work, such as Arma Christi, Madonna of the Seven Sorrows, and such like.
[12] Stefano Candiani, The iconography of Veronica in the Lombardy region, late XIII-early XV centuries, in A. Murphy et al. Ed., The European Fortune of the Roman Veronica in the Middle Ages, Convivium Supplementum 2017, Turnhout: Brepols. p. 264.
[13] "From the 14th century, wherever the Roman Church went, the veronica would go with it" (Neil MacGregor and Erika Langmuir, 2000, p.92)
[14] The moment in which the veil was imprinted was not initially linked to the ascent to Calvary, but to Jesus' public life.
[15] Her tomb is still preserved at Saint Seurin (Bordeaux), on the pilgrims' way to Santiago di Compostela.

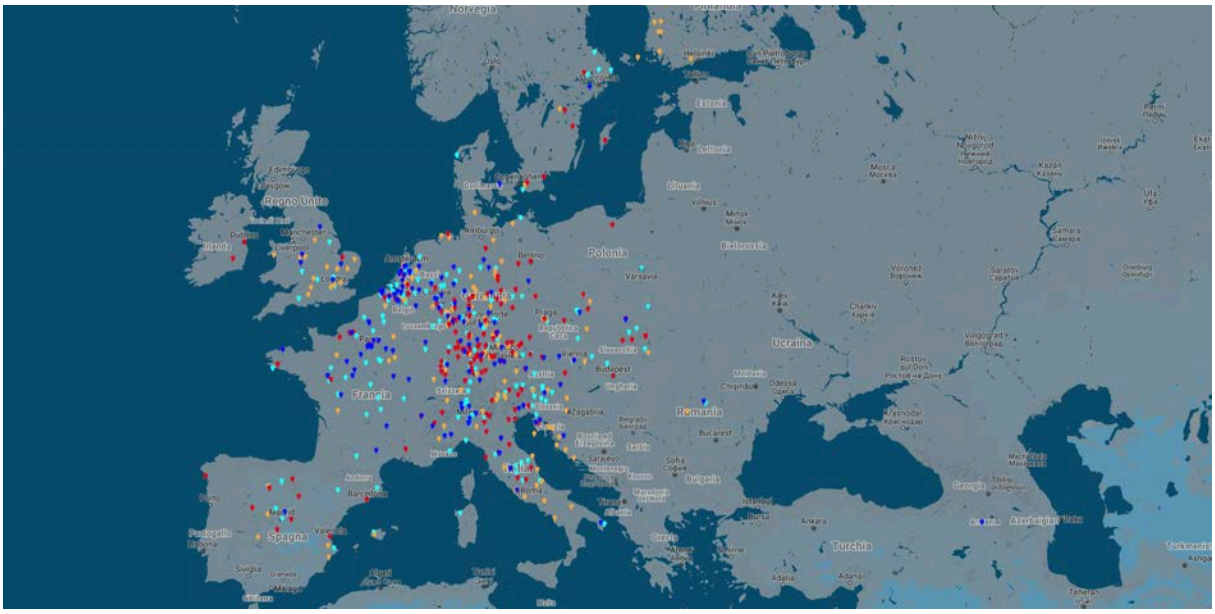


Figure 3. 1400s - distribution of the clusters

2.5 The Sixteenth century

There are 1075 works from the 1500s in Veronica Route. In a 4-cluster analysis of the 948 works, we find the SUDARIUM in cluster 1 (165 veronicas, including those characterised by DARK FACE and CUT OUT); cluster 2, SAINT VERONICA (always dominant in France), 295 veronicas; cluster 3, 281, SIGNS OF SUFFERING (a feature which becomes dominant in Italy); and in cluster 4, with 207 veronicas, the ASCENT TO CALVARY and BLANK VEIL. This feature, meaning that the face of Christ imprinted on the cloth is no longer visible, seems to shift attention away from the relic kept in Rome and onto the woman's pious gesture. In actual fact, the Protestant Reform and the Sack of Rome of 1527 interrupted – for various reasons – the history of the Roman relic.

3 Research on the Roman relic: validation

The last investigation concerns the relic kept in St Peter's, of which there are no photographic reproductions and which has never been an object of study.

In Veronica Route the tag ROMAN RELIC is assigned when the historical sources of the work refer directly to the relic. This feature is present only in 3.4% of the 1081 veronicas that are catalogued up to the end of the 1400s (excluding the veronicas with the tags TEXT and BADGE). These are decidedly small numbers for making a predictive analysis of the characteristics of the ROMAN RELIC.

The investigation therefore proceeded by evaluating the concentrations of the tag ROMAN RELIC in the other features and ordering them according to decreasing values. In the graph, the average concentration of 2% - index 100 – is indicated by the blue line and the characteristics with a higher index are those that correspond most to the ROMAN RELIC; the clear emergence of the features CUT OUT and DARK FACE can be seen, whereas the features DOUBLE-POINTED BEARD - CROWN OF THORNS - SUFFERING, positioned underneath the blue line are clearly separate from the Roman relic.

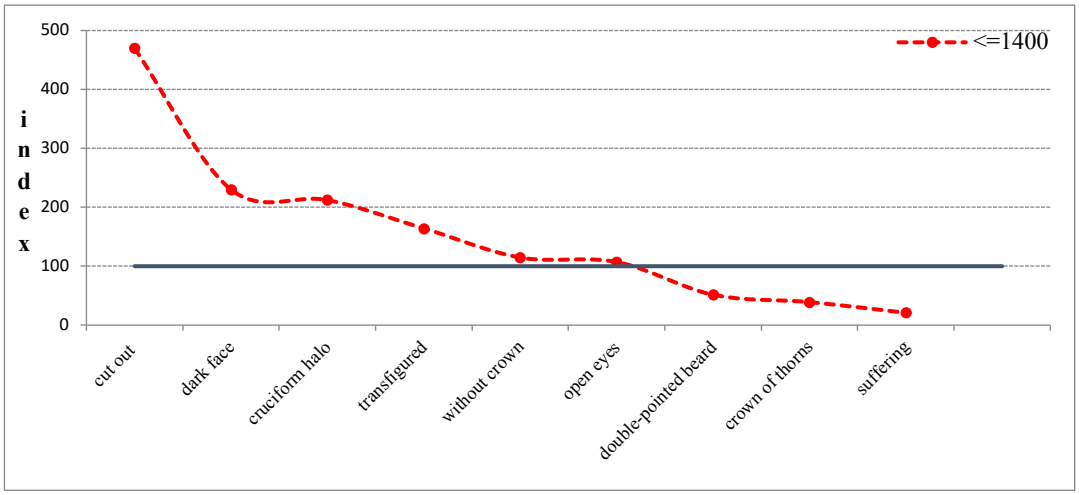


Figure 4. Concentration of the ROMAN RELIC

CUT OUT and DARK FACE are the features characterising the Mandylion in the Vatican (Figure 5),[16] a work likened to the medieval relic.[17] Identifying the Vatican Mandylion with the medieval Veronica would not be contradicted by the data, but the proportion of works tagged as CUT OUT and DARK FACE compared to the total works in the database (5 in 1200, 3 in 1300, 53 in 1400) and the late spread of the iconography, suggest the need for further research on this.

Figure 5. Left, the *Vatican Mandylion*, Lipsanoteca of the Pontifical Palaces, Vatican, next to works tagged CUT OUT and DARK FACE: *Veronica d'oro*, 1368 ca. Prague, Cathedral Treasury; *Santa Veronica col velo tra i SS. Pietro e Paolo*, 1430, altar Santa Maria del Monastero, Manta.

Conclusions

Firstly, the Veronica Route project intends to continue investigating the origins of the medieval relic's iconography. Secondly, we intend to break up the temporal arches (linked so far to centuries) so as to align them better with historical events, such as the Holy Years, in order to investigate the origins and development of the relic variations more precisely. Lastly, an exploration of the features which do not seem to derive from the Roman relic, but which have nevertheless become highly famous, would be an interesting new direction.

[16] The Mandylion was once considered the most ancient reproduction of the Face of Christ, even though it is documented in Rome only from 1517. Until 1870 it was kept in the Church of St Sylvester the First (San Silvestro in Capite), and is now kept in the Vatican.
[17] G. Morello (2012) p. 78.

References

Hans Belting. 1990. *Bild und Kult: eine Geschichte des Bildes vor dem Zeitalter der Kunst*, Beck, Munchen.

Michael J. A. Berry and Gordon S. Linoff. 1999. Mastering Data Mining, Wiley, Oxford.

Eugenio Burgio. 2001. Veronica e il volto di Cristo: Testi e immagini di una 'Legenda' tardomedioevale, in *Testo e immagine nel medioevo Germanico*, (Ed. Maria Grazia Saibene and Marina Buzzoni). Istituto Editoriale Universitario, Cisalpino, pp. 65-102.

Mario D'Onofrio (ed.). 1999. *Romei e Giubilei. Il pellegrinaggio medievale a San Pietro (350-1350)*, Electa, Milano.

Tiziana Maria Di Blasio. 2000. *Veronica, il Mistero del Volto*, Città Nuova, Roma.

Chiara Di Fruscia. 2014. Roma come Gerusalemme? Reliquie e memorie di Cristo nell'Urbe, in *Come a Gerusalemme. Evocazioni, riproduzioni, imitazioni dei luoghi santi tra Medioevo ed Età Moderna*, Benvenuti, A. and Piatti, P. Sismel, Firenze.

Arsenio Frugoni. 1999. *Pellegrini a Roma nel 1300, Cronache del primo Giubileo*, PIEMME, Casale Monferrato.

Giacomo Grimaldi. *Opusculum de sacrosancto veronicae sudario*, Biblioteca apostolica vaticana, Archivio Capitolo San Pietro H3, Città del Vaticano.

Jeffrey F. Hamburger. 1998. *The Visual and the Visionary*, ZoneBook, New York.

Herbert L. Kessler, Gerhard Wolf. (eds.) 1998. *The Holy Face and the Paradox of Representation, Papers from a Colloquium held at the Bibliotheca Hertziana*, Rome and the Villa Spelman, Florence.

Neil McGregor, Erika Langmuir. 2000. *Seeing Salvation. Images of Christ in Art*, Yale University Press, Newhaven.

Giovanni Morello. 2012. "Or fu sì fatta la sembianza vostra?" La Veronica di San Pietro: storia ed immagine, in Giovanni Morello. *La Basilica di San Pietro*, Gangemi Editore, Roma, pp. 39-80.

Giovanni Morello. 1997. "La Veronica nostra". In *La Storia dei Giubilei*, edited by Gloria Fossi, Giunti - Bnl Edizioni, Roma.

Giovanni Morello, Gerhard Wolf. 2000. *Il volto di Cristo, Catalogue of the Exhibition*. Electa, Milano.

Karl Pearson. 1887. *Die Fronica, Ein Beitrag zur Geschichte des Christusbildes im Mittelalter*, Strasbourg.

Alexander Sturgis. 2000. "The True Likeness", in Gabriele Finaldi (ed.), *Image of Christ*, The National Gallery, London.

Nicola Barbuti, Stefano Ferilli, Tommaso Caldarola. 2018. *Un innovativo Graphing Matching System per la ricerca in database di manoscritti antichi*. *Umanistica Digitale*, No 3. https://umanisticadigitale.unibo.it/article/view/8144

Alise Tifentale, Lev Manovich. 2015. *Selfiecity: Exploring Photography and Self-Fashioning in Social Media*. In Berry D.M., Dieter M. (eds.) *Postdigital Aesthetics*. Palgrave Macmillan, London.

FORMAL. Mapping Fountains over Time and Place. Mappare il movimento delle fontane monumentali nel tempo e nello spazio attraverso la geovisualizzazione

Pamela Palomba
Università degli Studi Suor Orsola Benincasa - Napoli
palomba.pamela@gmail.com

Roberto Montanari
Università degli Studi Suor Orsola Benincasa - Napoli
roberto.montanari@unisob.na.it

Emanuele Garzia
Università degli Studi Suor Orsola Benincasa – Napoli
garziaems@gmail.com

Abstract

English. Naples historical fountains are today testimonials of its huge and centenarian changes and urban transformations. Recording their distribution and movements in time and space, this project has the extent to analyze their spatial and diachronic contextualization in a relational mode using 'nodegoat', a web-based data management, analysis and visualisation environment, in order to get a deep understanding of the phenomenon both for researcher and for the public interested. Mapping in fact is here used as a form of spatial and data storytelling.

Italiano. La città con la sua evoluzione architettonica raccoglie un vasto insieme di informazioni che collegano diversi campi di ricerca: digital humanities, spatial humanities e beni culturali. Il progetto si propone di tracciare la forma della distribuzione dell'acqua pubblica nella città di Napoli nei secoli e, in particolare, di mappare il movimento delle fontane monumentali nel tempo e nello spazio facendo uso di un ambiente web-based di data management, network analysis & visualisation come Nodegoat sviluppato da Lab1100. Le fontane monumentali di Napoli sono state condizionate nel corso della loro storia secolare da numerose vicissitudini spaziali. L'avvicendarsi delle dominazioni e dei governatorati, specie nel periodo del Viceregno spagnolo (XVI - inizio XVIII sec.), ha comportato svariati cambiamenti all'arredo urbano che, nel caso delle fontane pubbliche, ha causato spesso una loro rimozione dal luogo originario per uno spostamento in contesti diversi. Attraverso l'impiego congiunto delle fonti storiografiche e cartografiche, del DBMS (database management system) e della GIS Science, lo studio si propone di creare una narrazione spaziale che evidenzi l'efficacia, sia per il ricercatore che per il fruitore, delle potenzialità di spatial storytelling e data storytelling per il recupero della memoria storica dei luoghi.

1 Introduzione

L'esistenza umana si organizza e pensa se stessa anche in termini spaziali, è nel suo rapporto con lo spazio che l'individuo struttura la sua esperienza, in maniera situata (Merleau-Ponty, 1965), tanto che si può a buon diritto affermare che l'esistenza è spaziale e lo spazio è esistenziale. Lo spazio in tal modo, inteso in senso antropologico, cessa di essere solo un'astrazione geometrica e definisce il campo di studi oggetto delle *spatial humanities*. L'interesse crescente per lo spazio negli studi storiografici è uno degli aspetti della svolta culturalista (Torre, 2008) dell'ultimo quarto del XX sec., che, con l'abbandono del concetto di spazio assoluto del sistema cartesiano in favore di un concetto di spazio relativo, ha comportato una svolta che viene definita *spatial turn* ossia il passaggio a un sistema tolemaico che distingue tra spazio assoluto e relativo, tra geografia e corografia (Cosgrove, 2003). La prima, con le sue capacità descrittive basate sulla matematica e sul rilievo scientifico, e la seconda, imperniata su un'impostazione di tipo più visuale e letterario, ma in grado di connettere però la dimensione storica a quella geografica. In sostanza si tratta di una svolta verso un'interpretazione simbolica del paesaggio che tenga conto della sua doppia anima di spazio naturale e culturale: quella che si riferisce a processi naturali e sociali, e quella che corrisponde alle conseguenze delle azioni umane che lo trasformano. Lo spazio in questo senso da entità geometrica astratta si trasforma in luogo, ossia l'espressione peculiare di uno spazio geografico.

La 'peculiare' natura del luogo lo definisce come l'ordine secondo il quale diversi elementi vengono distribuiti entro rapporti di coesistenza, gli uni affianco agli altri; un luogo è dunque una configurazione di posizioni, una indicazione di stabilità. Al contrario lo spazio è un incrocio di entità mobili, è in qualche modo il prodotto dell'insieme dei movimenti che si verificano al suo interno e lo animano, orientandolo, rendendolo contingente tanto da tramutarlo in unità polivalente di programmi conflittuali o di prossimità contrattuali (de Certeau, 2001). Lo spazio non è il semplice scenario dell'azione storica, ma un prodotto significativo e determinante di cambiamento. Ne deriva che un movimento produce uno spazio e ne traccia la storia. Ogni narrazione che si rispetti si struttura lungo una sequenza di eventi producendo quella che può essere chiamata 'la forma del tempo' e privilegiando così la componente cronologica, ma tutte le narrazioni implicano un mondo di estensione spaziale. Alcuni teorici infatti riconoscono un più che evidente collegamento tra spazio e tempo, come si evince dalla concezione di cronotopo di Mikhail Bakhtin (1981) intesa come «l'intrinseca connessione di relazioni spaziali e temporali», con il tempo che fornisce la quarta dimensione dello spazio. In questa prospettiva la narrazione è immaginabile come «la rappresentazione del movimento all'interno delle coordinate di spazio e tempo», con gli eventi marcati dall'intersezione di assi orizzontali e verticali in un intreccio dinamico tra superficie e profondità (Bakhtin, 1981).

Sulla base di queste premesse, con il presente lavoro di ricerca si è inteso rilevare nella catalogazione e visualizzazione della distribuzione e del movimento delle fontane storiche di Napoli le condizioni narratologiche per produrre una narrazione spaziale attraverso la geo-visualizzazione, un ramo della *Gis Science* che sviluppa tecniche e strumenti disegnati per rendere visuali dei fenomeni spaziali (Craine and Aitken, 2009).

L'operazione di traduzione dei concetti di spazio e luogo in quelli rispettivamente di mappa e itinerario, intesi come linguaggi simbolici e antropologici dello spazio e come i due poli dell'esperienza (de Certeau, 2001), ha condotto alla messa in scena della loro interazione, attraverso l'impiego di una piattaforma di *database management system* come *nodegoat*[1], in grado di produrre uno scenario in cui i luoghi figurano come 'OGGETTI' della mappa, e gli itinerari come tracciati degli spostamenti delle fontane nel tempo e nello spazio.

L'operazione effettuata consente così di arrivare a una più profonda comprensione dei contesti spaziali in cui i beni culturali sono inseriti come testimonianze vive e multilivello, capaci di raccontare una storia stratificata sia in senso cronologico che spaziale, in cui possano essere rilevabili quindi come parti in relazione complessa con i luoghi ai quali conferiscono valore di civiltà.

Per comprendere e diffondere queste informazioni con contenuti multidisciplinari il *data storytelling* si rivela uno strumento efficace nella divulgazione di fenomeni complessi. Per trasformare questi dati in informazioni utili alla conoscenza è necessaria la visualizzazione così come il racconto attraverso una storia.

Alcuni strumenti come penne, mappe e calcolatrici possono essere considerati artefatti cognitivi che migliorano la nostra conoscenza, amplificando i processi cognitivi coinvolti nell'interazione con le rappresentazioni tipiche del mondo esterno. Nell'*Information Visualisation* un ruolo fondamentale è dato dal *data storytelling*. Gli esseri umani hanno sempre utilizzato le storie per trasmettere informazioni, valori culturali ed esperienze attraverso mezzi tecnologici che si sono evoluti nel tempo (si pensi alla scrittura, alla stampa e oggi ai computer). Una buona narrazione trasmette una mole di informazioni in un formato facilmente assimilabile dal fruitore o dallo spettatore (Segel and Heer, 2010). La visualizzazione delle informazioni attraverso i dati, combinata con un adeguato storytelling, permette di produrre rappresentazioni visive molteplici. Ogni visualizzazione può essere utilizzata per raccontare una storia e le diverse modalità sono funzionali ai differenti tipi di storia.

Proprio per queste loro caratteristiche, le tecniche di visualizzazione dimostrano di essere un valido strumento nelle mani dello studioso interessato a valorizzare il patrimonio informativo collegato ai rapporti fra la città, la sua evoluzione e i beni culturali.

[1] Pim van Bree and Geert Kessels. 2013. nodegoat: a web-based data management, network analysis & visualisation environment, http://nodegoat.net from LAB1100, http://lab1100.com

2 Caso d'uso: FORMAL, Mapping fountains over time and place.

Il nome del progetto, Formal, oltre ad essere un acronimo delle parole che lo specificano (Mapping Fountains OveR tiMe And pLace), si ispira al termine 'formale' che indica uno dei segmenti di cui si compone la rete degli acquedotti ed è dunque anche usato per definire un canale principale di alimentazione delle fontane pubbliche. Testimoni di eccezione delle trasformazioni politiche, sociali e urbanistiche della città, le fontane pubbliche di Napoli hanno una storia complessa, fatta di traslazioni, mutilazioni, cambiamenti che si è tentato di restituire in forma semplessa grazie all'impiego di *nodegoat*, un DBMS *web-based* in grado di processare, analizzare e visualizzare dataset complessi in modalità relazionale, diacronica e spaziale.

La prima fase del progetto è consistita nel reperimento di tutti i dati della ricerca, di tipo bibliografico, cartografico e iconografico, necessari alla compilazione del database. Successivamente si è passati alla costruzione del modello dei dati (*data modelling*) , dapprima a livello concettuale attraverso la creazione del dataset e sulla base delle esigenze della domanda di ricerca e, in seguito, a livello logico tramite l'inserimento dei dati nelle schede di struttura approntate. La piattaforma *nodegoat* ha consentito la creazione di un database completamente personalizzato da parte di un utente non esperto e perfettamente rispondente alle esigenze di ricerca. Il *data modelling* nell'ambito umanistico è largamente percepito come un processo epistemologico, piuttosto che come un processo ontologico. L'interfaccia dell'applicazione del database può far nascere infatti nuove opportunità o creare sfide ulteriori.

La prima operazione necessaria è stata la definizione dei *type* principali di consultazione; il *type* principale consente di leggere nella scheda di ciascun oggetto (*object*) a esso riferito tutte le informazioni anagrafiche, geografiche e storiche (Figura 1). Sono state inserite 28 fontane storiche pubbliche, che coprono un arco cronologico che va dal XVI sec. al XXI sec., 14 di esse hanno subito almeno uno spostamento, con il caso eclatante della fontana del Nettuno che ha avuto ben otto trasferimenti (cfr. Figura 3).

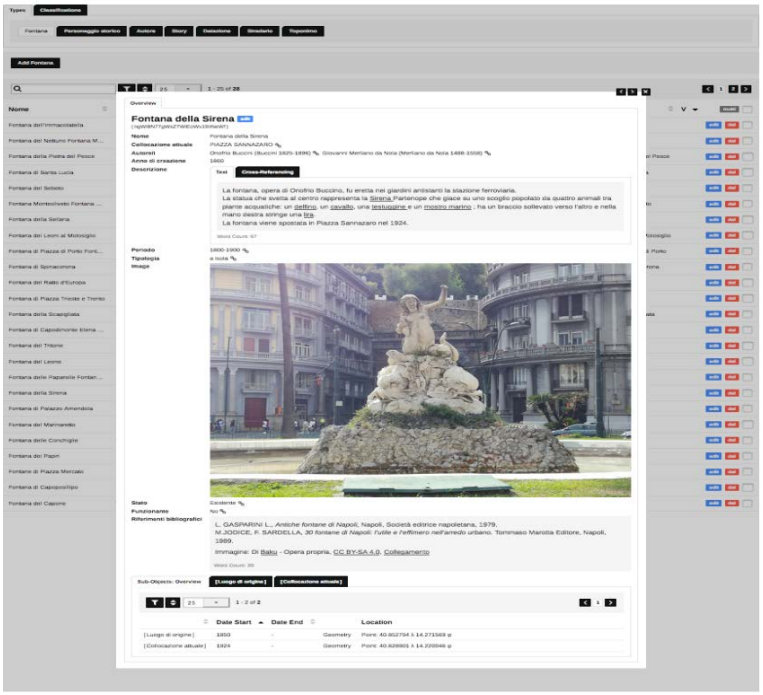


Figura 1. Scheda dell'*object* Fontana della Sirena

Gli obiettivi sono stati essenzialmente due: da un lato catalogare e ordinare il materiale di studio raccolto e sistematizzarlo attraverso la produzione di schede anagrafiche e multimediali che descrivessero le caratteristiche di ciascuna fontana (*object*); dall'altro ottenere per ciascun *object* la visualizzazione geografica della sua posizione nello spazio e nel tempo con il tracciamento dello spostamento da un luogo all'altro nelle

diverse epoche storiche (*sub-object*). Quest'ultima operazione è consistita in particolare nell'approntare l'apparato di coordinate leggibili attraverso la '*Geographical Visualisation*' (Figura 2).

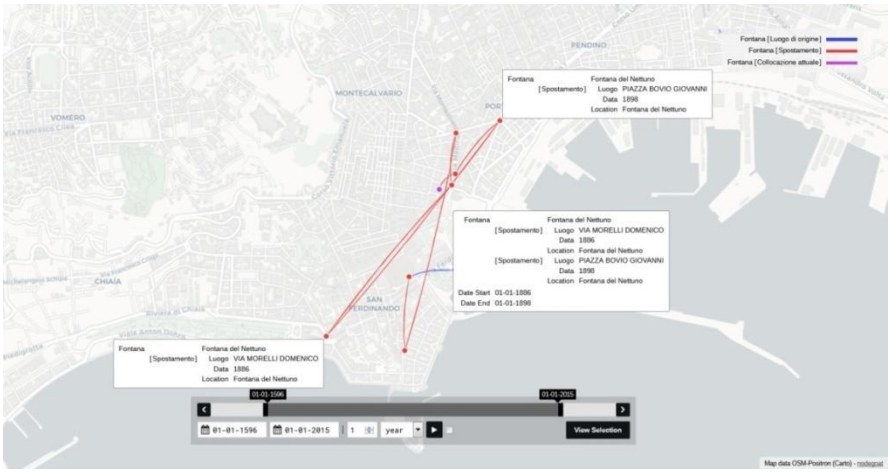


Figura 2. *Geographical visualisation* degli spostamenti della Fontana Medina

Per quanto riguarda la collocazione si è scelto di visualizzare sulla mappa, attraverso l'uso di punti e linee, la posizione e il percorso compiuto dalla fontana in caso di diverse collocazioni, usando il sistema di coordinate geografiche sia della collocazione originaria che delle collocazioni successive per georeferenziare il punto di interesse (*point of interest*). Nel caso di luogo non più esistente si è fatto ricorso alla cartografia storica per identificare il punto corrispondente da georeferenziare. Le coordinate geografiche del punto di interesse, così come le corrispondenti datazioni di collocazione e/o spostamento, sono inserite usando la sezione *sub-object* della relativa scheda dell'*object* di ciascuna fontana (Figura 3).

Sub-Objects: Overview | [Luogo di origine] | [Spostamento] | [Collocazione attuale]

25 1 - 9 of 9 1

	Date Start	Date End		Location
[Luogo di origine]	1596	-	Geometry	Point: 40.835831 λ 14.252375 φ
[Spostamento]	1629	-	Geometry	Point: 40.835494 λ 14.249188 φ
[Spostamento]	1634	-	Geometry	Point: 40.831622 λ 14.248883 φ
[Spostamento]	1639	-	Geometry	Point: 40.843006 λ 14.252458 φ
[Spostamento]	1659	-	Geometry	Point: 40.840303 λ 14.25216 φ
[Spostamento]	1886	-	Geometry	Point: 40.832359 λ 14.243439 φ
[Spostamento]	1898	-	Geometry	Point: 40.843666 λ 14.255492 φ
[Spostamento]	2000	-	Geometry	Point: 40.840885 λ 14.252425 φ
[Collocazione attuale]	2015	-	Geometry	Point: 40.840072 λ 14.251289 φ

Figura 3. Sezione *sub-object* con gli spostamenti della Fontana del Nettuno

Ciascun *type* è collegato in maniera incrociata (*cross-referencing*) agli altri *type* rilevanti, consentendo così l'esplorazione di una serie di relazioni (*network analysis*) tra i dati che ha rivelato la sua efficacia dal punto di vista della ricerca, come da quello del possibile impiego dei dati stessi per la costruzione di narrazioni. Nel primo caso infatti è stato possibile analizzare il dataset in base alle diverse domande di ricerca, ad esempio attraverso il filtro del personaggio storico (committente, artista), piuttosto che per cronologia, per elemento decorativo ricorrente o per toponimo.

Relativamente invece alla potenzialità narratologica è stato previsto il type '*Story*' che contiene una selezione di brani tratti dalla letteratura periegetica del XVII sec., nello specifico le *Notitie del bello, dell'antico e del curioso della città di Napoli per i signori forastieri date dal canonico Carlo Celano napoletano, divise in dieci giornate* di Carlo Celano, che ci informano sulla conformazione urbanistica della città nonché sulla descrizione delle fontane pubbliche della Napoli del 1692. Impiegando la *network analysis* per indagare il campo *Story* è possibile avere un quadro chiaro e immediato di tutte le fontane esaminate dal cronista nel

corso della specifica *giornata* e trarne considerazioni spazialmente orientate in merito alla fonte impiegata. A mero titolo di esempio si può rilevare che grazie alla *network analysis* è stato possibile ragionare sugli spostamenti programmati dal cronista per la stesura del testo letterario della singola *giornata*: attraverso la visualizzazione dei dati collegati in maniera relazionale appare chiaro il percorso spaziale che struttura la narrazione e nello stesso tempo pianifica e suggerisce itinerari al destinatario (quante e quali fontane sono descritte da Celano nella Giornata V con schema delle relazioni topografiche e topologiche che le legano, cfr. Figura 4).

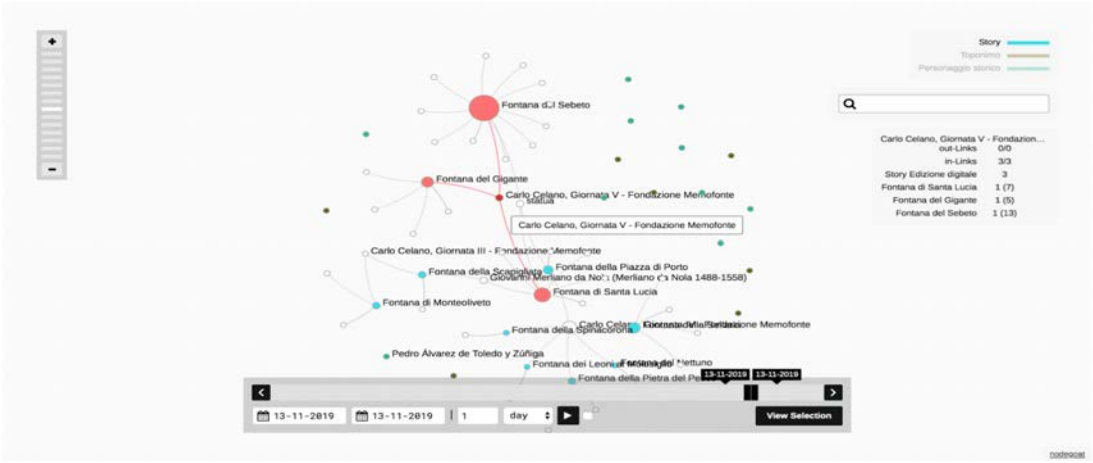


Figura 4. *Network analysis* del campo *Story*

Appare chiaro che un tale uso delle fonti spazialmente strutturato, relazionale e incrociato rende in questo caso le fontane un possibile espediente, dal punto di vista narrativo storicamente delineato, per costruire storie che si avvalgano dei documenti, digitali e non, per generare contesti narrativi scientificamente validi.

Nell'ottica del riuso creativo delle fonti digitalizzate e *open access* si è fatto ricorso ad esempio all'edizione digitale curata dalla Fondazione Memofonte alla quale si rimanda, all'interno della singola scheda, in maniera contestuale al luogo del testo. Allo stesso modo si è fatto ricorso allo stradario ufficiale del Comune di Napoli integrando nel DBMS il dataset con licenza IODL (*Italian Open Data License*) per effettuare il collegamento georeferenziato con l'attuale mappa cittadina.

Infine si è prestata particolare attenzione alla scelta delle opzioni di visualizzazione dello '*Scenario*' di geovisualizzazione allo scopo di creare un *data storytelling* efficace dal punto di vista della fruizione per la futura pubblicazione su web. Sulla base della modellizzazione proposta da Segel e Heer (Segel and Heer, 2010) si è optato per un approccio che prevede una posizione di controllo dei contenuti erogati (*Author-driven approach*), ideale per lo *storytelling* e la comunicazione di contenuti educativi, temperandolo con uno di tipo più interattivo (*Reader-driven approach*) che consentirà all'utente di esplorare lo scenario, interrogandolo in base a diverse chiavi di ricerca con la possibilità di produrre forme più complesse e personalizzate di analisi.

3 Sviluppi futuri

La fase successiva del progetto prevede l'implementazione di un'interfaccia pubblica per la fruizione web con la produzione di scenari dedicati in base alla fascia di utenza e l'approfondimento dell'indagine tramite la *network analysis* e il *data storytelling* per la scelta di una narrazione da visualizzare, valorizzare e divulgare. Sarà inoltre analizzato e geovisualizzato un secondo dataset relativo alle fontane scomparse e alla distribuzione delle acque affioranti cittadine, che in parte le alimentavano, rintracciate attraverso le fonti storiche.

4 Conclusioni

Ciò che di interessante emerge dall'interazione tra la mappa e l'itinerario è una dimensione spaziale narrativa in cui la mappa giustifica l'itinerario e l'itinerario definisce la mappa come spazio geografico e culturale insieme, non dunque solo una carta geografica, ma anche un libro di storia. Abbiamo infatti da un lato la mappa, che ha una funzione topica ossia di definitore di luoghi e, dall'altro, un racconto fatto di

spostamenti ossia topologico, relativo alla deformazione delle figure. La metodologia usata aiuta a preservare e presentare le complessità che sono insite nelle fonti storiche e nel loro impiego incrociato attraverso il *deep mapping*. La visualizzazione e il *data storytelling* invece si configura come un'interfaccia interattiva utile per la ricerca nelle *digital humanities*, in particolare nel *cultural heritage* per la storia della città e delle sue trasformazioni, con risvolti interessanti anche per un impiego a servizio della fruizione dei beni culturali e delle imprese creative.
Lo strumento scelto ci consente di esplorare contemporaneamente le due dimensioni, topica e topologica, in senso sincronico (es. analizzare e visualizzare quali e quante fontane nello stesso secolo) e diacronico (es. tappe del loro percorso all'interno della città nel corso del tempo). Tra queste due determinazioni vi sono dei passaggi che portano alla conclusione che i racconti spaziali effettuano un lavoro che trasforma i luoghi in spazi o gli spazi in luoghi, con un'azione creativa e performativa, delimitando con le loro attività su di essi una scena da narrare che ancora vive.

Bibliografia

Mikhail M. Bakhtin. 1981. *The Dialogic Immagination: four essays*, ed. Michael Holquist, trad. Caryl Emerson and Michael Holquist. University of Texas Press, Austen: 278.

David J. Bodenhamer, John Corrigan and Trevor M. Harris (Eds.). (2015). *Deep Maps and Spatial Narratives*. Bloomington; Indianapolis: Indiana University Press.

Carlo Celano. 1692. *Notitie del bello, dell'antico e del curioso della città di Napoli per i signori forastieri date dal canonico Carlo Celano napoletano, divise in dieci giornate*. Edizione digitale Fondazione Memofonte: https://www.memofonte.it/ricerche/napoli/#carlo-celano, Napoli.

Michel de Certeau. 2001. *L'invenzione del quotidiano.* Lavoro, Roma.

Denis E. Cosgrove. 2003. *Landscape and the European Sense of Sight: Eyeing Nature*. In: K. Anderson et al. eds., *Handbook of Cultural Geography*. Sage, London: 249–68.

James Craine and Stuart C. Aitken. 2009. *The emotional life of maps and other visual Geographies.* In: Martin Dodge, Rob Kitchin and Chris Perkins, eds., *Rethinking maps: New Frontiers in Cartographic Theory*. Routledge, New York: 168-185.

Maurice Merleau-Ponty. 1965. *Fenomenologia della percezione*. Il Saggiatore, Milano.

Edward Segel and Jeffrey Heer. 2010. *Narrative visualisation: Telling stories with data, Visualisation and Computer Graphics*, IEEE Transactions on, vol. 16, no. 6: 1139–1148.

Angelo Torre. 2008. Un « tournant spatial » en histoire : Paysages, regards, ressources. *Annales. Histoire, Sciences Sociales*, 63rd,(5): 1127-1144. https://www.cairn.info/revue-annales-2008-5-page-1127.html

Pim van Bree and Geert Kessels. 2013. nodegoat: a web-based data management, network analysis & visualisation environment, http://nodegoat.net from LAB1100, http://lab1100.com

Pim van Bree and Geert Kessels. 2013. Nodegoat documentation: https://nodegoat.gitbooks.io/documentation/content/, GitBook.

Pim van Bree and Geert Kessels. 2015. Mapping Memory Landscapes in nodegoat. In: Aiello L., McFarland D. (eds) Social Informatics. SocInfo 2014. Lecture Notes in Computer Science, vol 8852. Springer, Cham. https://doi.org/10.1007/978-3-319-15168-7_34

Matthew O. Ward, George Grinstein, Daniel Keim. 2010. *Interactive Data Visualisations Foundations, Techniques, and Applications*. A K Peters, Ltd. Natick, Massachusetts.

Paul is Dead? Differences and Similarities before and after Paul McCartney's Supposed Death. Stylometric Analysis on Transcribed Interviews

Antonio Pascucci, Raffaele Manna, Johanna Monti
L'Orientale University of Naples
UNIOR NLP Research Group
Naples, Italy
[apascucci,rmanna,jmonti]@unior.it

Vincenzo Masucci
Expert System Corp. Naples, Italy
vmasucci@expertsystem.com

Abstract

English. In this paper, we show the results of a stylometric analysis conducted on Paul McCartney's interview transcriptions using three different approaches in order to detect differences and similarities in his speeches before and after 9th November 1966, the date of his supposed death. Our research is based on the *Let IT Corpus*, a corpus of Paul McCartney's Interview Transcriptions. The corpus is a collection of texts from the *Beatles Interviews Database*, a repertoire of one-hundred sixty-three Beatles' interviews freely available on the Web (http://www.beatlesinterviews.org/), and from other interviews available on YouTube.

Italiano. In questo contributo mostriamo i risultati di un'analisi stilometrica con tre ap-procci differenti operata sulle trascrizioni delle interviste di Paul McCartney con lo scopo di individuare analogie e differenze stilistiche nelle trascrizioni delle interviste fatte prima e dopo il 9 novembre 1966, data della sua presunta scomparsa. La ricerca si basa sul *Let IT Corpus*, un corpus composto da trascrizioni di interviste fatte a Paul McCartney che abbiamo costruito con le trascrizioni delle interviste presenti sul *Beatles Interviews Database*, una raccolta di centosessantatré interviste ai Beatles disponibile su internet (http://www.beatlesinterviews.org/) e di altre interviste disponibili su YouTube.

1 Introduction

Paul McCartney's supposed death (dated 9th November 1966 because of a car accident) represents a legend which does not belong to the music business only but embraces other worlds, both for Paul McCartney's fame given by Beatles' everlasting success, and because of many stories born around this episode. Paul is dead (PID) theory represents one of the most controversial legends in the history of music, enough to be still debated after more than half a century, during which numerous stories are born, some feeding and some other damping its truth. In this paper, we show the results of a stylometric analysis conducted on Paul McCartney's interview transcriptions using three different approaches in order to detect differences and similarities in his speeches before and after 9th November 1966. Our research is based on:

- the Beatles Interviews Database (http://www.beatlesinterviews.org/), a collection of one-hundred sixty-three interviews from 1962 to 1984;
- YouTube subtitles, which we manually corrected, if necessary, listening to the audio.

To the best of our knowledge, this research represents one of the very first stylometric analyses on interview transcriptions. In this paper, we also present the *Let IT Corpus* (Paul McCartney's Interview Transcriptions), composed of fifty-two documents concerning interviews before 9th November 1966 and fifty-two documents concerning interviews after 9th November 1966. *Let IT Corpus* is still in its embryonic stage: we foresee to expand it with further texts so that it can be used for more accurate analyses in the near future.

The strongest supporters of PID theory claim that immediately after his death, Paul McCartney has been replaced by a lookalike. There are several theories that even today sustain the veracity of PID, many of which spread by the Beatles themselves, who sometimes enjoyed including subliminal messages in their songs. For example the celebrated John Lennon's whisper in the song *I'm so tired*, if listened backwards, seems to say *Paul is dead, man: miss him! miss him! miss him!*. The Abbey Road's album cover also shows at least ten references to Paul McCartney's death. On the other hand, some theories remove all doubts, claiming that Paul McCartney never died and all PID hypotheses are nothing but a business choice, which has contributed to add extra charm to Beatles' success. In October 1969, the Beatles' press office categorically denied PID rumours, labelling them as *a load of old rubbish*. PID theory has been investigated in literature (Cartocci, 2005) and in automatic-recognition (Holland et al., 2014). The present contribution is organized as follows: in Section 2 we show Related Work. The *Let IT Corpus* is described in Section 3, in Section 4 we describe three different approaches adopted in the analysis and their results. In Section 5 the stylistic differences and similarities detected by the linguistic analysis are thoroughly discussed. Conclusions are in Section 6. In Section 7 we introduce Future Work.

2 Related Work

CS is the statistical analysis of writing style (Zheng et al., 2006) and it is used to identify or profile the author of a text. The main assumption of Authorship Attribution (AA) is that each author operates choices which are influenced by sociological (age, gender and education level) and psychological (personality, mental health and being a native speaker or not) factors (Daelemans, 2013) which determine a unique writing style. In AA some studies are being conducted on speech transcriptions. In 2014 (Herz and Bellaachia, 2014) investigated the authorship of Barack Obama's speechwriters on a corpus composed by thirty-seven speech transcriptions. They based their research on the supposition that Barack Obama has four principal speechwriters and deal with the AA of Barack Obama's speeches with four different approaches, that reached different results, but still showing that CS can be used to differentiate authors who write in a similar style. (Airoldi et al., 2006) conducted a similar research on Ronald Regan's radio speeches. The corpus they used for their investigation is composed of a thousand thirty-two radio addresses delivered by Ronald Reagan between 1975 and 1979. The scholars focused the experiment on three-hundred twelve radio addresses for which no direct AA evidence is available, and they concluded that in 1975, Ronald Reagan drafted seventy-seven speeches and his collaborators drafted seventy-one, whereas over the years 1976-1979, Ronald Reagan drafted ninety speeches and his collaborator Peter Hannaford drafted seventy-four speeches. The study of (Herz and Bellaachia, 2014) and that of (Airoldi et al., 2006) share a problem: it is not possible to know the accuracy of the AA results of their study.

CS is also useful in studying changes in the style of an author over time. As argued by (Rybicki, 2015) time is one of the most significant factors for the evolution of the literary lexicon. With this in mind, some researches are conducted on stylochronometry (for a survey, see (Stamou, 2007)), namely the study of the change of style correlated to the passing of time. (Forsyth, 1999) differentiates the style of the poet William Butler Yeats between *younger Yeats* and *older Yeats*, devising along the way a measurement he calls a *youthful Yeatsian Index*. (Van Hulle and Kestemont, 2016) use sylometry to periodize Samuel Beckett's works, finding stylistically innovative change in his late style. Lastly, the findings of (Evans, 2018) show that the dramatic style of Aphra Behn over the course of her 20-year career, can be divided in three different phases. Obviously we must keep in mind that our analysis is based on transcription of speeches, and therefore not on written texts. Until now, to the best of our knowledge, no stylistic research analysis has been carried out to detect differences and similarities in interview transcriptions before and after Paul McCartney's supposed death.

3 *Let IT Corpus*

For our research we investigated the *Beatles Interviews Database*[1], a collection of one-hundred sixty-three transcription of Beatles' interviews from 1962 to 1984 created in 1997 by Jay Spangler and now managed by Jude Southerland Kessler and Suzie Duchateau. The website also contains a songwriting and

[1] http://www.beatlesinterviews.org/

recording database, a collection of Beatles' movies, quotes and pictures. We also investigated thirty-five Beatles' interviews available on YouTube: in this case we analyzed the automatic captions generated by speech recognition, and we corrected texts if necessary. In each interview, we isolated Paul McCartney's speeches and we created a document for each interview. The *Let IT Corpus* is a very small balanced corpus composed of one-hundred four documents belonging to two different classes: I) *before* (composed by fifty-two documents concerning interviews before 9th November 1966) and II) *after* (composed by fifty-two documents concerning interviews after 9th November 1966). A few texts belonging to the *after* class found on YouTube date after 2000. The majority of texts of the *Let IT Corpus* are from the *Beatles Interviews Database* (32 *before* texts and 25 *after* texts, including a few chunks). The corpus contains also texts from the *Beatles Interviews Database* concerning interviews involving the whole Beatles group, from which we isolated Paul McCartney's speeches. The remaining part of *Let IT Corpus* consists of Beatles' interviews freely available on YouTube. *Let IT Corpus* is still in its embryonic stage, since it is composed of approximatively one-hundred texts and it represents the first step in this field. Further work will be carried out as soon as *Let IT Corpus* will be expanded.

4 Our three approaches to stylometric analysis

We investigated this AA issue with three different approaches, in order to compare the results. For all the experiments we removed punctuation and symbols, and we lowercased all characters.

4.1 Hybrid approach

In this section we describe the first approach to stylometric analysis, namely a hybrid approach based on CS, Linguistic Rules and Machine Learning (ML). Thanks to the analysis of approximatively five thousand English documents from a variety of sources (newspapers, social media and books) we identified several stylistic features that we used to write linguistic rules for English. Here we report a short list of stylistic features: sentence length (Argamon et al., 2003), vocabulary richness (De Vel et al., 2001), word length distributions (Zheng et al., 2006), punctuation (Baayen et al., 1996), use of a specific class of verbs or adjectives, use of first/third person. The hybrid approach of CS, Linguistic Rules and ML consists in the following steps: I) *Linguistic Definition of Stylometric Features*: starting from the assumption that each author operates different grammatical choices when writing a text (Daelemans, 2013), we organized the grammatical characteristics of the case-study language (in this case, English) in a taxonomy. The work was carried out thanks to COGITO® by Expert System Corp., a semantic analysis software based on Artificial Intelligence algorithms. In each limb of the taxonomy it is possible to write linguistic rules concerning the language of the case study in order to recognize the grammatical characteristics of the analyzed texts (i.e. to detect modal verbs, we create the limb "modal verbs" and we associate to it linguistic rules that allow to find modal verbs in the texts); II) *Semantic Engine Development*: Expert System's semantic engine is trained in order to extract the aforementioned features from texts and is implemented thanks to COGITO®'s semantic network (called *Sensigrafo*); III) *Features Extraction*: texts are analyzed and all features (based on the grammatical characteristics of the texts) are extracted; IV) *Supervised ML Process*: the features extracted are used to train the model in order to detect the features in the untagged texts. For ML process we exploit WEKA (Hall et al., 2009), a software with ML tools and algorithms for data analysis.

The hybrid approach is evaluated through the 10-folds Cross Validation method. We tested two different algorithms, Random Forest (RF) and Tree J48 (J48). During previous AA investigations RF resulted to be the most performing algorithm for a binary classification. The results we obtained for 10-folds Cross Validation test confirm this result and Table 1 presents the performances in terms of Precision, Recall and F-Measure for both algorithms (namely, RF and J48). In order to evaluate the performances of the classifier, after this process, we tested both RF and J48, as well as 10-folds Cross-Validation (Table 1). Compared to the results obtained for the 10-folds Cross Validation (see Tables 2 and 3), J48 performances (Table 3) are better than RF performances (Table 2).

10-folds Cross Validation (RF)	Precision	Recall	F-Measure
	0.815	0.824	0.808
10-folds Cross Validation (J48)	Precision	Recall	F-Measure
	0.779	0.784	0.781

Table 1: 10-folds Cross Validation on the whole corpus with RF and J48

Test Set 80-20 (RF)	Precision	Recall	F-Measure
	0.781	0.750	0.764

Table 2: Performances in terms of Precision, Recall and F-Measure with 80% of the corpus as Training set and the remaining 20% as Test set randomly selected with the support of RF algorithm

Test Set 80-20 (J48)	Precision	Recall	F-Measure
	0.853	0.800	0.819

Table 3: Performances in terms of Precision, Recall and F-Measure with 80% of the corpus as Training set and the remaining 20% as Test set randomly selected with the support of J48 algorithm

4.2 Support Vector Machine (SVM)

For our second approach we exploited SVM with a Bag-of-Words (BoW) features set created using TF-IDF vectorization. As stated by (Diederich et al., 2003) SVM is capable to process thousands of inputs, which allows to use all the words of a text directly as features. SVM involves building a decision boundary to separate the data into classes (in our case, *before* and *after*), which may be non-linear if the kernel trick is used to transform our existing data into a higher dimensional space. As such, the right choice to take when fitting an SVM classifier is kernel in addition to others hyperparameters specific to that kernel. In applying SVM to AA, (Schwartz et al., 2013) used a linear kernel, while (Diederich et al., 2003) examined a range of different kernels. Since our AA is a binary classification problem we used the linear kernel for our model and considered C values in the set {1, 10, 100}. The optimal value of C was determined using GridSearchCV function with a default 3-fold Cross-Validation and accuracy used as the scoring metric. The optimal C value was determined to be C = 1. Results are in Tables 4 and 5.

SVM-BoW	Precision	Recall	F-Measure
	0.785	0.761	0.773

Table 4: 10-folds cross validation SVM - BoW features set.

SVM-BoW	Precision	Recall	F-Measure
	0.885	0.809	0.818

Table 5: Performances in terms of Precision, Recall and F-Measure with 80% of the corpus as Training set and the remaining 20% as Test set randomly selected.

4.3 Convolutional Neural Network (CNN)

To deal with the problem of AA of speech transcriptions, our third approach consists in a two-class text classification based on a deep CNN. We built a neural network that exploits the morpho-syntactic

information to improve the classification and correctly identify the given samples. The input data are preprocessed and tagged with linguistic information using the Part-of-Speech (PoS) tagger provided by the free NLP open-source library *Spacy*. Given the importance of function words (Kestemont, 2014), conjunctions, prepositions, interjections, adverbs and auxiliary verbs were taqken into account for this analysis. In fact, as proved by (Mosteller and Wallace, 1963) and confirmed by (Koppel et al., 2006), function words are discriminators of authorship, since the usage variations of such words are a strong reflection of stylistic choices. Our proposed architecture receives a sequence of tagged texts as input and then is transformed into padded sequences of fixed length. The sequences are then processed by four modules: an embedding module, a convolutional module and two max pooling layers to consolidate the output of the convolutional layer. The output of the three modules are processed by one Dense layer and an output layer. Results are shown in Tables 6 and 7.

CNN-PoS	Precision	Recall	F-Measure
	0.681	0.734	0.706

Table 6: 10-folds cross validation CNN + PoS.

CNN-PoS	Precision	Recall	F-Measure
	0.692	0.818	0.750

Table 7: Performances in terms of Precision, Recall and F-Measure with 80% of the corpus as Training set and the remaining 20% as Test set randomly selected.

5 Differences and Similarities before and after 9th November 1966

Thanks to the linguistic analysis of the texts belonging to the two different classes, we detected stylistic differences and similarities in the speech transcriptions before and after 9th November 1966. We started by dividing Paul McCartney's interviews into separate sentences. A number of stylistic features are extracted from these sentences and then all features are used for K-Means clustering. Here we report a list of some features extracted: number of function words, number of verbs and a number of interjections. For clustering, the average of each feature is calculated. Further, a SVM classifier is trained on 70% of the interviews and tested on the remaining 30%. Performing this process means to see whether a link is present and consistent over time through Paul McCartney's style. Accuracy is shown in Table 8.

Accuracy on test set
0.561

Table 8: SVM to test stylometric similarity

Here we report some examples of interview transcriptions before and after 9th November 1966 and we highlight the most noticeable differences and similarities. It is very important to consider that all the interviews collected are from different sources (TV, radio, newspaper), that means that speeches can differ from source to source as well as according to the historical moment in which they were done.

- You know like, number one records, Sunday Night At The Palladium, Ed Sullivan Show, go to America, you know. All kinds of ambitions like that. (**Carnegie Hall - New York, 1964 February 12th**);
- The only thing is that we've gotta do a lot from London, 'cuz a lot of the TV shows are down in London, you know. And so, we're forced to do a lot down in London. I mean, it's like someone said the other day Why doesn't Harry Secombe go to Cardiff? You know, he never does. But no

one ever moans about Harry never going...You know what I mean? (**BBC-TV by Gerald Harrison - Liverpool, 1964 July 10th**);

- Personal differences, business differences, musical differences, but most of all because I have a better time with my family. Temporary or permanent? I don't really know. (**Break-up - 1970 April 10th**);
- It was like a gesture to Russia because normally records are released first in America and England in Europe and then Russia gets them last and because Gorbachev and Reagan were talking about glasnost and we're talking about arms reduction. I think a lot of us in Europe were very happy to hear this so I had the opportunity to release this record so I wrote a little note on the record saying this is the peace gesture the hand of friendship from the west to the east and I just felt it might just help a bit of glasnost it's my little bit of glasnost. (**Flemish Public Television Interview - 1989**)

As we can see, slang expressions and fillers such as *'cuz*, *I mean*, *You know?* and *You know what I mean?* completely disappear in interviews after 9th November 1966. The use of slang disappears also in other interviews after this date, in which we can find a different Paul McCartney, who seems to be more serious and not only because of an older age. Changes can be brought about by the different topics addressed in the interviews, but we also believe that speech preserves some characteristics (such as slang) in different contexts. In texts belonging to the *after* class, sentences are longer compared to those of the *before* class. We noticed also that in texts belonging to the *after* class style changes occur continously not allowing for the identification of a specific style. For these reasons we also report the date and the source. Our research highlights some similarities in *before* and *after* texts: the overuse of expressions such as *We are gonna do* and *a lot/a lot of* is confirmed in both periods. These represent the most used expressions by Paul McCartney in his speeches. In the interviews in the *Let IT Corpus* we also noticed that Paul McCartney is inclined to rely on lists both in *before* and in *after* periods.

6 Conclusions

In this paper we have presented the *Let IT Corpus*, namely a corpus of one-hundred four transcriptions from speech to text of Paul McCartney's interviews collected from the *Beatles Interviews Database* and YouTube. The aim of this research is to detect possible differences and similarities in Paul McCartney's speeches before and after 9th November 1966 (date of his supposed death). For this reason texts have been organised in two classes: I) *before* and II) *after*. We investigated three different text classification approaches and we detected that all methods achieved high percentage of accuracy classifying texts in two different classes referring to two different periods. To reinforce these results and on the basis of the analysis of the stylistic features set out above, it is clear that the way of modulating the words of Paul McCartney is quite distinguishable between the two periods examined.

7 Future Work

The corpus is in its embryonic stage, since it is composed of approximatively a hundred texts. Future work therefore concerns the expansion of the *Let IT Corpus*, so to allow a more thorough investigation. To corroborate our hypothesis it might be interesting to see if the differences we detected between the two classes represent a pure coincidence. A possible experiment in this respect can be carried out considering a different temporal division of the texts.

Acknowledgements

This research has been partly supported by the PON Ricerca e Innovazione 2014-20 and the POR Campania FSE 2014-2020 funds. Authorship contribution is as follows: Antonio Pascucci is author of Sections 1, 2, 3, 4.1 and 7 and Raffaele Manna is author of Sections 4.2 and 4.3. Section 5 is in common. This research has been developed in the framework of two Innovative Industrial PhD projects in Computational Stylometry (CS) by "L'Orientale" University of Naples in cooperation with Expert System Corp. We are grateful to Vincenzo Masucci and Expert System Corp. for providing COGITO® for research and to Prof. Johanna Monti for supervising the research.

References

Edoardo M Airoldi, Annelise G Anderson, Stephen E Fienberg, Kiron K Skinner, et al. 2006. Who wrote ronald reagan's radio addresses? *Bayesian Analysis* 1(2):289–319.

Shlomo Argamon, Marin Šarić, and Sterling S Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 475–480.

Harald Baayen, Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3):121–132.

Glauco Cartocci. 2005. *Il caso del doppio Beatle: il dossier completo sulla" morte" di Paul McCartney*. Robin Edizioni IT.

Walter Daelemans. 2013. Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 451–462.

Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining e-mail content for author identification forensics. *ACM Sigmod Record* 30(4):55–64.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence* 19(1-2):109–123.

Mel Evans. 2018. Style and chronology: A stylometric investigation of aphra behn's dramatic style and the dating of the young king. *Language and Literature* 27(2):103–132.

Richard S Forsyth. 1999. Stylochronometry with substrings .

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11(1):10–18.

Jonathan Herz and Abdelghani Bellaachia. 2014. The authorship of audacity: Data mining and stylometric analysis of barack obama speeches. In *Proceedings of the International Conference on Data Mining (DMIN)*. The Steering Committee of The World Congress in Computer Science, Computer . . . , page 1.

Jerremy Holland, Jan Erik Solem, and William E Hensler. 2014. Auto-recognition for noteworthy objects. US Patent 8,755,610.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. pages 59–66.

Moshe Koppel, Navot Akiva, and Ido Dagan. 2006. Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology* 57(11):1519–1525.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association* 58(302):275–309.

Jan Rybicki. 2015. Vive la différence: Tracing the (authorial) gender signal by multivariate analysis of word frequencies. *Digital Scholarship in the Humanities* 31(4):746–761.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1880–1891.

Constantina Stamou. 2007. Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing* 23(2):181–199.

Dirk Van Hulle and Mike Kestemont. 2016. Periodizing samuel beckett's works: A stylochronometric approach. *Style* 50(2):172–202.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology* 57(3):378–393.

Digital Projects for Music Research and Education from the Center for Music Research and Documentation (CIDoM), Associated Unit of Spanish National Research Council[1]

Juan José Pastor Comín
University of Castilla-La Mancha (Spain)
juanjose.pastor@uclm.es

Francisco Manuel López Gómez
University of Castilla-La Mancha (Spain)
franmalogo@hotmail.com

Abstract

English. This paper focuses on the description of the two digital databases developed by the CIDoM (Centro de Investigación y Documentación Musical –Center for Music Research and Documentation–), intended for the cataloguing of the musical heritage of the region of Castilla-La Mancha (Spain), on the one hand, and for the study of musical presence in the work of the most important writer in Spanish language, Miguel de Cervantes.

Italiano. Questo lavoro è focalizzato sulla descrizione dei due database digitali sviluppati dal CIDoM (Centro de Investigación y Documentación Musical –Centro di Ricerca e Documentazione Musicale–), destinati, da una parte, alla catalogazione del patrimonio musicale della regione di Castilla-La Mancha (Spagna) e, dall'altra, allo studio della presenza musicale nell'opere del più importante scrittore in lingua Spagnola, Miguel de Cervantes.

1 The Center for Music Research and Documentation of Spain

The Centro de Investigación y Documentación Musical is an Associated Unit of Centro Superior de Investigaciones Científicas (CSIC, Spanish National Research Council). Founded in 2012 and formed by an interdisciplinary team of PhDs in Musicology, History, History of Art and Hispanic Philology coordinated by Professors Paulino Capdepón and Juan José Pastor, CIDoM has among its objectives to replace and restore the musical heritage in one of the most important regions of Spain, Castilla-La Mancha, a large area of 80.000 Km^2, that hosts civil and religious centres of a great and historical musical activity. Centres as the music chapel of the Toledo and Cuenca Cathedrals attracted a large number of composers. This musical legacy has remained, unfortunately, mostly unknown. In the last years CIDoM has developed several national I+D+i Research Projects (Research + Development + innovation) focused in the Musical Heritage of Castilla-La Mancha and its Critical Analysis, Reception and Digital Edition.

[1] This work is inscribed within the context of the projects HAR2017-86039-C2-2-P. "El patrimonio musical de la España moderna (siglos XVII-XVIII): recuperación, digitalización, análisis, recepción y estructuras retóricas de los discursos musicales" [The Musical Heritage of the Modern Spain (17th and 18th centuries): recovery, digitalisation, análisis, reception and rhetorical structures of musical discourses] and PTA2016-13106-I, Catalogación y Digitalización del Patrimonio Musical de Castilla-La Mancha [Cataloguing and Digitalisation of the Musical Heritage of Castilla-La Mancha].

2 The digital database for Musical Heritage of Castilla-La Mancha

The digital database gathers the inventory of the musical sources to be consulted by the scientific community by means of a page web (beta.cidom.es) in order to analyse a little-known musical heritage: the Castilla-La Mancha's musical legacy during the Renaissance and Baroque period. The main aim is to cover the enormous gap existing in the region as regards the lack of an institution responsible for research and musical documentation. In this sense CIDoM is presented as a qualified proposal capable of fulfilling its responsibilities in the directions proposed by the International Council on Archives (ICA)[2], the International Association of Music Libraries, Archives and Documentation Centers (IAML)[3], the Spanish Association of Musical Documentation (AEDOM)[4], the Spanish Society of Musicology (SEdeM) and the National Institute of Performing Arts and Music (INAEM)[5], following the standards of quality and criteria of documentation and research sanctioned by these institutions to which the members of the Center belong. This essential task must allow, on the one hand, the analysis and study of the works of the fundamental composers of our region –recognized by international musicology– belonging to different periods, such as Diego Ortiz, Sebastián de Covarrubias, Alonso Xuárez, Torrejón and Velasco. On the other hand, it will allow to order, classify and catalogue the sources of information and musical work ignored and contained in the cathedral and administrative archives of the region, Toledo, Sigüenza, Talavera de la Reina, Cuenca, Guadalajara, Pastrana, etc., with the aim to provide the professional community with the opportunity to publish and record this musical heritage.

The digital catalogue is managed by two independent databases with which information relating to composers is recorded (including basic data of interest for the user's search, biography, registered works, bibliography and discography), on the one hand, and musical works, on the other. The process of cataloguing the latter has been developed from the standards ISAD(G) (1999), ALA (2004), and the guidelines by González-Valle (1996), Miliano (1999), Schultz and Shaw (2003), so that the following fields are considered: composer, title, type of document (manuscript or printed), genre, year of composition (or time bracket, when unknown), date of the source (or time bracket, when unknown), author of the text, literary source, premiere date, file in which it is kept, signature, coded musical incipit (allows works with the same musical beginning to be located, even if they are transported to another mode or tonality, so that it can automatically detect borrowings between different works)[6], literary incipit, vocal and instrumental template, facsimile (allows the user to download directly a scanned copy of the source in PDF format), transcription (if available), data on the edition (if applicable), and observations (including data such as the diplomatic title of the work, information on the copy and the physical state of the medium, its measurements, the number of folios or pages, the tone or mode of the work and the time signature)[7].

3 The digital database for Cervantes and Music

Every musical adaptation of a literary work has to be considered as an exercise of perception and interpretation that provides additional information on the hermeneutics of a writer's works. These adaptations allow us to understand and explain how the literary work has been recreated and transformed in each epoch. Once a play or a novel has been put in music, every musical version offers the audience a sort of critical and at the same time musical thought that reflects a new conception –even misconception– of the work. It cannot be denied that Cervantes' works have provided composers excellent material for their musical compositions and this fact has to be taken into account to describe the process of his musical

[2] They have developed the ISAD(G), General International Standard Archival Description, a guidance for cataloguing activities that has been considered for CIDoM's projects.
[3] That includes the MARC 21 standard for the representation and exchange of bibliographic information data.
[4] In particular, taking into account the orientations offered in Eudom (2010).
[5] Under the conclusions of the symposium *La gestion del patrimonio musical* (Management of the Musical Heritage), expressed in Álvarez Cañibano et. al. (2014).
[6] The coding adopted consists in recording the intervallic distances between the musical notes of the incipit. To this purpose, the quality of each interval (ascending or descending) in number of semitones is recoded. For example: C-E flat-D-G would be noted as 3a-1d-5a.
[7] You can access the catalogue of the historical musical heritage in Castilla-La Mancha, made by the CIDoM, in the following link: http://beta.cidom.es/patrimonio-musical/patrimonio-musical-historico/bases-de-datos-de-compositores-y-obras-de-castilla-la-mancha

reception: how the characters and the episodes of his works have been selected by composers and perceived by the audience and what kind of musical treatment –genres, musical patterns, etc.– each composer provides (Pastor, 2007; Pastor, 2009).

At the same time, there cannot be any doubt that Cervantes' works reflect faithfully the Spanish musical world of 16th and 17th centuries: musical instruments, dances and *bailes*, romances and songs are often cited and performed in his pages in order to depict not only a special and picturesque environment in which his characters evolve such as a gypsy's world in *La gitanilla* or Muslim's traditions in *La gran sultana* or *Los baños de Argel*, but in addition assign a particular semantic value to each musical element adding a supplementary meaning to the work's understanding (Pastor, 2005; Pastor, 2006). Our digital project distinguishes between three different aspects considered as a powerful educative instrument:

3.1. Musical instruments

This first point of the project will provide a catalogue of the musical instruments cited by Cervantes in his works, explaining their social functions in the texts and offering, from an educative point of view, different sound files, image files and text files in order to familiarize the users with the musical world around Don Quixote's author[8]. Let's consider some examples. In the First Part of Don Quixote (I, XXVI), the mad knight says to his squire:

> […] for know, Sancho, that all or most of the knights-errant of times past were great poets and great musicians; these two accomplishments, or rather graces, being annexed to lovers-errant. True it is, that the couplets of former knights have more of passion than elegance in them. (Don Quixote, I, XXVI)

In the Second Part, in the adventure in Duke's Palace, Don Quixote requests a lute to console Altisidora:

> "Do me the favour, señora, to let a lute be placed in my chamber to-night; and I will comfort this poor maiden to the best of my power; for in the early stages of love a prompt disillusion is an approved remedy;" and with this he retired, so as not to be remarked by any who might see him there.
>
> He had scarcely withdrawn when Altisidora, recovering from her swoon, said to her companion, "The lute must be left, for no doubt Don Quixote intends to give us some music; and being his it will not be bad."
>
> They went at once to inform the duchess of what was going on, and of the lute Don Quixote asked for, and she, delighted beyond measure, plotted with the duke and her two damsels to play him a trick that should be amusing but harmless (Don Quixote, II, XLVI)

There cannot be any doubt that Cervantes's works faithfully reflect the Spanish musical world of the 16th and 17th centuries: musical instruments, dances and *bailes*, romances and songs are often mentioned and performed in his books depicting not only the environment in which his characters evolve but they also add a particular semantic value to each musical element. Participants in this galaxy of musical performance are representatives of all walks of life, from the highest noble to the lowliest peasant, and the number of instruments one encounters in Cervantes's writings is truly extensive. Cervantes groups them in pastoral, military, popular and aristocratic and there are fifty different instruments cited in his works. Let's go to see some examples, but I would caution previously that the English translations consulted don't respect exactly the nature of musical instruments.

We see in Cervantes that harps and lutes are playing together. We have several texts in Cervantes that describe the performance of harps and lutes together:

> *But the instant the car was opposite the duke and duchess and Don Quixote the music of the clarions ceased, and then that of the lutes and harps on the car, and the figure in the robe rose up, and flinging it apart and removing the veil from its face, disclosed to their eyes the shape of Death itself,* fleshless and hideous, at which sight Don Quixote felt

[8] To access the catalogue of musical instruments in Cervantes, use the following link: http://beta.cidom.es/musica-y-literatura/cervantes-y-la-musica/instrumentos-musicales-en-cervantes.

uneasy, Sancho frightened, and the duke and duchess displayed a certain trepidation. Having risen to its feet, this living death, in a sleepy voice and with a tongue hardly awake, held forth as follows:

I am that Merlin who the legends say
The devil had for father, and the lie
Hath gathered credence with the lapse of time. (*Don Quixote,* II, XXXV)

The harp is used too as an aristocratic instrument for ladies:

Calliope

With so much peculiarity, with so much sweetness, with such harmony, *she touched the harp of the graceful muse.* She, having sounded the strings awhile, with a voice sonorous past conception, then gave utterance to these stanzas:

Song of Calliope
To the sweet sound of my attempered lyre
Oh shepherds listen with attentive ear (*La Galatea,* V)

Lucinda

I passed in such employments as are not only allowable but necessary for young girls, those that the needle, embroidery cushion, and spinning wheel usually afford, and if to refresh my mind I quitted them for a while, I found recreation in reading some devotional book or *playing the harp, for experience taught me that music soothes the troubled mind and relieves weariness of spirit*. (*Don Quixote,* I, XXVIII)

Altisidora

He trembled lest he should fall, and made an inward resolution not to yield; and commending himself with all his might and soul to his lady Dulcinea he made up his mind to listen to the music; and to let them know he was there he gave a pretended sneeze, at which the damsels were not a little delighted, for all they wanted was that Don Quixote should hear them. *So having tuned the harp, Altisidora, running her hand across the strings, began this ballad:*

O thou that art above in bed,
Between the holland sheets,
A-lying there from night till morn,
*With outstretched legs asleep; (Don Quixote,*II, XLIX)

All these elements studied and analyzed can be consulted on the digital platform: http://beta.cidom.es/musica-y-literatura/cervantes-y-la-musica/instrumentos-musicales-en-cervantes/instrumentos/1/arpa.html

3.2. Songs, romances, dances and bailes

This second point deals with the accomplishment of the digital edition of the scores related with Cervantes' texts (Pastor, 2017). For example, some chapters of the First Part of Don Quixote begin with the first verse of a sung poem. This interactive frame will be accompanied by sound files, facsimile editions, bibliographical information about composers, and different articles explaining the significance of the relationship between music and poetry in Cervantes' works[9]. I would like to underline that some chapters of the *First Part of Don Quixote* begin with the first verse of a sung poem. Many chapters of both parts begin with one, two or several "accidental verse-lines" –prose lines that may be read and, consequently, sung– as endecasyllables, octosyllables, heptasyllables: there are also so many indeed that we must assume they are not there by chance but deliberately. It shouldn't be overlooked that chapter one of the *First Part of Don Quixote* also begins with a ballad-line to identify the place where Don Quixote lived: "En un lugar de la Mancha" [In a place in La Mancha]. Although in this last case we haven't got any evidence or proof of its musical performance, it's easy to imagine that Cervantes might have conceived the beginning of his novel like an epic poem composed to be sung (Pastor, 2005).

[9] Use the following link to see the database of songs, romances, dances and *bailes* in Cervantes' texts: http://beta.cidom.es/musica-y-literatura/cervantes-y-la-musica/danzas-y-bailes-en-cervantes

The same thing happens with another romance, "Mira Nero de Tarpeya" ("Nero fiddled while Rome burned"), that relates the history of an indolent Nero playing the harp from Tarpeian hill while Rome was burning. This romance was very famous in Iberian Peninsula and was put in music by Bermudo (*Declaración de instrumentos musicales*, 1555) and Venegas de Henestrosa (*Libro de cifra nueva para tecla, harpa y vihuela,* 1557). Cervantes introduces and intersperses in several episodes of *Don Quixote* this musical reference as an echo of the musical romance emphasizing the semantic value of the *madness*. First time it appears, is after *Desperate song* of Grisóstomo:

> Or comest thou to triumph in the cruel exploits of thy inhuman disposition, or to behold from that eminence, like another *pitiless Nero, the flames of burning Rome; or insolently to trample on this unhappy corpse, as did the impious daughter on that of her father Tarquin? (Don Quixote*, I, XIV)

Second occurrence, it appears as parody, when Sancho gets stuffed in Camacho's Wedding:

> Sancho beheld all this, *and was nothing grieved thereat*; but rather, in compliance with the proverb he very well knew, *When you are at Rome, do as they do at Rome*, he demanded of Ricote the bottle, and took his aim, as the others had done, and not with less relish. (*Don Quixote,* II, LIV)

The last occurrence of the romance is part of the fun of Altisidora, who makes mock of Don Quixote, integrated in another long romance she sings:

> *Manchegan Nero, look not down*
> *From thy Tarpeian Rock*
> *Upon this burning heart, nor add*
> *The fuel of thy wrath.* (*Don Quixote,* II, XLIV)

All these elements studied and analysed can be consulted too on the digital platform: http://beta.cidom.es/musica-y-literatura/cervantes-y-la-musica/canciones-y-topicos-musicales-en-cervantes/canciones/8/mira-nero-de-tarpeya.html

3.3. Musical reception of Cervantes' works

Finally, the development of this project will provide a complete catalogue of musical compositions based in Cervantes' texts. Information included will be articulated by genres, countries, and musical periodization and it will be the first step to seriously study how the Cervantes' literary genius has encouraged the composers' creative imagination[10]. Some composers who have put the work of Cervantes into music can be consulted on our digital platform: http://beta.cidom.es/musica-y-literatura/cervantes-y-la-musica/la-recepcion-musical-cervantina/recepcionmusical/3/millan-de-las-heras-manuel-1971--.html

4 Conclusions

For the CIDoM, the main objective is the cataloguing and digitalisation of the musical heritage of the region of Castilla-La Mancha, as well as facilitating the researcher's search and relation between data, and providing access to primary sources. In addition, it is crucial to project the results of our research on the area of Music Education and to disseminate this information to the educational community, in order to create and to implement educational tools –demanded by music teachers– concerned with music heritage, thus increasing the quality and the cross-sectional relations of the musical education in the different educational levels. In this sense, the project about musical reception of Cervantes' works has a high pedagogical project for us (Pastor, 2016). For this reason, the interdisciplinary vocation with which the digital projects presented here are born seeks in the educational field the adequate space to project university research on the reality of other academic levels.

[10] You can access to the database about musical reception of Cervantes' works in the following link: http://beta.cidom.es/musica-y-literatura/cervantes-y-la-musica/la-recepcion-musical-cervantina

References

ALA (American Library Association). 2004. *Reglas de Catalogación Angloamericana* (2nd. ed). Bogotá, Rojas Eberhard Editores.

Antonio Álvarez Cañibano, Eugenio Gómez del Pulgar, Mª José González Ribot, Pilar Gutiérrez Dorado and Cristina Marcos Patiño. 2014. *La gestión del patrimonio musical. Situación actual y perspectivas de futuro*. Madrid, Centro de Documentación de Música y Danza, INAEM. Retrieved from http://www.musicadanza.es/ficheros/documentos/actas-simposio

EUDOM (ed.). 2010. *Clasificación sistemática de libros de música, partituras y grabaciones sonoras*. Valencia, AEDOM.

ISAD(G). 1999. *General Internacional Standard Archival Description* (2.ª ed.). Retrieved from https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_ EN.pdf

José González Valle (ed.). 1996. *Normas internacionales para la catalogación de fuentes musicales históricas (Serie A-II, Manuscritos musicales, 1600-1850)*. RISM-España, Arco Libros, S.A.

Juan José Pastor Comín. 2005. *Por ásperos caminos. Nueva música Cervantina*. Cuenca, Ediciones de la Universidad de Castilla-La Mancha.

Juan José Pastor Comín. 2006. "*Terminorum musicae index.* La organología musical en la obra cervantina y su proyección didáctica". In Pastor Comín, J. J, and Ángel G. Cano (Coords.) *Don Quijote en el Aula. La aventura pedagógica.* Ciudad Real, Servicio de Publicaciones de la UCLM. pp. 231-248.

Juan José Pastor Comín. 2007. *Cervantes: Música y Poesía. El hecho musical en el pensamiento lírico cervantino*. Vigo, Editorial Academia del Hispanismo.

Juan José Pastor Comín. 2009. *Loco, trovador y cortesano: bases materiales de la expresión musical en Cervantes.* Vigo, Editorial Academia del Hispanismo.

Juan José Pastor Comín. 2016. "Las músicas de Cervantes: recursos digitales para la recuperación del patrimonio musical" in *Innovación Universitaria: digitalización 2.0 y Excelencia en Contenidos.* Madrid, McGraw-Hill Education, sub-colección "Innovación y Vanguardia Universitarias, pp. 649-663.

Juan José Pastor Comín and Paulino Capdepón (coords.). 2017. "*Trabajos que nacen del espíritu". Estudios sobre música y literatura en las obras cervantinas.* Madrid, Alpuerto.

Juan José Pastor Comín. 2018. "Creación poética y escritura musical: lectura, interpretación, recepción y canon", en Lolo, Begoña y Presas, Adela (eds.), *Musicología en el siglo XXI: nuevos retos, nuevos enfoques*. Madrid, SEdeM, pp. 719-727.

Lois Schultz and Sarah Shaw. 2003. *Cataloging Sheet Music Guidelines for Use with AACR2 and the MARC Format*. Lanham, Md., Scarecrow Press, Music Library Association.

Mary Miliano (ed.). 1999. *The IASA Cataloguing Rules*. Stockholm, Baden-Baden, International Association of Sound and Audiovisual Archives.

Prospects for Computational Hermeneutics

Michael Piotrowski
Deptartment of Language and Information Sciences Faculty of Arts
University of Lausanne Lausanne, Switzerland
michael.piotrowski@unil.ch

Markus Neuwirth
Digital and Cognitive Musicology Lab
Digital Humanities Institute
College of Humanities
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
markus.neuwirth@epfl.ch

Abstract

English. The central concern of the humanities is the understanding of human artifacts. This goal requires interpretation and makes hermeneutics a core element of their methodology. So far, the digital humanities have been mostly concerned with providing tools that either support human interpretation or that permit scholars to record results of their interpretation through annotation. However, if we understand digital humanities as the construction of formal models in the humanities, we must also strive to integrate hermeneutics into these models. In this paper, we reflect on the role of hermeneutics in digital humanities and sketch an approach for combining human interpretation with formal models.

Italiano. La preoccupazione centrale delle scienze umane è la comprensione dei artefatti umani. Questo obiettivo richiede interpretazione e fa dell'ermeneutica un elemento centrale della loro metodologia. L'informatica umanistica si è finora occupata soprattutto di fornire strumenti che supportano l'interpretazione umana o che permettono agli studiosi di registrare i risultati della loro interpretazione attraverso l'annotazione. Tuttavia, se intendiamo l'informatica umanistica come la costruzione di modelli formali nelle scienze umane, dobbiamo anche cercare di integrare l'ermeneutica in questi modelli. In questo lavoro riflettiamo sul ruolo dell'ermeneutica nell'informatica umanistica e tracciamo un approccio per combinare l'interpretazione umana con i modelli formali.

1 Introduction: The Theoretical DH

The interpretation of human artifacts in order to understand their "meaning" is the central concern of the humanities. They are therefore often characterized as being "qualitative-hermeneutical," in contrast to the natural sciences and to computer science, which are supposedly "empirical," "quantitative," and much less dependent on interpretation. However, as Piotrowski (2018) argues, disciplines are not defined by their methods alone but rather by a "unique combination" of a research object and a research objective; research methods, he notes, are "secondary in that they are contingent on the research object and the research objective." In addition, technical and scientific progress not only enables methods to evolve, but also requires them to adapt, while research objects and objectives largely remain stable. Even though particular methods may be "typical" for a particular discipline, all disciplines can, in principle, use any methods, including computational ones, as long as they fit their research objectives. As Orlandi puts it, "un po' di aritmetica ha sempre fatto parte delle discipline umanistiche" (Orlandi, 1990, 114). Conversely, other fields also use methods commonly associated with the humanities. For example, Frodeman (1995) has argued that geology is really a "historical and interpretive science," rather than a "derivative science, relying on the logical techniques exemplified by physics" (see also Comet, 1996). Similarly, artificial intelligence (e.g., Winograd, 1981) and computer science more generally (e.g., West, 1997) attempted to formalize and apply the key concept of hermeneutics, namely, *understanding*.

The research carried out under the heading of "Digital Humanities" (DH) currently tends to focus on quantitative analyses, which have long been difficult or even impossible in the humanities and which

yield important new insights complementing traditional ("manual") qualitative analyses. However, if we understand DH as the construction of formal models in the humanities (Piotrowski, 2018; McCarty, 2014; Orlandi, 1990), we must not neglect the qualitative-hermeneutic dimension. If the humanities want to succeed in answering their research questions—which are primarily qualitative in nature—they cannot rely on quantitative methods alone. Instead, a multilayered research process is required, one in which quantitative and qualitative analyses continuously alternate.

One of the main challenges for the *theoretical digital humanities* (Piotrowski, 2018) remains to find ways to *integrate* hermeneutic methods and insights into formal models, rather than keeping interpretation detached, as a kind of afterthought to automatic analyses (or vice versa). In this regard, there are noteworthy initiatives to exploit the computer as a "modeling machine" (McCarty, 2014, 256) while continuing the long philosophical tradition of hermeneutics (e.g., Dilthey, Heidegger, Gadamer, Ricœur, Iser, Jauss etc.). However, there does not seem to be a transfer back in the other direction. The goal of this paper is thus to outline the prospects for a novel approach that might be called *computational hermeneutics* and to stimulate a wide discussion on the possibility of a unified science bridging the gap between the humanities and the sciences.

2 Hermeneutics and Understanding

The main goal of any hermeneutic approach is to achieve what Dilthey called *Verstehen*, an "understanding" of human artifacts in order to answer questions about the "Why?" and "How?" and to uncover underlying patterns (Bod, 2015). But what exactly is *understanding*, and the *understanding* of what? One may argue that hermeneutic interpretation aims at uncovering the meaning of a given text by reference to the author's intention (whether empirical or idealized), the envisaged reader, and the dense web of meanings invoked by the text. *Understanding* thus involves a reconstruction of (a) the reasons why a given author (or group of authors) produced a particular text (*text* understood in a broad sense), (b) the overt or hidden layers of meaning of a given text, (c) the type of recipient envisaged by the author and/or the text, and (d) the potentially infinite number of contexts in which the reconstruction of meaning can take place (Ricœur's "conflit des interprétations").

Whenever we interpret language, we need to rely on some *pre-understanding* that provides the basis on which to build an interpretation. It is thus a recursive process, in which (pre-)understanding is necessary for interpretation, which in turn produces understanding, and so forth—hence the term *hermeneutic circle* (for an illuminating discussion see Göttner (1973)). Since this process leads to a progressive approximation to an (ideally exhaustive) understanding of a given text, Bolten (1985) has proposed the more apt metaphor of a *hermeneutic spiral*.

Despite the supposed "death of the author" (as famously heralded by Roland Barthes), authorial intention remains an important component of *pre-understanding*: the interpretation of texts cannot be successful when one just relies on lexical meanings and sentence semantics alone, both of which may not even be available when we understand texts in a broad sense. For an interpretation to be sound, one has to make complex inferences that rely on vast knowledge about the world and on the attribution of mental states (especially intentions) to the author. Here Grice's conversational maxims play a particularly important role in guiding the inferential process. It is characteristic of the ways in which hermeneutics is commonly construed that these inferential processes about *the other mind*, however foreign, are rarely reflected in depth (see Winograd's *model of the speaker* as part of the reader's model of the world (Winograd, 1981)). Nonetheless, despite the importance of authorial intention, it is necessary to draw a distinction between the meaning of a text and the meaning as intended by the (empirical rather than ideal) author. In other words, the meaning of a text cannot be reduced to the intended meaning either.

3 Hermeneutics and Digital Humanities

There are essentially two "native" strands of interpretation in DH, which both now have long traditions going back to the beginnings of computing in the humanities.

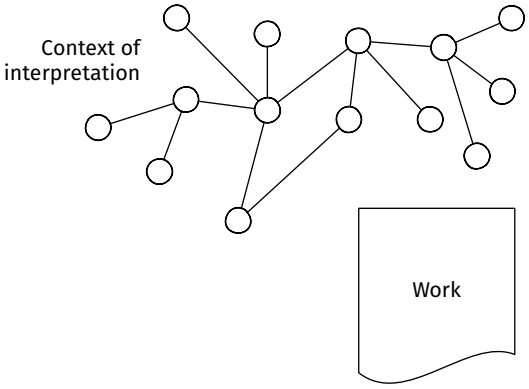


Figure 1 – Hermeneutics considers a work in a particular context of interpretation, peculiar to a reader, here modeled as a network of concepts. Links between concepts can be of various types; they can be thought of as mental associations.

One strand is that of *annotation*, exemplified by the Text Encoding Initiative (TEI, 1987).[1] It goes back to an even longer editorial tradition in philology, focusing primarily (though not exclusively) on a single text and the textual phenomena therein. It is thus a relatively "weak" form of interpretation, in the sense that it makes only limited connections to the extra-textual (which for Ricœur (2017, 103) is a defining feature of interpretation)—but intentionally so: editions are generally used as a basis for a later interpretation of the text.

The other strand, which Rockwell and Sinclair (2016) call "computer-assisted interpretation," builds on an equally long-standing tradition in literary studies, in particular concordancing, stylometry, and other quantitative analyses, and belongs to the first applications of computers in the humanities (see, e.g., Kroeber, 1967). The modern evolution of this strand can be exemplified by Rockwell and Sinclair (2016) and their work on Voyant.[2] This strand is oriented towards tools and automatic analyses informing human interpretation. Rockwell and Sinclair's notion of the "hybrid essay, an interpretive work embedded with hermeneutical toys" (Rockwell and Sinclair, 2016, 17) illustrates well the idea of the computer providing scholars with new evidence.

Both strands are not limited to philology and literary studies; they can also be found in other humanities disciplines, and images or other artifacts may replace texts as research objects. Outside of DH, computer-assisted interpretation (in the above sense) remains controversial (see, e.g., the debates following the publication of Da, 2019); critics typically question the legitimacy of quantitative methods in general.

However, both annotation and computer-assisted interpretation have an inherent limitation in common, which is rarely, if ever, discussed: human interpretation remains outside of the formal framework. In the case of annotation, the (formal) annotation is the result of a preceding human interpretation that motivates a particular annotation (say, that tagging of some text as "deleted"), but only the result (in the form of a `<del>` tag) is formally documented, the reasoning for this choice generally remains inaccessible, at least to the computer. Furthermore, it is usually difficult, or even (practically) impossible, to record alternative interpretations.

4 Proposal

How, then, could we link hermeneutics to formal models, so that human interpretations can be taken into account as well and different types of methods can be combined to truly complement each other? The idea of *mixed methods*, which originated in the social sciences (Kuckartz, 2014), certainly cannot be transferred to the domain of the humanities without modification. It is important to stress that the goal cannot be to "automate" interpretation; the bedrock of *Verstehen* is a shared understanding of the *conditio humana*.

[1] https://tei-c.org
[2] https://voyant-tools.org

The goal must rather be to support the scholar by making it possible, for example, to process qualitative human interpretations alongside the results of automatic quantitative analyses.

The basic idea of our proposal is to model the context of interpretation—i.e., a reader's knowledge of cultural concepts and the associations between them—as a *semantic network* or *knowledge graph* (see Fig. 1), and interpretation as the linking of features of the interpreted object to nodes of this network, i.e., the construction of a new network, as illustrated in Fig. 2. Understanding can thus be defined as the integration of the object's properties into a preexisting network.

Computationally, this model can be represented using Semantic Web and Linked Data technologies, which has the advantage that existing tools and methods can be leveraged. In particular, we propose to use *nanopublications*, a knowledge representation approach originally developed in bioinformatics (Groth et al., 2010), although the conceptual model is neutral with respect to a particular implementation. Nanopublications were developed as a common framework for describing scientific statements together with contexts (e.g., original publication, authors, organisms involved) in a machine-readable fashion, so that scientific results are easier to discover, unambiguously referenced and connected to particular scholars, and can be automatically aggregated and analyzed.

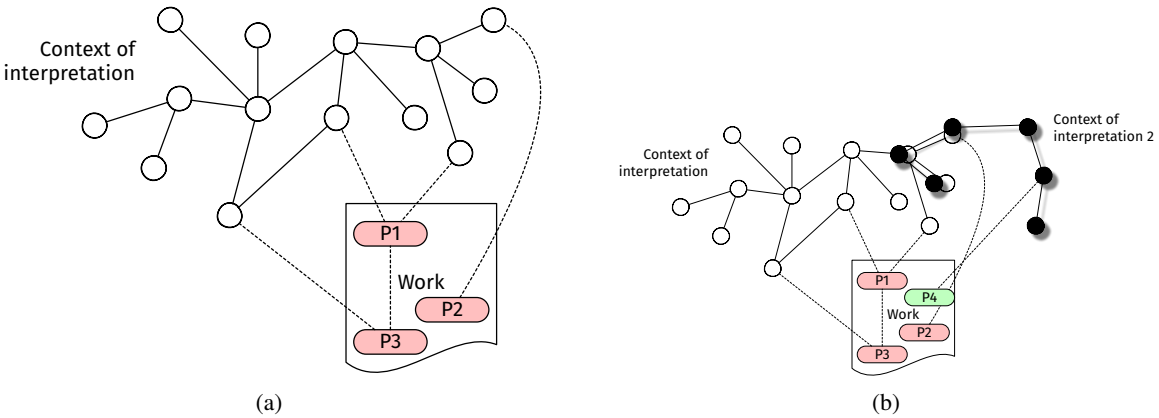


Figure 2 – (a) Interpretation links features of the work (here: passages P1–3) to concepts in the reader's context of interpretations. (b) The contexts of interpretation of different readers may partly overlap (and thus share associations) but may also have different relations and thus come to different interpretations of the same work.

5 Case Study

Is it possible to model the temporal (and geographical) dynamics of the *horizon of expectations*? Let us consider an example from music history to demonstrate our approach to computational hermeneutics. In his review of a symphony by Robert Volkmann (which is little known today), Selmar Bagge wrote: "Volkmann's Dmoll-Symphonie ist eine durchaus pathetische Production" (AmZ 48, 1863, col. 806). Suppose this sentence originated from a present-day source. In this case, a translation such as the following would be perfectly possible: "Volkmann's symphony in D minor is a quite emotive work". However, since a model of understanding contains assumptions about the author and the time of his or her writing, such a translation would ignore that the German word *pathetisch* has undergone a significant semantic shift. Today, *pathetisch* has a rather negative connotation and would thus have to be translated as 'melodramatic' or 'pompous.' To reveal the (historical) meaning likely to be intended by Bagge, we need to explore and model the contexts in which *pathetisch* has been used. These contexts have to be distinguished according to their distance to the target object of interpretation. Generally, an interpretation is more likely if it is supported by sources that show proximity in terms of time and space. In other words, sources that have been written around the same time and in the geographical vicinity of the source under investigation are to be preferred over sources that show greater temporal and geographical distance. Both temporal and geographical distances can best be modeled using network approaches (see above).

When consulting one central source, the approximately contemporaneous *Deutsches Wörterbuch* by the Brothers Grimm, we find *pathetisch* glossed as 'powerful,' 'dignified,' or 'solemn.' In addition, the word is linked to both the *passionate* and Schiller's concept of the "pathetic-sublime" (1793). Both of these usages are confirmed by much earlier sources: In Johann Georg Sulzer's *Allgemeine Theorie der Schönen Künste* (1793), the "pathetic" is considered a synonym of the "passionate." In Heinrich Christoph Koch's *Musikalisches Lexikon* from 1802, the reader interested in the meaning of "patetico, pathetisch" is directed to the entry on the "sublime" (Koch, 1802), thus suggesting that *pathetisch* and "sublime" are synonyms. More distant 18th century sources even suggest an association of the sublime with the "(delightful) horror." Given the historical distance, this connotation is less likely to be conveyed in Bagge's statement. Considering this complex semantic history, the modeling task consists in (1) linking related semantic concepts, (2) qualifying these links (e.g., as synonym, as super- and subcategory, or as semantic overlap), and (3) weighing links according to temporal proximity.

The model reader that Bagge had in mind when making his statement about Volkmann's symphony as being pathetic is somebody who had a certain prior knowledge of that concept (as reconstructed from the sources just mentioned). In addition to the semantic history of words, further contexts that need to be considered concern a dense web of musical works. The prototypes of a pathetic work, as invoked by Bagge, are Beethoven's 5th and 9th symphonies. Readers of the time likely understood this to be the primary context of Volkmann's symphony without which a proper understanding could not be achieved. Further works featuring "pathetic" in their titles are Beethoven's piano sonata op. 13 and, much later, Tchaikovsky's 6th symphony, the distance between these two works being roughly a hundred years. However, despite the lack of a title, many earlier symphonies (by other composers) from the late 18th century on have been referred to as invoking the "sublime," and hence are "pathetic" in Koch's sense. The reason for Bagge's aesthetic judgment thus lies in the shared musical properties of all the works contained in the set of pathetic or sublime symphonies: the minor mode, the orchestral setting, a particular tempo, etc. As a result, a hermeneutic reconstruction must consider both the semantic tradition (and change) of the word "pathetic" and the corresponding musical production.

6 Conclusion: Implications and Prospects

As outlined at the outset of our paper, the humanities and the sciences are widely assumed to be separated from each other by their respective methods, objects, and objectives. However, as suggested above, the humanities and the sciences face a common challenge: both have to address explicitly the issues of interpretation and decision-making under uncertainty. In particular, they need to formalize and model the contexts of interpretation and the inferential processes under uncertainty, seeking to exploit the rich potential of the computer as modeling machine (Piotrowski, 2019). The development of suitable probabilistic tools (Pearl, 2000) for modeling network-like relationships between objects is a crucial task for the whole scientific community, one that brings us closer to the ideal of a truly unified science.

The use of formalization and modeling is often met with a certain hostility in the humanities. Many humanities scholars subscribe to the notion that interpretation can in principle never come to a conclusion, and indeed the fascination of hermeneutics seems to lie in its inherent incompleteness. In addition, it is assumed that multiple interpretations can exist alongside each other without the need (or even the possibility) to prefer one over the other; this is in keeping with the cherished notion of plurality and multiplicity of perspectives in the humanities. Yet exactly in this respect a computational approach may offer obvious advantages, as the possibilities of formally representing interpretations, their contexts, and the inference procedures allow scholars to better compare different interpretations and assign different probability values to them (for applying a Bayesian approach to historiography and the problems of assigning prior probabilities (see Tucker, 2004; Carrier, 2012). More generally, this approach can give rise to the idea of *progress* in the humanities (something that is notoriously rejected by many humanities scholars). Thus the essential challenge of the theoretical digital humanities is to come up with a convincing approach to a "hermeneutic computer science" (West, 1997), whose tasks involves modeling interpretation contexts, inferential processes, and uncertainty.

References

Rens Bod. 2015. *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present*. Oxford University Press, Oxford.

Jürgen Bolten. 1985. Die Hermeneutische Spirale. Überlegungen zu einer integrativen Literaturtheorie. *Poetica* 17(3–4):355—371.

Richard C Carrier. 2012. *Proving History: Bayes's Theorem and the Quest for the Historical Jesus*. Prometheus Books.

Paul A. Comet. 1996. *Geological reasoning: Geology as an interpretive and historical science: Discussion*. Geological Society of America Bulletin 108(11):1508–1510. https://doi.org/10.1130/0016-7606(1996)108%3C1508:grgaai%3E2.3.co;2

Nan Z. Da. 2019. The computational case against computational literary studies. *Critical Inquiry* 45(3):601–639. https://doi.org/10.1086/702594

Wilhelm Dilthey. 1900. Die Entstehung der Hermeneutik. In *Philosophische Abhandlungen. Christoph Sigwart zu seinem 70.* Geburtstage gewidmet, J. C. B. Mohr (Paul Siebeck), Tübingen, pages 185–202.

Robert Frodeman. 1995. *Geological reasoning: Geology as an interpretive and historical science*. Geological Society of America Bulletin 107(8):960–968. https://doi.org/10.1130/0016-7606(1995)107%3C0960:grgaai%3E2.3.co;2

Heide Göttner. 1973. *Logik der Interpretation. Analyse einer literaturwissenschaftlichen Methode unter kritischer Betrachtung der Hermeneutik*. Fink, Munich.

Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, Academic Press, New York, volume 3 of *Syntax and Semantics*, page 41–58.

Paul Groth, Andrew Gibson, and Jan Velterop. 2010. The anatomy of a nanopublication. *Information Services and Use* 30(1–2):51–56. https://doi.org/10.3233/ISU-2010-0613

Heinrich Christoph Koch. 1802. *Musikalisches Lexikon*. Hermann, Frankfurt am Main.

Karl Kroeber. 1967. Computers and research in literary analysis. In Edmund A. Bowles, editor, *Computers in Humanistic Research*, Prentice-Hall, Englewood Cliffs, NJ, USA, chapter 13, pages 135–142.

Udo Kuckartz. 2014. *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Springer-Verlag.

Willard McCarty. 2014. *Humanities Computing*. Palgrave Macmillan, Basingstoke, paperback edition.

Tito Orlandi. 1990. *Informatica umanistica*. La Nuova Italia Scientifica, Rome.

Judea Pearl. 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Michael Piotrowski. 2018. Digital humanities: An explication. In Manuel Burghardt and Claudia Müller-Birn, editors, *Proceedings of INF-DH 2018*. Gesellschaft für Informatik. https://doi.org/10.18420/infdh2018-07

Michael Piotrowski. 2019. Accepting and modeling uncertainty. *Zeitschrift für digitale Geisteswissenschaften* Sonderband 4. https://doi.org/10.17175/sb004_006a

Paul Ricœur. 2017. *Le conflit des interprétations: Essais d'herméneutique*. Seuil, Paris.

Geoffrey Rockwell and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press, Cambridge, MA, USA.

Aviezer Tucker. 2004. *Our Knowledge of the Past: A Philosophy of Historiography*. Cambridge University Press.

Dave West. 1997. Hermeneutic computer science. *Communications of the ACM* 40(4):115–116. https://doi.org/10.1145/248448.248467

Terry Winograd. 1981. What does it mean to understand language. In Donald A. Norman, editor, *Perspectives on Cogniscience;* Norwood, NJ, USA pages 231-263

EModSar: A Corpus of Early Modern Sardinian Texts

Nicoletta Puddu
University of Cagliari
Cagliari, Italy
nicoletta.puddu@unica.it

Luigi Talamo
Saarland University
Saarbrücken, Germany
luigi.talamo@uni-saarland.de

Abstract

English. The article introduces the Early Modern Sardinian Corpus (EModSar), a corpus featuring nine manuscripts from the Early Modern Period (16th-17th centuries) written in Sardinian with passages in Catalan and Latin. Manuscripts are encoded according to the TEI-P5 guidelines, annotated for bibliographic, philological and linguistic features and published on-line using TEITOK, a software aimed at combining digital philology and corpus linguistics.

Italiano. Presentiamo EModSar (Early Modern Sardinian), un corpus composto da nove manoscritti della prima età moderna (XVI-XVII secolo) scritti in sardo con inserti di catalano e latino cancelleresco. I manoscritti sono codificati secondo le linee-guida TEI-P5, annotati per caratteristiche bibliografiche, filologiche e linguistiche, e resi disponibili on-line tramite il software TEITOK, che combina le esigenze della filologia digitale con la flessibilità di ricerca degli strumenti della linguistica dei corpora.

1 Introduction

In this paper[1] we present the Early Modern Sardinian Corpus (EModSar: http://corpora.unica.it/TEITOK/emodsar)[2], a historical corpus developed within a more general project whose aim is to describe the linguistic repertoire of Sardinia in the Modern Era[3] (see section 2.1). Our main research question addresses the impact of language contact on Sardinian, and, in order to answer this question, we decided to build a pos tagged and lemmatized corpus covering texts from the 16th to the 17th century which also contains extralinguistic information about the chosen texts (see section 2.2). Moreover, given that our texts are written in Sardinian, but also contain sections written in Catalan and Latin, we wanted to both preserve multilingualism, and also ensure that our corpus tools focused on the linguistic analysis of Sardinian. As Pahta et al. 2018:10 point out, multilingual historical corpora are rarer than monolingual ones, and have not been used extensively in historical linguistics. However "embracing a multilingual approach to language history leads the researcher to look beyond the main language of a text and consider what a holistic overview of all the languages in it reveals about the 'grammar' of non-monolingual writing on the one hand or individual identity or social practice on the other" (Pahta et al. 2018:5). Consequently, we decided to adopt the TEI-P5 guidelines to code our documents in order to accomplish Lass 2004's three desiderata for a proper historical corpus (i.e. "maximal information preservation", "no irreversible editorial intervention", and "maximal flexibility"). On the one hand, the use of TEI-P5 for our corpus allowed editorial choices to be preserved in the text at the philological level, while, all the relevant information could be inserted in the header. The use of a TEI-P5 encoding is not a common standard in historical corpora. As Jenset and McGillivray 2017:125 note, "TEI is not very widely used for historical corpora, where there is a stronger emphasis on linguistic annotation rather than on paleographic and historical markup. However, in the case of historical texts, the information contained in these tags can

[1]For Italian academic purposes only, Nicoletta Puddu was responsible for Sections 1 and 2 and Luigi Talamo for Sections 3 and 4.
[2]The corpus is currently composed of nine manuscripts, for a total of 6495 tokens.
[3]EModSar has been developed under the project *System for developing and annotating a corpus of ancient Sardinian texts*, funded by the Regione Autonoma della Sardegna (*Capitale Umano ad alta qualificazione*, L.R. 7/07, year 2015).

be crucial to the interpretation of the text and should be considered by the language processing tools. [. . .]." A convenient solution is the use of softwares such as TEITOK (Janssen 2016), a tool which can handle both textual mark-up and linguistic annotation. Since our texts have been annotated at three different levels (at the document-level, at the section-level and at the token-level (see section 3.2), queries in EModSar can combine different levels in order to connect linguistic information with extralinguistic information.

2 Language and texts

2.1 Sardinian in the Modern Era

The linguistic repertoire of Sardinia in the Modern Era is largely understudied, but it is extremely interesting since it sees the presence of many different languages within the same period. From 1324 onwards, the kingdom of Aragon gradually took possession of the Island, and as a consequence Catalan became the official language. After the unification of the Kingdom of Aragon with the kingdom of Castile, Castilian began to spread, but Catalan actually remained in use for juridical and administrative purposes, while Castilian became the language of Universities and of the Church (Virdis 2017). Thus, between 1324 and 1720, when the Island was conceded to the House of Savoy and started its process of Italianization, Sardinia was under Iberian domination. However, Sardinian continued to be used in juridical documents of both a public and, especially, private nature particularly in the countryside. Both Catalan and Castilian deeply influenced Sardinian during the Iberian domination.

The Sardinian language of this period is documented through two typologies of documents: literary sources and juridical sources. The Sardinian literati in the Modern Era usually wrote in the dominant languages (mainly Castilian). However, some of them (like Antonio Lo Frasso) inserted some Sardinian sections in their works or even wrote entire compositions in Sardinian (like Girolamo Araolla). What is clear, however, is that all the literati living in Sardinian were highly plurilingual (Marci 2006).

As for juridical documents, Sardinian was used during trial courts, not only for testimonies, but also for other stages of the trial. In private documents Sardinian appears in notary deeds mainly containing sales, donations, debit notes, last wills and testaments (Cadeddu 2013). While we have critical editions of literary texts from the Modern Era, juridical documents are mainly kept in a number of archives in Sardinia. Only a small part of these documents have been published, mainly in historical studies: there are very few critical editions and no systematic linguistic studies.

2.2 The choice of texts

In our project, we wanted to study Sardinian of the Modern era, in the perspective of historical sociolinguistics in Romaine 1992's terms. In order to do so, we decided to create the Early Modern Sardinian Corpus by encoding and annotating juridical documents of the Modern Era, annotated by POS and lemma, and accompanied by contextual information. To date, we have encoded nine documents written in Sardinian dating from the 16th to the 17th century retrieved from the *Archivio storico del Comune di Cagliari* and the *Archivio di Stato di Cagliari*. Most of the retrieved documents come from villages in the Northern Sardinian area and from the towns of Sassari and Bosa. However, we know for certain that documents written in Sardinian datable to those centuries also exist in southern Sardinia. We do not expect to find any documents in Sardinian for the city of Cagliari where Catalan was widespread in all the written domains.

Our documents have presented many problematic aspects typical of historical corpora which we will exemplify by discussing document Osp250 which contains the last will of Canonigu Montixi, the priest of the diocese of Arborea, who, in 1569, leaves a "fellowship" to one of his relatives so that he can study grammar, philosophy and theology.

First of all, our documents are characterized by a high level of orthographic variation, both between different documents and within the same document. For instance, in Osp250 the preposition 'in' can have different orthographic realizations (*in*, *jn*, *en*). Moreover, we have many cases of univerbation, such as *insu* 'in the', *inpodere* 'in power', *etinsu* 'and in the'.

Secondly, our documents are multilingual and we can have code-mixing both at the intersentential level and at the intrasentential level (on different levels of code-switching in historical texts see Kopaczyk 2018). Different codes often correlate with different sections of the document. If we adopt the traditional subdivision in the *formulae* which make up the document, we can see that the *datatio* and the *dispositio* (the core of the document) in Osp250 are written in Sardinian, while the *roboratio testes* and the *completio* are in Catalan. However, we also have intrasentential code-mixing. First of all, as could be expected in juridical documents, we have Latin expressions, such as *ut supra*, *qui supra fidem facio*. But, even more interestingly, we have Catalan and Sardinian code mixing. The *datatio* in Osp250 is in Sardinian, but we find the form *en* for the preposition 'in' , and the name of the month 'June' in the Catalan form *junny*. By contrast, in the *completio*, written in Catalan, the name of the month 'July' is in the Sardinian form *treulas*. Given the close affinity between the different languages present in the document, it is worth noting that, it is not always simple to identify the instances of code-switching, nor to distinguish code-mixing from borrowing.

Finally, our documents are 'stratified', since they have come to us via several passages. Osp250 contains the last will of Canonigu Montixi, but the codicil was redacted by another scribe-priest, Antiogo Molarja. Moreover, the document we have was actually copied by the scribe Sebastià Polla in 1648 at the request of another citizen from Villanovafranca. The document finally arrived in the Archives of the Hospital of Sant'Antonio, since Canonigu Montixi had decided that, were the chain of heirs to die out, his house would have gone to the hospital.

3 Corpus building and annotations

3.1 Corpus building

Due to the mixed nature of our corpus, we needed a software that was able to combine philological aspects i.e., faithful rendering of the manuscripts, bibliographic and historical information with the standard tools used in corpus linguistics i.e., a powerful and flexible query engine. Our choice fell on TEITOK[4] (Janssen 2016), a software developed by Marteen Janssen at the CELTA-ILTEC institute (University of Coimbra, Portugal); in a nutshell, TEITOK is organized in two main components: (i) a web-based application that renders XML files annotated according to the TEI-P5 guidelines and (ii) a suite of executable binaries that convert XML files into the Open Corpus WorkBench (CWB: Evert and Hardie 2011) file format. The first component of Teitok fits our philological needs, as we were able to reproduce our manuscripts with the original page and line breaks, ligatures and graphic variants of linguistic forms (words), while the second component allows us to search our corpus using the Corpus Query Processor (CQP), either from the standard command line facility or using the web application.

Although Teitok is also a powerful XML editor, we employed external XML editors such as oXygen in order to deal with the TEI encoding and annotation processes. Once annotated according to the TEI-P5 guidelines[5], TEI-XML files are uploaded to the web application where they are automatically split into tokens by the Teitok tokenizer. As for the linguistic annotations, Teitok contains some in-development pos-tagging and lemmatization facilities, which have been proven to perform well on historical varieties of languages (Janssen et al. 2017); however, the parts of speech tagging and lemmatization processes, as well as the difficult process of the annotation of graphic variants are all performed manually: at the moment the creation of annotation tools for Sardinian is work in progress (Puddu and Stein 2018) and no annotated corpus is available even for contemporary Sardinian.

Summing up, our corpus building process can be summarized as follows:

1. creation of the XML files: encoding of manuscripts;
2. XML files become TEI-XML files: text annotation according to the TEI-P5 guidelines (TEI header and text elements);
3. automatic tokenization of the TEI-XML files, which are stored in the web application (Teitok);

[4] http://www.teitok.org

[5] The EModSar corpus complies with the latest version of the TEI-P5 guidelines, 3.6.0 released on 16/07/2019. Whenever relevant, we have indicated the URL for the online documentation in the footnotes.

4. manual pos-tagging, lemmatization and annotation of graphic variants.

3.2 Annotations

The annotations featured in EModSar can be conveniently divided into three types: (i) document-level annotation, (ii) section-level annotation and (iii) token-level annotation.

The first type of annotation corresponds to the TEI element known as 'header' and contains bibliographic and, to a lesser extent, linguistic and sociolinguistic information; out of the five principal components described by the TEI-P5 guideline[6], we have compiled the 'file description', the 'text profile' and the 'revision history' components. The 'file description' component[7] contains bibliographic information such as the repository, collection and archival reference of the manuscript, a brief history of the manuscript tradition and the name(s) of the author and copyist. In the 'text profile' component[8], we have gathered information about the place and redaction of the manuscript, the language(s) employed and a summary of the content. As we have pointed out in the previous section, this kind of information is of paramount importance for historical corpora. Finally, the 'revision history' component[9], as the name suggests, works as a change log displaying the date when the TEI-XML file was last changed; the component is most useful during the process of corpus building, which is usually characterized by many versions of the same TEI-XML file, often shared between several collaborators.

Annotations at the section-level are performed within the TEI element known as 'text', which in turn is divided into different sections, marked up by the tag. Note that this text arrangement does not reproduce any formal elements of the original manuscript, but was carried out by the archivist during the encoding process. As mentioned earlier, we decided to mark this structure since it appears to be related to code switching. The tag contains two attributes: the section attribute, describing one of the *formulae* in which a notary document is customarily arranged and the language attribute, giving the language used in the section. For instance, the following text snippet represents the section-level annotation of Osp250, whose *formulae* were mentioned in Sect. 2.2:

```
...
<div n="1" type="datatio" lang="srd" id="div-1"> ... </div>
<div n="2" type="dispositio" lang="srd" id="div-2">...</div>
<div n="3" type="notitia testium" lang="cat" id="div-3">...</div>
<div n="4" type="subscriptiones" lang="cat" id="div-4">...</div>
<div n="5" type="completio" lang="cat" id="div-5">...</div>
<div n="6" type="completio" lang="cat" id="div-6">...</div>
<div n="7" type="dispositio" lang="srd" id="div-7">...</div>
<div n="8" type="completio" lang="cat" id="div-8">...</div>
<div n="9" type="dispositio" lang="cat" id="div-9">...</div>
<div n="10" subtype="dorsale" lang="ita" id="div-10">...</div>
<div n="11" subtype="dorsale" lang="cat" id="div-11">...</div>
...
```

The third type of annotation takes place at the token level and, just like the previous section-level annotation, is implemented through the attributes of the <tok> tag; the tag is not described in the TEI-P5 guidelines and is added by Teitok during the automatic process of tokenization. Each token is annotated for graphic variants and for linguistic information, for a total of five different attributes; as for the graphic variants, we have distinguished between (i) 'written form', corresponding to the graphic variant as found in the manuscript, (ii) 'extended form', which is a written form with expanded abbreviations and (iii) 'normalized form', showing a tentative normalization of the graphic variant. For example, the annotation of the three different orthographic realizations of the preposition 'in', which we have discussed in Section 2.2 is given as follows:

[6] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD1 Last accessed on 23/11/2019.
[7] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD2 Last accessed on 23/11/2019.
[8] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD4 Last accessed on 23/11/2019.
[9] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD6 Last accessed on 23/11/2019.

```
<tok id="w-10" form="en" fform="en" nform="in" pos="PRE" lemma="in">en</tok>
<tok id="w-100" form="jn" fform="jn" nform="in" pos="PRE" lemma="in">jn</tok>
<tok id="w-513" form="in" fform="in" nform="in" pos="PRE" lemma="in">in</tok>
```

As for the linguistic information, we provide annotations for (iv) parts of speech and (v) lemma; the parts-of-speech tagset is an adaptation of the tagset used in the Medieval Sardinian Corpus, which contains texts written in an earlier stage of Sardinian (Puddu 2015, Puddu and Stein 2018), and features 25 tags, some of which are specified for morpho-syntactic properties such as verbal mode and nominal definiteness.

Finally, let us just briefly mention how we handled linguistic expressions - mostly, noun and prepositional phrases - written without spaces between words in the manuscripts. In order to faithfully reproduce the manuscripts, these linguistic expressions are encoded without spaces in the written form of EModSar, with a correspondence between a linguistic expression and a single token; at the same time and for the purpose of linguistic queries, the linguistic expression is split into tokens in the normalized form of our corpus by means of another non-standard TEI tag, <dtok>, which is introduced by Teitok (Janssen 2016:4038) and nested into the <tok> tag. Take for instance the prepositional phrase *in podere*, which was originally written as a single word in one of the manuscripts:

```
<tok id="w-280" form="inpodere" fform="in podere" nform="in podere">
inpodere
<dtok id="d-280-1" form="in" fform="in" nform="in" pos="PRE" lemma="in"/>
<dtok id="d-280-2" form="podere" fform="podere" nform="podere" pos="NOUN"
lemma="podere"/></tok>
```

4 Further developments

In building the Early modern Sardinian Corpus we have already achieved several objectives, summarized as follows:

- we established an annotation schema for Early Modern Sardinian notary deeds which allows all the relevant external information to be preserved;
- we have inserted our documents into Teitok which, not onlymakes it easy to use for different kinds of users, but also permits linguistic searches to be performed with standard corpus tools;
- since the documents will be freely downloadable, they can be re-used for other searches (for instance, personalized queries through XPath, or through other platforms like TXM).

The first studies on the languages used in the documents show the importance of being able to combine linguistic information and extralinguistic information and of considering texts in a multilingual perspective. For instance, we were able to confirm our idea that, some sections in our documents,such as the *completio* and the *subscriptiones*, are generally in Catalan while in others, like the *datatio*, Sardinian alternates with Latin. The use of Catalan and Latin thus seems to be correlated to more "formal" discourse moves and is used to add authority to the document. Moreover, since we also collected extralinguistic information, we were able to correlate linguistic phenomena with different levels of linguistic variation. For example, some of our documents show variants that mantain the original Latin consonant cluster *-pl-/-bl-* (as *complimentu* and *obligare*) while others have the innovative form in *-pr-/-br-* (like *comprimentu* and *obrigare*). Our corpus allowed us to see that the forms in *pr/br* tend to appear in documents which also show some other "lower" phenomena like the methathesis of *-r-* (as in *frimadu* for *firmadu*) and it can consequently be hypothesized that both correlate with diastratic variation.

Future work will focus on two points:

- at a more general level we need to develop the structural coding of more complex documents such as court trials, which arrived in the form of a summary report containing different documents such as letters, trial witness statements, and attestations relative to the delivery of convocations;

- some issues on normalization and lemmatization are still to be discussed, especially if we want to place our corpus in a diachronic and ambitious perspective as one of the steps for the construction of a diachronic corpus of Sardinian.

It goes without saying that, only by increasing the size of our corpus, can we confirm the already noticed tendencies and give a more detailed picture of the multilingual practices in Modern Sardinia.

Acknowledgements

The authors wish to thank Maarten Janssen for his wonderful support on Teitok.

References

Maria Eugenia Cadeddu. 2013. Reperti di plurilinguismo nell'Italia spagnola (sec.XVI-XVII). In T. Krefeld, W. Oesterreicher, and V. Schwägerl-Melchior, editors, *Scritture di una società plurilingue: note sugli atti parlamentari sardi di epoca moderna*, Berlin-Boston: DeGruyter, pages 13–26.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference, Birmingham, UK*.

Maarten Janssen. 2016. Teitok: Text-faithful annotated corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Maarten Janssen, Josep Ausensi, and Josep M. Fontana. 2017. Improving pos tagging in old spanish using teitok. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Linköping University Electronic Press, Linköpings universitet, 133, pages 2–6.

Gard B. Jenset and Barbara McGillivray. 2017. *Quantitative Historical Linguistics*. Oxford: Oxford University Press.

Joanna Kopaczyk. 2018. Administrative multilingualism on the page in early modern Poland. In Päivi Pahta, Janne Skaffari, and Laura Wright, editors, *Multilingual practices in language history*, De Gruyter, Berlin-Boston.

Roger Lass. 2004. Ut custodiant litteras. Editions, Corpora and Witnesshood. In M. Dossena and R. Lass, editors, *Methods and data in English historical dialectology*, Bern: Peter Lang, pages 21–48.

Giuseppe Marci. 2006. *In presenza di tutte le lingue del mondo. Letteratura sarda*. Cagliari: CUEC.

Päivi Pahta, Janne Skaffari, and Laura Wright. 2018. From historical code-switching to multilingual practices in the past. In Päivi Pahta, Janne Skaffari, and Laura Wright, editors, *Multilingual practices in language history)*, De Gruyter, Berlin-Boston.

Nicoletta Puddu. 2015. Costituzione del Sardinian Medieval Corpus: prime proposte per la codifica e l'annotazione. In Piera Molinelli and Ignazio Putzu, editors, *Modelli epistemologici, metodologie della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*, Franco Angeli, pages 282–299.

Nicoletta Puddu and Achim Stein. 2018. Word-level and higher level annotation of the sardinian medieval corpus. In Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, editors, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities)*. Gerastree Proceedings, Vienna.

Suzanne Romaine. 1992. *Socio-historical Linguistics. Its Status and Methodology*. Cambridge: Cambridge University Press.

Maurizio Virdis. 2017. Superstrato spagnolo. In M. Dossena and R. Lass, editors, *Manuale di linguistica sarda*, Berlin: DeGruyter, pages 168–183.

Shared Emotions in Reading Pirandello. An Experiment with Sentiment Analysis

Simone Rebora
University of Verona
University of Basel
simone.rebora@univr.it

Abstract

English. The paper reports on an experiment conducted with a group of students, aimed at verifying the effectiveness of Sentiment Analysis on Italian literary texts. Students were asked to annotate each paragraph of the short story "Ciàula scopre la luna" (1907) by Luigi Pirandello with a numeric evaluation of the sentiment and a free comment. Analysis of the annotations shows how, while inter-annotator agreement is still low, (a) emotional shifts in the story heighten the agreement in sentiment detection; (b) Sentiment Analysis works better for the comments than for the text, thus confirming its efficiency in reader response studies.

Italiano. L'articolo riporta un esperimento condotto con un gruppo di studenti, volto a verificare l'efficacia della Sentiment Analysis sui testi letterari in lingua italiana. Agli studenti è stato chiesto di annotare ogni paragrafo del racconto "Ciàula scopre la luna" (1907) di Luigi Pirandello con una valutazione numerica del Sentiment e un commento libero. L'analisi delle annotazioni mostra come, mentre l'Inter-Annotator Agreement resta basso, (a) le variazioni emotive nella storia aumentano l'accordo nella rilevazione del Sentiment; (b) la Sentiment Analysis funziona meglio per i commenti che per il testo, confermando così la sua efficienza negli studi sulla ricezione.

1 Introduction

Sentiment Analysis (SA) has recently grown in relevance in Digital Humanities. This computational technique, originally developed with the goal of analyzing "people's opinions, sentiments, appraisals, attitudes, and emotions towards entities and their attributes" (Liu, 2015), has found multiple applications in literary studies. From the widely discussed "shapes of stories" by Jockers (2014) and Reagan et al. (2016), to the study of fairy tales (Mohammad, 2012; Rotari, 2018), literary criticism (Rebora, 2017; Mellmann and Du, 2018), genre (Kim et al., 2017; Henny-Krahmer, 2018), and narrative structure (Zehe et al., 2016), SA has become one of the key methodologies in computational literary studies. For an extensive survey, see (Kim and Klinger, 2018). However, criticisms abound, both in theoretical (Ciotti, 2017) and practical (Sprugnoli et al., 2016) terms.
With this paper, I will report on an experiment aimed at verifying the efficiency of the approach in the study of two related phenomena: the narratological structure of a story and its associated reader response.

2 The Experiment: Bringing Research and Didactics Together

The experiment was conducted during the Digital Humanities course (*Informatica per gli studi umanistici*) held at the University of Verona in the academic year 2018/2019. Students were asked to read the short story (novella) "Ciàula scopre la luna" (1907) by Luigi Pirandello and were provided with an XML file with the following structure:

- the <novella> root tag;
- the child <frase>, containing one paragraph from the short story;

- the child <sentiment>, which the students were asked to fill with a numeric evaluation of the sentiment of the paragraph, ranging between -5 and +5;

- the child <commento>, where the students could write a free comment on the effects produced by reading the passage.

Full text of the novella was downloaded from LiberLiber and based on the 1986 Mondadori edition (Pirandello, 1986). Sentences were automatically split using the SA software Syuzhet, that was adopted as a groundwork for the entire experiment[1].
At the end of the annotation process, a total of 51 students wrote at least one comment or sentiment evaluation, for a total of 1,884 comments (36.94 per student) and 1,401 sentiment evaluations (27.47 per student). The 51 XML documents were then anonymized and merged into a single file, available for consultation (together with the R scripts for its analysis) on Github.
The experiment had the didactic purpose of letting students familiarize with the XML markup language and with SA computational techniques (both, in a very simplified form). In terms of research purposes, their annotations proved precious for a verification of the efficiency of SA approaches.

3 Analysis

3.1 Agreement on Sentiment Annotations

Figure 1 shows all sentiment annotations by the students. As evident, annotations are widely spread throughout the 111 paragraphs of "Ciàula", with a dominance of the central levels of emotionality and a deviation towards the most positive levels only at the end of the novella.

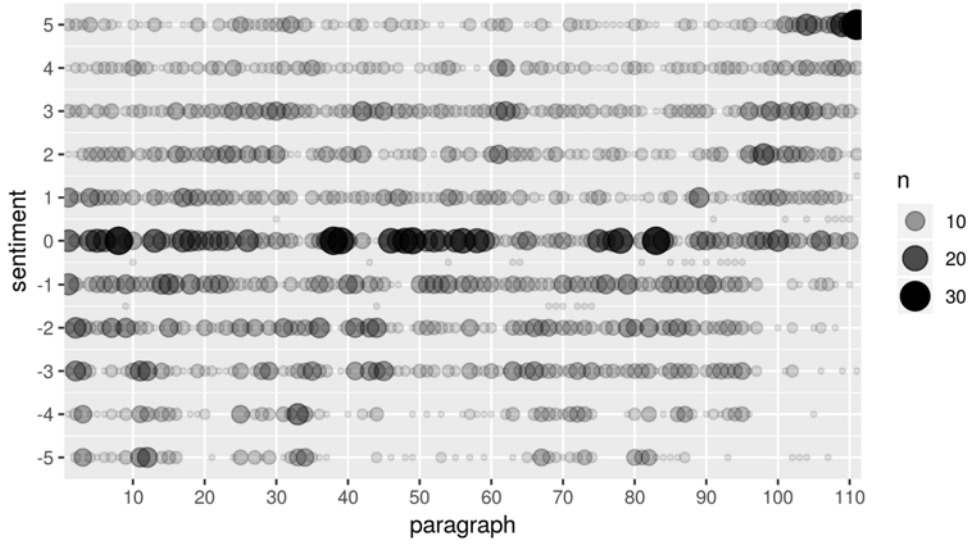


Figure 1: Sentiment annotations on "Ciàula scopre la Luna" (51 annotators: -5/+5)

For a more detailed understanding of the level of agreement, Krippendorff's Alpha (Krippendorff, 2018) was adopted. To maximize the possible agreement, annotations were reduced to a binary selection:

- annotation value < 0: "negative" tag;

- annotation value = 0: no tag;

- annotation value > 0: "positive" tag.

[1]Note that Syuzhet was designed to work on a sentence level (in fact, repeated words do not count towards the total sentiment). This is why annotation was performed on a sentence (and not paragraph) level.

However, Krippendorff's Alpha was still substantially low (0.19), thus confirming the result of Sprugnoli et al. (2016), who showed how inter-annotator agreement advises against the application of SA to historical (or literary) texts.
Given this acknowledgment, still, some interesting outcomes can be derived from the experiment.

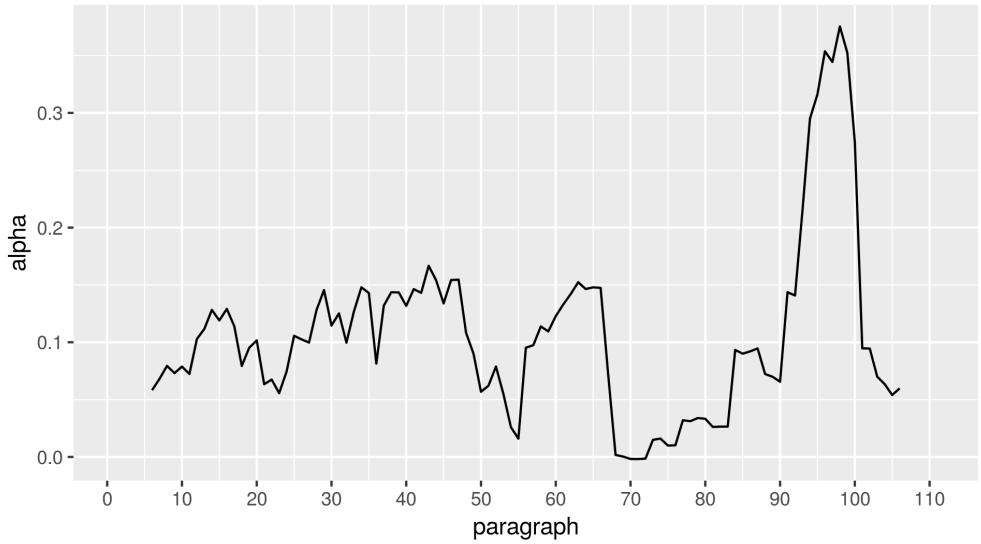


Figure 2: Krippendorff's Alpha for sentiment annotation (51 annotators: POS/NEG; 11-paragraph moving window)

Figure 2 shows the evolution of inter-annotator agreement through a moving window procedure: Krippendorff's Alpha is calculated on just 10% of the text (corresponding to 11 paragraphs), moving from its beginning to its end. The most striking result is in the peak of inter-annotator agreement, that comes not at the very end of the novella (marked by a dominance of positive emotions), but a few paragraphs before. Precisely, it happens around paragraph 98 ("Dapprima, quantunque gli paresse strano, pensò che fossero gli estremi barlumi del giorno"), that signals the beginning of the emotional shift (the "plot twist") in the novella: Ciàula, still fearing the blank darkness of the night, gradually discovers the presence of the moon. This result confirms how an actual sharing of emotions happens not at their climax but with their modification, and transformation—as already noted by Oatley (2012)—is the driving force of narratives.

3.2 Correlation in Sentiment Analysis

SA of text and comments was performed using the simplest method (wordcount) implemented by the Syuzhet package. Being Syuzhet designed for the analysis of English language and given the much more inflected nature of Italian language, analysis was performed on lemmatized texts, that were prepared through the UDpipe software. Two Italian sentiment dictionaries were prepared and uploaded in Syuzhet:

- Sentix, where sentiment values were calculated as the product of polarity and intensity;
- OpeNER (Russo et al., 2016), where sentiment values were calculated as the product of sentiment and confidence.

To keep a direct connection between text and comments, for each paragraph of the novella:

- a single sentiment value was calculated for the text;
- the mean of all sentiment values was calculated for the comments.

Figure 3 shows a comparison between the analyses of text and comments with the two sentiment dictionaries. A reference point (the black, dashed line) was set by calculating the means (per paragraph) of

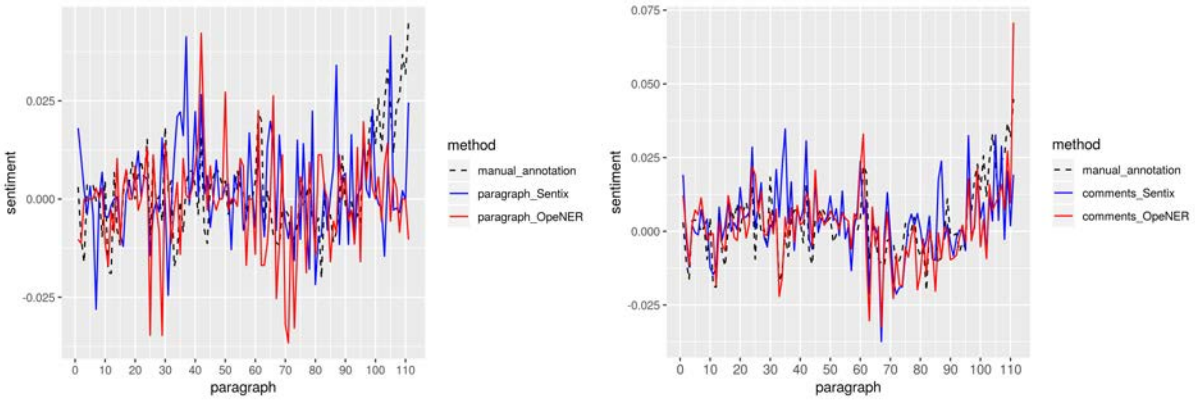


Figure 3: Sentiment analysis of "Ciàula": text (left) and comments (right)

Focus	Sentix	OpeNER
paragraph	0.135	0.266*
comments	0.454*	0.659*

Table 1: Pearson correlations between SA results and mean values of manual annotation. Asterisks indicate significant correlations (p-value < 0.05)

the sentiment values annotated by the students. A mathematical evaluation of the similarity between the plots was provided by Pearson correlation tests. See Table 1 for an overview of the results.
At least two phenomena call for attention. First, OpeNER seems to achieve better results than Sentix. Second, and most importantly, analyses of comments show much higher correlations than analyses of the commented text. This may be considered as a confirmation of the fact that SA is much more effective when studying reader response, than when analyzing narrative structure, as already shown by Rebora and Pianzola (2018). These results become even more striking when applying the "rolling mean" procedure, implemented in Syuzhet to harmonize plots (see Figure 4): here the similarity can be noticed with the naked eye.

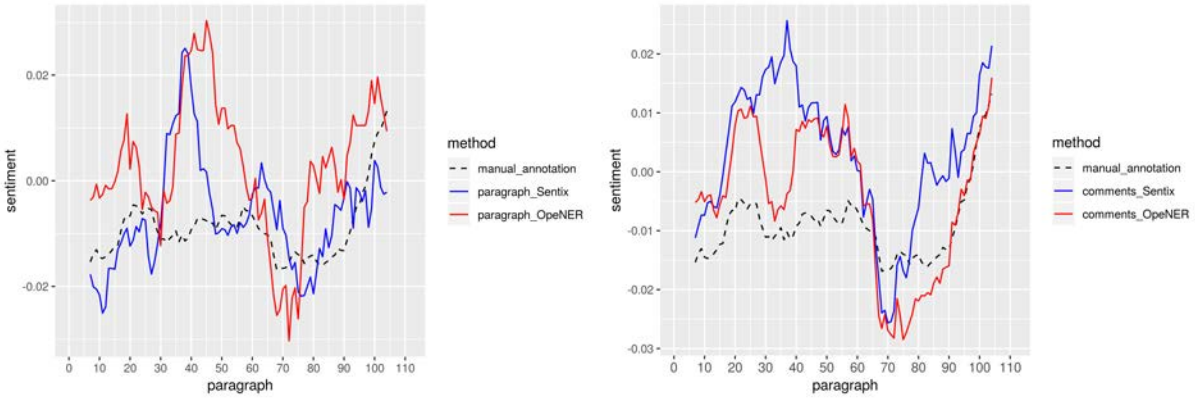


Figure 4: Sentiment analysis of "Ciàula": text (left) and comments (right), normalized with rolling mean

4 Conclusion

The small dimensions of the analyzed corpus call for caution when trying to generalize such results. However, they are in line with evidence already presented in previous studies and they call for new research on the topic. In particular, the high correlation in the SA of comments suggests how, notwithstanding the low agreement between readers when trying to evaluate the sentiment of a text, SA is still able to catch

general trends in reader response. At this point, two main lines of enquiry seem advisable: one, that focuses on improving the methodologies further[2]; another, that tries to tighten the connection between computational methods and literary theory.

In conclusion, while being still a very problematic and disputable technique, SA offers multiple stimuli for theoretical and methodological reflection, revealing how, through a direct confrontation with its limitations and imperfections, research in Digital Humanities can still progress towards unexplored grounds.

5 Acknowledgments

I thank Tiziana Mancinelli for allowing me to perform this experiment with her students.

References

Fabio Ciotti. 2017. Modelli e metodi computazionali per la critica letteraria: lo stato dell'arte. In B. Alfonzetti, T. Cancro, V. Di Iasio, and E. Pietrobon, editors, *L'Italianistica oggi*, Adi Editore, Roma, pages 1–11.

Ulrike Edith Gerda Henny-Krahmer. 2018. Exploration of sentiments and genre in spanish american novels. ADHO, Mexico City. https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/

Matthew Jockers. 2014. A Novel Method for Detecting Plot. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/

Evgeny Kim and Roman Klinger. 2018. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *arXiv:1808.03137 [cs]* ArXiv: 1808.03137. http://arxiv.org/abs/1808.03137

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the Relationship between Literary Genres and Emotional Plot Development. In Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Association for Computational Linguistics, Vancouver, Canada, pages 17–26. https://doi.org/10.18653/v1/W17-2203

Klaus Krippendorff. 2018. *Content analysis: an introduction to its methodology*. SAGE, Los Angeles, fourth edition edition.

Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Katja Mellmann and Keli Du. 2018. Sentimentanalyse in unstrukturierten Texten (am Bsp. literaturgeschichtlicher-Rezeptionsanalyse). Köln, pages 305–308.

Saif M. Mohammad. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems* 53(4):730–741. https://doi.org/10.1016/j.dss.2012.05.030.

Keith Oatley. 2012. *The passionate muse: exploring emotion in stories*. Oxford University Press, New York.

Luigi Pirandello. 1986. *Novelle per un anno*. A. Mondadori, Milano. OCLC: 14516480.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5(1):31.

Simone Rebora. 2017. A Software Pipeline for the Reception of Italian Literature in Nineteenth-Century England. Preliminary Testing. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage (DATeCH)*. ACM, New York, pages 129–134. https://doi.org/10.1145/3078081.3078102

Simone Rebora and Federico Pianzola. 2018. A New Research Programme for Reading Research: Analysing Comments in the Margins on Wattpad. *DigitCult - Scientific Journal on Digital Cultures* 3(2):19–36. https://doi.org/10.4399/97888255181532

Gabriela Rotari. 2018. Digital Analysis of Emotions in the Brothers Grimm's Fairy Tales. EADH, Galway. https://eadh2018.exordo.com/programme/presentation/12

[2]It is undeniable that the method here adopted (that ignores sentiment shifters, does not parse sentences or use machine learning) is quite basic. However, as noted by Kim and Klinger (2018), this is the state of the art for SA in Digital Humanities.

Irene Russo, Francesca Frontini, and Valeria Quochi. 2016. OpeNER Sentiment Lexicon Italian - LMF. http://www.opener-project.eu; https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-73

Rachele Sprugnoli, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. 2016. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities* 31(4):762–772. https://doi.org/10.1093/llc/fqv027

Albin Zehe, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of happy endings in German novels based on sentiment information. In *Proceedings of the workshop on interactions between data mining and natural language processing*. volume 2016, pages 9–16.

DH as an Ideal Educational Environment: the Ethnographic Museum of La Spezia

Letizia Ricci
University of Pisa
l.ricci29@studenti.unipi.it

Francesco Melighetti
University of Pisa
f.melighetti@studenti.unipi.it

Federico Boschetti
CNR-ILC, Pisa &
VeDPH, Ca' Foscari Venezia
federico.boschetti@ilc.cnr.it

Angelo Mario Del Grosso
CNR-ILC, Pisa & University of Pisa
angelo.delgrosso@ilc.cnr.it

Enrica Salvatori
LabCD,
University of Pisa
enrica.salvatori@unipi.it

Abstract

English. The authors present the outcomes of an educational experimentation that took place in the academic year 2018-2019 at the degree course in Informatica Umanistica at the University of Pisa. The first objective of the project concerned the digitization of a corpus of postcards from the period of the First World War owned by the ethnographic Museum of La Spezia "G. Podenzana". The aims of the work are not only the historical study of the corpus, but also the organization of a public history project with the Museum.

Italiano. Gli autori presentano i risultati di una sperimentazione didattica svolta durante l'anno accademico 2018-2019 presso il Corso di Laurea in Informatica Umanistica dell'Università di Pisa. Il primo obiettivo del progetto riguarda la digitalizzazione di un corpus di cartoline del periodo della Prima Guerra Mondiale, di proprietà del Museo Etnografico di La Spezia "G. Podenzana". Gli obiettivi del lavoro non sono solo lo studio storico del corpus, ma anche l'organizzazione di un progetto di Public History con il Museo.

1 Introduction

The authors present the outcomes of an educational experimentation that took place in the academic year 2018-2019 at the degree course in Informatica Umanistica at the University of Pisa. The experimentation involved the courses of Digital Public History, Digital Text Encoding as well as Digital Philology, and at the beginning concerned the digitization of a corpus of postcards from the period of the First World War owned by the ethnographic Museum of La Spezia "G. Podenzana".

The postcards have been historically contextualized, digitized, placed on a collaborative web platform and distributed to the students in order to be recorded, transcribed and encoded in XML-TEI. Students have been involved in the development of the web platform by tracking the usability issues as beta-testers. Students contributed also to the requirement analysis and the definition of the specifications necessary to extend the platform to a broader audience of users without specific skills in Digital Humanities.

Indeed, the project aims were not only the historical study of the corpus, but also the organization of a public history project with the Museum, its targeted audience and the High School students of La Spezia. Arriving almost at the end of this educational experiment, we propose now to discuss the current achievements, based on the common educational statement of "learning by doing" and announced months ago at AIUCD 2019.

2 Background

Within the previous annual conference of the Italian Digital Humanities Association held in Udine (AIUCD2019), a preliminary work towards a profitable collaboration between students and teachers of different DH classes at the University of Pisa was presented. The context of the project (Booth, 1996; Cati, 2006; Cole, 2016; Delle Cave, 2013) gave the actors the opportunity to collaborate with each other and with the "G. Podenzana" museum of La Spezia outside the formal classroom constraints, involving also some activities carried out by non academic communities (Salvatori, 2017). In Salvatori et al. (2019), the authors discussed the objectives and the outcomes that have been achieved during the bootstrap phases of the project. In particular, within that paper, they pointed out the main problems and the added values of the initiative introduced above.

To date, the collaboration has been getting wider and a few internships have been activated to improve the design and development skills of the interested students in order to enhance the tools already developed within the *Euporia* platform (Mugelli et al., 2016). These new activities facilitate students and general users in digital encoding historical documents, which have an inherently complex nature. This objective has been possible by adopting a formal but "common" rules for annotating and for processing textual data (Fowler, 2010). Moreover, as far as the actual TEI-XML encoding work (Burnard, 2014; Pierazzo, 2015) which concerns the digitization, the recording and transcription of the postcards provided by the involved cultural institutions, the main problems have been overtaken by putting in place a chain of document processing tasks. This process has been developed by using XSLT technology and by implementing a Web environment to publish the encoded documents (Del Turco and Di Pietro, 2016).

3 Methods

The project involves students, teachers and representatives of the Ethnographic Museum of La Spezia, which collaborate by sharing information, resources and tools. The Museum has made available two large corpora of postcards dating back to the Great War period, which have been digitized, uploaded onto the *Euporia* platform and encoded in XML-TEI by the students under the supervision of the teachers, according to a custom subset of tags declared in the *CartolineXML* schema. Furthermore, the Museum played the role of an interdisciplinary meeting center among High School students of La Spezia, students and teachers of the University of Pisa and the Museum managers, in order to share ideas about the project from different perspectives.

Two students of the aforementioned courses and co-authors of this contribution, made an internship at the CNR-ILC to improve their skills in text encoding. They focused on the simplification of the annotation process, in order to involve High School students and volunteers that could actively collaborate in transcribing and annotating postcards, even if they have not specific skills in XML-TEI encoding.

Therefore, they have defined a Domain Specific Language (DSL), *CartolineDSL*, with the same expressivity of its counterpart in XML-TEI but much less verbose. A DSL is a formal language with a simple, understandable and suitable syntax for the domain of interest we are dealing with and based on a limited and controlled vocabulary. A DSL must be defined by a Context-Free Grammar (CFG), that is a set of recursive rewriting rules used to generate string patterns. Therefore *CartolineDSL* is a language suited to the domain of postcards characterized by a series of "*attribute: value" fields that users can easily fill in.

```
doc: title description? body note;
title: TITLE_HEADER text;
description: DESCRIPTION_HEADER figure? notes?;
body: bodyText recipient;
note: NOTE_HEADER text?;
bodyText: TEXT_HEADER opener letterbody closer notes?;
recipient: RECIPIENT_HEADER address notes?;

opener: date placeName? initialSalute?;
closer: finalSalute? signed;
notes: NOTES_TAG text?;

figure: FIGURE_TAG text;
date: DATE_TAG dateValue;
placeName: PLACE_NAME_TAG text;
initialSalute: INITIAL_SALUTE_TAG text?;
finalSalute: FINAL_SALUTE_TAG text?;
letterbody: LETTERBODY_TAG (text|linebreak)+;
signed: SIGNED_TAG text;

address: persName street city province?;
persName: NAME_TAG text;
street: STREET_TAG text;
city: CITY_TAG text;
province: PROVINCE_TAG text?;
dateValue: DATE;

linebreak: '//';
```

Figure 1: Formal grammar code snippet for the postcard corpus within the *Euporia* digital environment

Fig. 1 illustrates some rewriting rules in the CFG of *CartolineDSL*, whereas Fig. 2 shows an example of data and metadata encoded in *CartolineDSL*.

```
===TITOLO===
Sommeil Interrompu

===DESCRIZIONE===
*immagine: Una donna interrompe il sonno di un uomo che si riposa vicino ad un albero.
*note: Sono presenti tre timbri e un francobollo.

===TESTO===
*data: 24/07/1913
*luogo: Ravenna
*saluto:
*corpo: A Ravenna piove sempre: è una gioia per chi non è andato ai bagni. Per ora niente di nuovo, tutti bene. Stasera qui c'è musica, come ieri che fu S.Apollinare: ad Arona, niente
*commiato: Saluti affettuosi,
*firma: Giuseppe
*note:

===DESTINATARIO===
*nome: Signorina Oliva Turtura
indirizzo: Via Cavour 12
*località: Arona (Lago Maggiore)
*provincia:
*note:

===NOTE===
La cartolina si presenta in un buono stato di conservazione.
```

Figure 2: *CartolineDSL* snippet

The conversion of *CartolineDSL* to XML-TEI is performed in two steps. In the first phase, the annotations encoded in *CartolineDSL* are parsed by the ANTLR compiler compiler (Parr, 2013) and converted in XML. The proprietary *CartolineML* schema allows the serialization in XML of the Abstract Syntactic Tree (AST) parsed by ANTLR. In the second phase, the proprietary XML document is converted to XML-TEI by an XSLT transformation. The XSLT style-sheet has been created by the students on the latest part of their internship at the ILC-CNR.

As usual, other XSLT style-sheets are necessary to transform XML-TEI in HTML for visualization purposes. Fig. 3 shows the designed interface.

Figure 3: Mockup sketch of the ongoing web-app aimed at publishing the archive

4 Results

The main achievements of this didactic experimentation are listed below: 1. coordination of three courses, in order to work on the same materials from different perspectives (Public History for the historical contextualization of the project, Text Encoding for XML-TEI models and technologies, and Digital Philology for the treatment of uncertain readings and for the creation of an optimized human readable DSL); 2. engagement of students in the annotation process of a large sample of the postcards corpus (learning by doing); 3. transfer of knowledge and experience from students of the University of Pisa and High School students of La Spezia during meetings at the Museum; 4. involvement of students in the creation of Domain-Specific Languages meant to bridge the gap between the best practices of Digital Humanists and the simple practices of unskilled citizens that desire to participate in projects of Public History.

5 Conclusion and Future Work

We guess that the educational model that we experimented can be easily exported in other contexts, with a broader involvement of multidisciplinary communities of practice and applied to different textual and/or iconographic (or multimedia) digital resources.

In the next academic year we will release an updated version of *Euporia*, which currently is just a prototype, in order to allow High School students and volunteers to annotate further postcards in *CartolineDSL*.

References

A. Booth. 1996. *Postcards from the Trenches: Negotiating the Space Between Modernism and the First World War*. Oxford University Press.

L. Burnard. 2014. *What is the Text Encoding Initiative? Encyclopédie numérique 3*. OpenEdition Press.

I. Cati. 2006. *Cara mamma ti scrivo: le cartoline dei soldati della grande guerra*. Gasparri.

K.J. Cole. 2016. *Postcards from the Front 1914-1919*. Amberley Publishing Limited.

R. Del Turco and C. Di Pietro. 2016. *Between innovation and conservation: the narrow path of UI design for the DSE*, University of Graz, Graz.

L. Delle Cave. 2013. *Orme di guerra : lettere e cartoline dal fronte (1912-1919)*. Samus.

M. Fowler. 2010. *Domain-Specific Languages*. Addison-Wesley Signature Series (Fowler). Pearson Education.

G. Mugelli, F. Boschetti, R. Del Gratta, A.M. Del Grosso, F. Khan, and A. Taddei. 2016. A user-centred design to annotate ritual facts in ancient greek tragedies. *BICS* 59(2):103–120.

T. Parr. 2013. *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf, 2nd edition.

E. Pierazzo. 2015. *Digital Scholarly Editing : Theories, Models and Methods. Digital Research in the Arts and Humanities*. Ashgate, Farnham Surrey.

E. Salvatori. 2017. Digital (public) history: la nuova strada di una antica disciplina. *RiMe* 1-I:57–94.

E. Salvatori, F. Boschetti, and A.M. Del Grosso. 2019. From collaborative transcription to interdisciplinary education: the postcards of the great war case. In *AIUCD2019 Book of Abstracts*. Udine.

A Digital Review of Critical Editions: A Case Study on Sophocles, *Ajax* 1-332

Camilla Rossini
Università di Genova - DIBRIS
camilla.rossini@studenti.unige.it

Abstract

English. The paper describes a framework for publishing reviews of critical editions of classical works in a digital environment. After an account of its advantages over 'traditional' reviews, the paper outlines its modelization, realization, and criticalities. Finally, some possible developments are listed.

Italiano. Nell'articolo si descrive un modello per pubblicare recensioni di edizioni critiche in un ambiente digitale. Dopo un'analisi dei vantaggi rispetto alle recensioni 'tradizionali', se ne espongono le fasi di modellizzazione e realizzazione e gli elementi di difficoltà. Infine, si elencano possibili sviluppi futuri.

In the next pages I will give an account of a still ongoing project conducted at the University of Leipzig under the supervision of professor G. Crane. The project aims at modelling a framework for publishing reviews of critical editions of classical works in a digital environment (*Smart Reviews* - SR), thus re-thinking the review genre at its roots. For a first experimental mock-up, the chosen case study is Sophocles, *Ajax*, 1-332. The sequential steps we designed are: 1) to select from two or more editions some noticeable readings; 2) to link them to the corresponding places in a text chosen as a base reference, with an unambiguous reference system; 3) to compare them with the aid of external tools, in order to explain the editorial choices behind them. After a paragraph on the purposes, the uses and the shortcomings of 'traditional' reviews on paper, I will proceed to show in further detail the advantages of a SR and the passages to realize it. Finally, I will list some future developments.

Whenever a new critical edition of an ancient text is published, other scholars carefully read it, compare it to previous texts and finally publish reviews of it on academic journals. Besides overall judgments on the edition's quality, bibliographic suggestions and further comments on specific editor's remarks, a review of a critical edition usually provides an account of the most noticeable editorial choices on the text. Textual renditions of controversial readings, new conjectures or the recovery of old ones, and maybe the comparison with the latest edition(s) on some crucial passages, are what really defines the work of the editor on the text itself, and are thus the ultimate object of the reviewer's judgment.

The reviews of critical editions, finally, play an irreplaceable role for the users as well. Not only are they often, at a practical level, the only way to access a new edition in the absence of it, while waiting, for example, for University libraries to purchase it; even more importantly, they provide a list of the differences between critical texts in different editions, thus saving the time for the reader to detect them by manually comparing two or more printed books.

Nevertheless, such important tasks in this kind of reviews are, at least, hard to perform on a less-than-abstract level. An example will explain why:

> Finglass often succeeds in defending transmitted text: he agrees with OCT against Dawe's Teubner in about 22 cases (for example 446, 771, 782, 790, 988, 1027, 1059, 1282, etc.), the reverse occurring about 15 times (for example 114, 191, 420, 630, 1357, etc.)[1].

This passage, from a review to Finglass' 2011 edition of Sophocles' *Ajax*, is just one of many similar ones. Finglass' work is compared with the two previous major editions (Lloyd-Jones – Wilson's and

[1] Catrambone, 2013, p.169 on Finglass, 2011.

Dawe's[2]), but only some specimens of agreement or disagreement are quoted, and for each of them the mere verse number is provided. The job of finding out where and how the three editions are unanimous or less so, is for the reader to do. Of course, the limited space of a review requires conciseness, and an extensive - rather than intensive - approach.

Such shortcomings are intrinsically linked to the printed (or printed-like, for the PDF distributed journals) format that the review articles have had so far. The main contents of a review, though, can be described as links between corresponding passages in different editions. In such a way, they could perfectly support a digital metamorphosis of the genre. Moreover, a fully digital distribution (what we could call a *Smart Review*, SR) could provide more effective comparisons between editions, and links to external resources could give the reader insights on the editors' choices. This way, not only the old, consolidated tasks of the 'traditional' reviews are performed better and in a more feasible way; but what is more, a SR could improve and widen the usefulness of the reviewing and comparison on multiple editions[3].

This shift in perspective is even more desirable if we think about the Scholarly Digital Editions (SDEs). More and more as we move on, new versions of the same ancient texts become available online: not only as scanned out-of-copyright editions, but also as new uploads in large online repositories for plain or annotated texts, like treebanks[4]. Even though the sense to assign to the expression 'SDE' is controversial, each of those new documents bears a specific version of a text that becomes available to a large public; moreover, both the digital-born plain texts and the linguistic annotated ones, often imply a critical revision by the digital editor. Unfortunately, the communication problems that have been acknowledged between printed editions, stand for digital publications as well. The artificial sense of fixedness of each of those 'base texts' is often reinforced by the absence of critical apparatuses, that flattens the editor's opinions and textual decisions in favour of a totally illusory objectiveness. It has been said that the technical possibility to publish all the witnesses and all the editions would lead to «a sort of 'Bédier effect'»[5], where everyone publishes an edition or a witness without establishing a critical text.

Although the study of single editions or single manuscripts can have great applicability in many fields, the differences among SDEs (broadly intended) is often underaddressed, and a great number of divergent passages remains unnoticed. This problem becomes even more visible when translations are involved. Not infrequently, the translations are made available online without their corresponding original text, making it difficult to address and explain the textual choices behind them[6]. To sum up, each digitally published text is liable of becoming an arbitrary base text.

The idea behind a SR is the opposite. Its goal is to show the diverging readings in traditional or digital editions by juxtaposition, thus not necessarily stating a hierarchy between them, similarly to what happens in traditional reviews. It is true, though, that we can not do without a base text to anchor each reading to its proper position, because a section of the text where the two or more editions diverge doesn't have, by its definition, a *lemma* to unequivocally refer to. Thus, the first criticality to address is the need to provide an unambiguous anchoring of the noticeable readings. The most frequently implemented solution, the XML AppCrit module, is not suitable for our purpose. Firstly, it has a binary (and thus, hierarchical) distinction between lemma and reading. Secondly, a core need of a SR is to be flexible, updatable, reusable, and for those necessities a standoff markup seems like a better choice[7].

[2] Lloyd-Jones and Wilson, 1994; Dawe, 1996.

[3] Gabler, 2010.

[4] See Crane et al., 2014. On editorial interventions on treebanks see *e.g.* Bamman et al., 2009, 10: «A scholarly treebank [...] reflects an interpretation of a single scholar». On textual variation and ambiguity in treebank annotation see also Bamman and Crane, 2010, p. 548; Beaulieu et al., 2012, p. 400.

[5] Bartoli, 2015. In 1928, J, Bédier suggested that, as the Lachmannian method was practically unreliable, a single witness (*codex optimus*) should be chosen and edited. See Bédier, 1928.

[6] A basic example will show it. Accessing Soph., *Aj.* 35 on Perseus, one will find: σῇ κυβερνῶμαι χερί ('hand'). The corresponding English translation perfectly matches the text: «it is your hand that steers me». Oppositely, if we take Romagnoli, 1926, whose Italian translation is freely available *e.g.* on Wikisource, we read: «il senno tuo per guida io prenderò», whic translates as «I will always take your *wisdom* as a guidance», and not «your hand». Poetic license? No, only a *varia lectio* that is recorded in most editions. The tradition is divided between χερί and φρενί. Finglass, 2011, 80 chooses the former, Dawe, 1996, 3, the latter.

[7] See the fundamental benchmark of the database of latin texts by the Digital Latin Library (LDLT, 2019) that, in a much wider perspective, modified the XML TEI P5 module 12 for Critical Apparatus (Guidelines, 2019) for its own purposes (Cayless

For these reasons, I tokenized and corrected an OCRed file of Pearson's 1922 out-of-copyright edition[8]. From this, I provided an automatically compiled list of references to each word, with unique identifiers (see fig. 1). To do so, my benchmark has been the CTS URNs model as implemented by the Perseus Catalog[9]. Each work in the Perseus Library (and in the new Scaife Viewer as well) has a string that identifies it. For example, the greek edition of Soph., *Aj.* 1-332 is referenced by *urn:cts:greekLit:tlg0011.tlg003.perseus-grc2:1-332*, where *tlg0011* and *tlg003* are the traditional codes assigned by the TLG project respectively to the author Sophocles and to the work *Ajax*, and *perseus-grc2* identifies the edition digitized by the Perseus team. The reference goes as far as pointing at a verse or a group of verses (in the example above, verses 1-332). Basing on the work already done on texts from the Perseus Digital Library and the *First Thousand Years of Greek Project*, I extended the unique reference system down to the word level[10]. Thus, each word has an identifier with this ideal structure:

urn.soph.ajax.pearson@134Τελαμώνιε[1]

Firstly conventional abbreviations of the author, the work, and the edition are listed, separated by a mark (I used a dot); then, after an @, the verse and the word are reported and, finally, a number between square brackets that indicates the occurrence of the same word form in that verse. This formulation of the CTS URN is totally conventional. For our purposes here, it could be cited also in its abbreviated form: 134Τελαμώνιε[1].

```
<?xml version="1.0" encoding="utf-8"?>
<cts_group id="urn.ajax.pearson">
    <cts id="1">urn.ajax.pearson@1Αεὶ[1]</cts>
    <cts id="2">urn.ajax.pearson@1μέν[1]</cts>
    <cts id="3">urn.ajax.pearson@1,[1]</cts>
    <cts id="4">urn.ajax.pearson@1ὦ[1]</cts>
    <cts id="5">urn.ajax.pearson@1παῖ[1]</cts>
    <cts id="6">urn.ajax.pearson@1Λαρτίου[1]</cts>
    <cts id="7">urn.ajax.pearson@1,[2]</cts>
    <cts id="8">urn.ajax.pearson@1δέδορκά[1]</cts>
    <cts id="9">urn.ajax.pearson@1σε[1]</cts>
```

Figure 1: A section of the CTS file from Pearson's edition, referencing Soph., *Aj.* 1: Ἀεὶ μέν, ὦ παῖ Λαρτίου, δέδορκά σε. Note the *[2]* in the cts with id 7, that denotes the second comma in the same verse.

I then divided the material into four sections: an ordered list with vv. 1-332 of Pearson's edition, where each word is assigned with such a CST URN reference (see fig. 2a); a database containing the noticeable readings found in the editions under analysis, and their position in reference to file 1 (see fig. 2b); another database containing the matches between each edition and the readings that could be found in it[11] (see fig. 2c); finally, in another database, the broadly meaning commentary material has been linked to the corresponding readings (see fig. 2d).

Linking the noticeable passages of each edition to the correct unit of text is not an easy matter. I came up with a conventional set of rules. I considered lexical substitutions, additions, subtractions and movements. For each of them I had to keep in mind that both the reading and the referenced passage could be formed by one word (see fig. 3a) or by a group of words (see fig. 3c). To each reading I added two attributes: *from* and *to*. They respectively mark the point in the CTSized text where the variant begins and ends; if they coincide, it means that the reading modifies only a word in the base CTSized

and Huskey, 2018). See also the XML structure of the *Euripides Scholia Project* (Mastronarde, 2010), whose editor chose not to use the TEI module for the Critical Apparatus «because in a project of this kind it seems to me that it would involve an unjustifiably large overhead of markup». About it, see Driscoll and Pierazzo, 2016, 213. For a theoretical comparison between inline and standoff markup see *e.g.* Schmidt, 2012; Eide, 2014; Petersen, 2016; Boschetti, 2007; Monella, 2008. For an overview of the criticalities of the XML TEI module 12, see the report issued by the Critical Apparatus Workgroup (Workgroup, 2014).

[8] Pearson, 1924.

[9] See the usage of CTS URNs and the Cite Architecture, both developed by the Homer Multitext project, by the Perseus Catalog. See Blackwell and Smith, 2014; Babeu, 2015; Blackwell and Smith, 2019b; Architecture, 2019; Tiepmar and Heyer, 2019; Blackwell and Smith, 2019a; Babeu, 2019.

[10] See Celano, 2017 on texts taken from the Perseus Digital Library (Perseus, 2019a,b) and the *First Thousand Years of Greek Project* (OGL, 2016). See also the new *Scaife Viewer* (Perseus, 2019c).

[11] Thanks to this organization of the material, I reduced the redundance to much less than if, say, I had to list the noticeable readings for each edition.

```
<cts id="247">
    urn.ajax.pearson@35χερί[1]
</cts>
```

(a)

```
<rdg id="rdg5"
    from="urn.ajax.pearson@0035χερί[1]"
    to="urn.ajax.pearson@0035χερί[1]">
    φρενί
</rdg>
```

(b)

```
<group id="Dawe1996">
    <choice ref='rdg2'/>
    <choice ref='rdg3'/>
    <choice ref='rdg5'/>
    <choice ref='rdg6'/>
```

(c)

```
<comment source="Dawe1996" ref="rdg5">χερί is a "literal-minded
substitute"</comment>
<comment source="Finglass2011" ref="rdg5">φρενί looks like a
correction</comment>
```

(d)

Figure 2: A reading (**b**) linked to its initial and final CTS URNs (**a**) and chosen in Dawe's edition (**c**), with comments on it by Finglass and Dawe (**d**).

```
<rdg id="rdg1"
    from="
    urn.ajax.pearson@0008εὔρινος[1]"
    to="
    urn.ajax.pearson@0008εὔρινος[1]">
    εὔρινος
</rdg>
```

(a)

```
<rdg id="rdg2"
    from="urn.ajax.pearson@0028.[1]+1"
    to="urn.ajax.pearson@0029καί[1]-1">
    <movement
        from="
        urn.ajax.pearson@024κάγὼ[1]"
        to="urn.ajax.pearson@024.[1]"
    />
</rdg>
```

(b)

```
<rdg id="rdg26"
    from="urn.ajax.pearson@0191ὄμμ'[1]"
    to="urn.ajax.pearson@0191ἔχων[1]">
    ἐμμένων
</rdg>
```

(c)

```
<rdg id="rdg61"
    from="urn.ajax.pearson@0243,[1]"
    to="urn.ajax.pearson@0243,[1]">

</rdg>
```

(d)

Figure 3: Types of variation. Interpretive (**a**), movement (**b**), substitution (**c**), subtraction (**d**).

text. This method works fine for substitutions (see fig. 3a[12]). For subtractions as well, it was enough to clearly show the reading as empty (see fig. 3d).

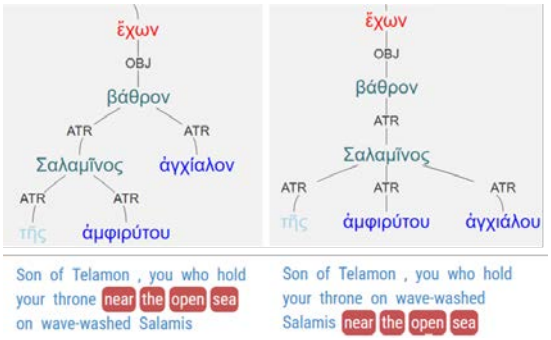
In the case of the word(s) addition, one needs to use a clear way to show it. I pointed at the space between two words by using the conventional formula 134Τελαμώνιε[1]+1 (to refer to the position after the word Τελαμώνιε) or 134Τελαμώνιε[1]-1 (to refer to the position before it). Finally, movements have been pointed at with the self-closing element *movement* (see fig. 3b, that also shows the use of +1 and -1).

This system has multiple advantages: in the first place, it becomes machine-inferable (but quite clear to the human reader as well) where and how each edition differs from the chosen base text, and from each other. The material is kept separate and clean, with an easy way to add, change and modify parts of it without having to alter the structure of the existing files. Moreover, the overlapping of variants becomes possible without complex systems as it is in the XML TEI. The basic types of intervention adopted by each edition can be easily inferred by an algorithm, by comparing the reading with the *from* and *to* attributes and, if necessary, by directing the reader to the comments (see footnote 12 about fig. 3a). Whatismore, in the exact same way as a group of readings is connected to an edition, other groups may be figured out and collected under specific types that go beyond the core distinction between orthographic,

[12] When the reading is identical to the 'base text', the comment material could tell us if the word is listed as a variant because it is a homograph - like in the case of fig. 3a - or because it is just an interpretive variant on the same word form.

```
<!-- Treebanks of Finglass' edition -->

<comment id="Fin_treebank1"
    ref="rdg6"
    from="urn.ajax.pearson@134Τελαμώνιε[1]"
    to="urn.ajax.pearson@136·[1]">
    https://www.perseids.org/tools/arethusa/
    app/#/perseids?chunk=1&amp;doc=63111
</comment>
```

(a)



(b)

Figure 4: Link to treebank (**a**). Treebank and aligned translation: Finglass *versus* Pearson (**b**)

morphological and lexical variants that is provided, for example, by the Digital Latin Library[13].

Another advantage of a SR is that it can point to external sources in order to give the reader insights about the differences between texts. The variants chosen by each editor alter the surrounding text in different ways. Some of them may generate syntactic differences, some other may remain on the lexical level. Finally, other variants are only due to different interpretations, and don't affect the texts themselves, but are only visible in the translations. Through the 'comments' section, the available online tools can be linked to specific passages in the considered editions to show these differences.

For the variants that have an impact on the morphology and the syntax, links to their treebank annotation and graphical visualization on the Arethusa Treebank Editor can be provided in the 'comment' database. In this experimental case, the treebanks for each critical edition have been compiled using as a base the file uploaded by the Ancient Greek and Latin Dependency Treebank project[14]. The comparison between treebanks of corresponding passages in different editions makes us able to encode precisely the difference between editorial choices. Finally, not all variations affect the translation. For the ones that do, links to parallel translation alignments can be provided[15] (see fig. 4).

A theoretical framework for a SR addresses, on the one hand, some problems that are known to the long-lasting debate over Scholarly Digital Editions (SDEs) and, more broadly, to the field of annotated texts. The comparison between editions, core element of the SR, urges to find a way for handling the textual variation in a digital fashion, i.e. to represent variants and to link them to the base text, which is itself the object of a dispute[16]. On the other hand, though, the SR's intrinsic differences from SDEs compel us to find new solution. The main distinction is probably the programmatic desultoriness of the provided data. Only the important readings, and not all the text as in SDEs, are named in 'printed' reviews, hence the same principle should apply to SRs as well.

A model for a SR, besides being a useful improvement of the current printed reviews, can prove to be a valid testing ground for the cooperation and co-existence of various instruments to annotate and encode different features of the texts that are edited in critical editions. Moreover, such a model proves once again that 'linguistic' instruments such as the treebank annotation can and should be integrated into strictly speaking philological resources, as precious means to gain a better understanding of the text and the critical editors' choices[17]. Finally, the possibilities offered by the SR to its users would increase significantly from those of a traditional review, in what we could call a re-purposing of a known instrument through digital means. At the same time, though, its final goal of helping the reader in assessing the degree of innovation or conservativity of an edition, and in evaluating specific editorial choices, would

[13] See the LDLT Guidelines (Cayless and Huskey, 2018). One could group together, *e.g.*, variants that affect the translation or the staging, or particular types or variants according to one's specific needs.

[14] See Alpheios, 2019. For the Guidelines for Greek Treebanking see Celano, 2014. See also Celano and Crane, 2015; Celano, 2019.

[15] I used Ugarit, 2019.

[16] On the base text see *e.g.* Andrews and Macé, 2013, p. 506. About variants see *e.g.* Boschetti, 2007; Monella, 2012; Lana et al., 2017.

[17] See Berti, 2019; Passarotti, 2019; Mambrini, 2016; Beaulieu et al., 2012; Bamman et al., 2009.

not be altered; quite the opposite, they might be enhanced.

From this starting ground, some crucial points need to be addressed. The connections traced between readings, base text and editions could be properly defined semantic. Should the path of semantic annotation be embraced more fully, by developing an ontology[18]? What can (or should) the role of automated processes both in variant detection and in word analysis be[19]? What can the visualization and the dissemination of the project be? Which platform will best suit the open source paradigm? The previous pages only provided a first, experimental model that is still under development and that may take various directions. As for now, my hope is that this paper might provide some additional discussion material for some long known questions, more than answers to those very doubts.

Acknowledgements

I would really like to thank the members of the Department of Digital Humanities at University of Leipzig and especially professor G. Crane, professor M. Berti and professor T. Köntges for their patient teaching and advising throughout my months as a visiting PhD student.

References

Alpheios. 2019. The alpheios project. https://alpheios.net/pages/tools/

Tara L. Andrews and Caroline Macé. 2013. Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing* 28(4):504–521.

CITE Architecture. 2019. http://cite-architecture.org/

Alison Babeu. 2015. CTS URNs and Work Identifiers: Overview and Perseus Catalog Usage. https://github.com/PerseusDL/catalog_pending

Alison Babeu. 2019. The perseus catalog: of frbr, finding aids, linked data, and open greek and latin. In Monica Berti, editor, *Digital Classical Philology*, De Gruyter, pages 53–72. https://doi.org/10.1515/9783110599572-005

David Bamman and Gregory Crane. 2010. Corpus linguistics, treebanks and the reinvention of philology. *INFORMATIK 2010. Service Science. Neue Perspektiven für die Informatik. Band 2* .

David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)*. pages 5–15.

Elisabetta Bartoli. 2015. The Mechanic Reader. Digital Methods for Literary Criticism. *Semicerchio. Rivista di poesia comparata* LIII(2):125–135.

Marie-Claire Beaulieu, Francesco Mambrini, and Matthew Harrington. 2012. Toward a Digital Editio Princeps. Using Digital Technologies to Create a More Complete Scholarly Edition in the Classics. In *Lire demain Des manuscrits antiques à l'ère digitale. Reading Tomorrow From Ancient Manuscripts to the Digital Era*, Presses polytechniques et universitaires romandes.

Monica Berti. 2019. *Digital Classical Philology, Ancient Greek and Latin in the Digital Revolution*. De Gruyter Saur, Berlin, Boston. https://doi.org/10.1515/9783110599572

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory R Crane. 2014. The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. volume 2014, pages 1140–1147.

Christopher N. Blackwell and Neel Smith. 2014. Specification of the Canonical Text Services protocol (CTS). https://github.com/cite-architecture/cts_spec

Christopher N. Blackwell and Neel Smith. 2019a. The cite architecture: a conceptual and practical overview. In *Digital Classical Philology*, DeGruyter - Saur, pages 88–101. https://doi.org/10.1515/9783110599572-007

[18] Romanello et al., 2009; Tomasi et al., 2015; Andrews and Macé, 2013; Ciotti and Tomasi, 2016; Oren et al., 2006.
[19] See Bizzoni et al., 2014; Boschetti, 2007; Passarotti, 2006.

Christopher N. Blackwell and Neel Smith. 2019b. *Homer multitext project*. http://www.homermultitext.org/about/

Federico Boschetti. 2007. Methods to extend Greek and Latin corpora with variants and conjectures: Mapping critical apparatuses onto reference text. In *Corpus linguistics*. pages 1–11.

Joseph Bédier. 1928. La tradition manuscrite du "lai de l'ombre": réflexions sur l'art d'éditer les anciens textes. Romania 54(214):161–196.

Marco Catrambone. 2013. Finglass P.J. Ed. Sophocles. Ajax: Edited with Introduction, Translation, and Commentary (Cambridge Classical Texts and Commentaries 48). Cambridge: Cambridge University Press, 2011. Pp. x + 612. £110/$180. 9781107003071. *Journal of Hellenic Studies* 133:169–170.

Hugh Cayless and Samuel J. Huskey. 2018. Guidelines for Encoding Critical Editions for the Library of Digital Latin Texts. https://digitallatin.github.io/guidelines/LDLT-Guidelines.html

Giuseppe G. A. Celano. 2014. Guidelines for Greek Treebanking.

Giuseppe G. A. Celano. 2017. Tokenized and sentence-splitted CTSized Ancient Greek texts (v1.1.0) [Data set]. https://github.com/gcelano/CTSAncientGreekXML

Giuseppe G. A. Celano. 2019. Standoff Annotation for the Ancient Greek and Latin Dependency Treebank [in press]. *LiLa Conference, MILAN* .

Giuseppe G. A. Celano and Gregory Crane. 2015. Semantic role annotation in the ancient greek dependency treebank. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*. page 26.

Fabio Ciotti and Francesca Tomasi. 2016. Formal Ontologies, Linked Data, and TEI Semantics. *Journal of the Text Encoding Initiative* 9.

Gregory Crane, Giuseppe G. A. Celano, and Bridget Almas. 2014. The Ancient Greek and Latin Dependency Treebank by PerseusDL. http://perseusdl.github.io/treebank_data/

Roger D. Dawe. 1996. *Sophoclis Aiax*. Teubner, Stuttgart.

Matthew James Driscoll and Elena Pierazzo. 2016. *Digital Scholarly Editing: Theories and Practices*. Open Book Publishers. Google-Books-ID: qW_jDAAAQBAJ.

Øyvind Eide. 2014. Ontologies, data modeling, and TEI. *Journal of the Text Encoding Initiative* 8.

Patrick J. Finglass. 2011. *Sophocles. Ajax*, edited with introduction, translation and commentary. CUP, Cambridge.

Hans Gabler. 2010. *Theorizing the digital scholarly edition*. Literature Compass 7:43 – 56. https://doi.org/10.1111/j.1741-4113.2009.00675.x

TEI Guidelines. 2019. 12 Critical Apparatus - The TEI Guidelines. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html

Maurizio Lana, Raffaella Afferni, Alice Borgna, Paolo Monella, and Timothy Tambassi. 2017. *". . . But What Should {I} Put in a Digital Apparatus?{"A} Not-So-Obvious Choice: New Types of Digital Scholarly Editions*. Sidestone Press. https://iris.unipa.it/handle/10447/284241

LDLT. 2019. The Library of Digital Latin Texts | Digital Latin Library. https://digitallatin.org/library-digital-latin-texts

Hugh Lloyd-Jones and Nigel G. Wilson. 1994. *Sophocles*, volume 1. Harvard University Press.

Francesco Mambrini. 2016. *The Ancient Greek Dependency Treebank: Linguistic Annotation in a Teaching Environment*. Digital Classics Outside the Echo-Chamber: Teaching, Knowledge pages 83–99.

Donald J. Mastronarde. 2010. Euripides scholia: The xml structure. https://euripidesscholia.org/EurSchStructure.html

Paolo Monella. 2008. Towards a digital model to edit the different paratextuality levels within a textual tradition. Digital Medievalist 4(0). https://doi.org/10.16995/dm.62

Paolo Monella. 2012. Why are there no comprehensively digital scholarly editions of classical texts? In *Proceedings of the 4th Meeting of Digital Philology*. volume 15.

OGL. 2016. OpenGreekAndLatin | First1kgreek. https://github.com/OpenGreekAndLatin/First1KGreek

Eyal Oren, Knud Möller, Simon Scerri, Siegfried Handschuh, and Michael Sintek. 2006. What are semantic annotations. *Relatório técnico. DERI Galway* 9:62.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology*, De Gruyter, Berlin, Boston, pages 299–320. https://doi.org/10.1515/9783110599572-017

Marco Carlo Passarotti. 2006. Towards Textual Drift Modelling in Computational Philology. *LINGUISTICA COMPUTAZIONALE* 24(A):63–86.

Alfred C. Pearson. 1924. *Sophoclis Fabulae: recognovit brevique adnotatione critica instruxit*. Clarendon Press, Oxford.

Perseus. 2019a. Digital library. http://www.perseus.tufts.edu/hopper/

Perseus. 2019b. Perseus Digital Library | canonical greek literature. https://github.com/PerseusDL/canonical-greekLit

Perseus. 2019c. Scaife viewer. https://scaife.perseus.org/

Jens Østergaard Petersen. 2016. Merula: A Standoff Implementation of TEI (Text Encoding Initiative) markup. https://github.com/jensopetersen/merula

Ettore Romagnoli. 1926. *Sofocle, Le Tragedie. Aiace, Filottete*, volume 1. Zanichelli, Bologna.

Matteo Romanello, Monica Berti, Federico Boschetti, Alison Babeu, and Gregory Crane. 2009. Rethinking Critical Editions of Fragmentary Texts by Ontologies. *ELPUB2009. Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proceedings of the 13th International Conference on Electronic Publishing held in Milano, Italy 10-12 June 2009* pages 155–174.

Desmond Schmidt. 2012. The role of markup in the digital humanities. *Historical Social Research/Historische Sozialforschung* pages 125–146.

Jochen Tiepmar and Gerhard Heyer. 2019. The Canonical Text Services in Classics and Beyond. In *Digital Classical Philology*, DeGruyter - Saur, pages 102–124. https://doi.org/10.1515/9783110599572-007

Francesca Tomasi, Fabio Ciotti, Marilena Daquino, and Maurizio Lana. 2015. Using Ontologies as a Faceted Browsing for Heterogeneous Cultural Heritage Collections. In *Proceedings of 1st AI*IA Workshop on Intelligent Techniques at Libraries and Archives*.

Ugarit. 2019. Translation alignment editor. http://ugarit.ialigner.com/

AppCrit Workgroup. 2014. Critical apparatus workgroup | teiwiki. https://wiki.tei-c.org/index.php/Critical_Apparatus_Workgroup

Strategie e metodi per il recupero di dizionari storici

Eva Sassolini[1], Marco Biffi[2,3]
[1]Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa
[2]Accademia della Crusca, Firenze
[3]Università degli Studi di Firenze
eva.sassolini@ilc.cnr.it marco.biffi@unifi.it

Abstract

English. The article describes ongoing work on the digitization of an authoritative and historically important Italian dictionary, namely Il Grande Dizionario della Lingua Italiana (GDLI) of S. Battaglia, with a focus on the stages of the conversion of this text into structured digital data. We report on the preliminary results of a collaboration between the Accademia della Crusca and Istituto di Linguistica Computazionale "A. Zampolli", which aims to extract the contents of the GDLI to convert them into structured digital data for human use, and/or to be integrated with other language resources, both dictionaries and corpora. The extraction process is articulated on the one hand in the definition of data extraction procedures, on the other hand in the adoption of strategies aimed at supporting the correction of errors.

Italiano. L'articolo descrive un approccio sperimentale all'estrazione, da formato digitale non standard, della completa struttura delle entrate lessicali del Grande Dizionario storico della Lingua Italiana (GDLI) di S. Battaglia. Sono riportati i risultati preliminari di una collaborazione tra l'Accademia della Crusca e Istituto di Linguistica Computazionale "A. Zampolli" del CNR, che mira a convertire i contenuti testuali in dati digitali strutturati per offrirli alla consultazione e allo studio degli utenti e/o per la successiva integrazione con altre risorse linguistiche, sia dizionari che corpora. Il processo di estrazione si articola da un lato nella definizione di procedure di estrazione dei dati, dall'altro nell'adozione di strategie finalizzate al supporto alla correzione degli errori.

1 Introduzione

Il progetto, nato per strutturare l'intero elenco di voci del dizionario GDLI[1], ha richiesto un articolato procedimento di estrazione, data la complessità dei dati e la disponibilità di un formato digitale non standardizzato. Il testo digitale da cui siamo partiti era costituito da un formato Word parzialmente strutturato, ottenuto sottoponendo l'originale cartaceo a procedure di OCR[2], senza nessun tipo di collazione, parziale o totale. Il processo di acquisizione ha evidenziato caratteristiche stilistiche e scelte di layout derivate dall'originale che hanno reso l'OCR estremamente complicato. La versione edita presenta una suddivisione della pagina in 3 colonne, un colore della carta non sufficientemente bianco, nonché un carattere tipografico relativamente piccolo e una altrettanto minima interlinea. Per ragioni legate a tempo e costi dell'impresa non abbiamo potuto migliorare la qualità dell'OCR, almeno in questa fase del progetto, come viene attualmente proposto in letteratura nei nuovi approcci. Nel caso specifico, utilizzando tecniche di pre- e/o post-elaborazione dell'output eseguita attraverso l'uso di un singolo o più motori di OCR. Inizialmente abbiamo valutato l'utilizzo di sistemi di estrazione automatici, sia basati su regole che su tecniche di *machine learning* (Khemakhem et al. 2017) ma l'analisi dei dati ha escluso l'opzione. La complessità strutturale non è l'impedimento maggiore, più rilevante è il numero e la varietà degli errori, a cui si aggiunge la mancanza di un training corpus opportuno per l'addestramento. Tutto questo ci ha spinto verso un approccio sperimentale, basato su strategie di definizione di regole di estrazione dal formato Word. Abbiamo inoltre evidenziato una distribuzione non uniforme delle tipologie di errore nei vari volumi, probabilmente influenzata dalla lunga gestazione dell'opera editoriale complessiva (vedi Tab. 1).

[1] *Grande Dizionario della Lingua Italiana*, di Salvatore Battaglia (poi diretto da Giorgio Bàrberi Squarotti), Torino, UTET, 1961-2002, 21 voll.; con *Supplemento 2004*, diretto da Edoardo Sanguineti, Torino, UTET, 2004, e *Indice degli autori citati nei volumi I-XXI e nel Supplemento 2004*, a cura di Giovanni Ronco, Torino, UTET, 2004.

[2] La Optical Character Recognition è una tecnologia che permette di convertire un'immagine PDF o di altro tipo in testo digitale.

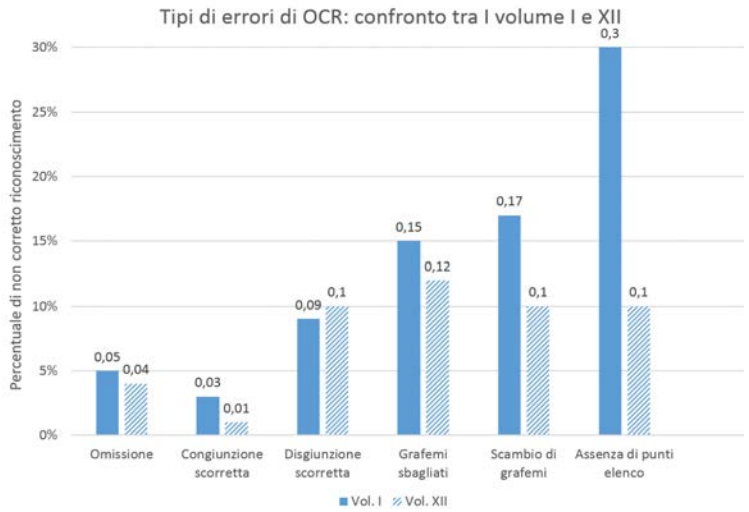


Tabella 1: confronto dei tipi di errore tra volumi diversi

In un lavoro realizzato in un arco temporale di 40 anni, era probabilmente inevitabile la presenza di cambiamenti e aggiustamenti (anche minori), introdotti nel tempo sia a livello delle voci che nel corpus di riferimento del GDLI, che hanno avuto influenze sulle procedure di OCR. Questo contesto ha reso difficile appoggiarsi ad esperienze di altri, pur se indirizzati come noi verso approcci non standard. In alcuni progetti simili si parla di "digitalizzazione attraverso una procedura primitiva" (Bausi, 2016), ma ci si appoggia poi principalmente alla ri-digitazione manuale da parte di studiosi ed esperti qualificati. Nel nostro caso, data la dimensione e complessità dei dati, è necessario limitare il ricorso alla correzione manuale, per le stesse ragioni di contenimento di tempi e costi indicate sopra.

2 L'approccio

Abbiamo impostato un piano di lavoro a lungo termine, che comprendesse 'tappe' da raggiungere progressivamente: 1) riconoscimento del lemma; 2) identificazione di tutti i campi del lemma principale; 3) numero di sensi principali; 4) numero di sensi annidati; 5) campi di ogni senso principale; 6) campi di ciascun senso annidato 7) *mapping* in formato TEI. L'approccio seguito consiste in fasi di riconoscimento successive che, partendo dall'identificazione del lemma dell'entrata lessicale, ne eseguono la segmentazione progressiva individuando, attorno a questo nucleo, gli altri campi dell'entrata. Potremmo definirlo un processo di parsing a più livelli. Ogni campo ha richiesto strategie specifiche per l'identificazione delle caratteristiche distintive, che, tradotte in vincoli di corretta attribuzione e impostati in modo incrementale, hanno portato ad un riconoscimento sempre più granulare della struttura dell'entrata. Oltre a definire procedure software di estrazione e codifica abbiamo implementato metodi di supporto alla correzione manuale e un sistema efficiente di revisione e riallineamento successivo dei dati estratti, per contenere il più possibile l'intervento manuale. L'articolo descrive l'approccio generale e le prime tappe del progetto; la conversione in un formato standard di rappresentazione, pur essendo un impegno rilevante e dall'impatto non trascurabile sul progetto, esula tuttavia dai nostri intenti.

2.1 L'analisi dei dati

Il GDLI è il principale dizionario storico dell'italiano pubblicato da UTET. I 21 volumi che lo compongono, terminati di pubblicare nel 2002, sono corredati da due supplementi integrativi, il primo del 2004 e l'altro del 2009, e da un *Indice degli autori citati*. Recentemente, è stato firmato uno storico accordo tra la UTET e l'Accademia della Crusca, che ha concesso a quest'ultima i diritti per un'edizione elettronica dell'opera, destinata alla consultazione gratuita. Dal maggio 2019 è quindi possibile consultare e interrogare il GDLI con un motore di ricerca per forma applicato al testo in formato Word sopra citato (www.gdli.it). Per quanto il testo elettronico presenti molte debolezze, l'approdo finale di ogni ricerca è la riproduzione in immagine dell'originale a cui si rimane del tutto fedeli, anche in questa edizione, consentendone una comoda lettura con l'ingrandimento a video, a differenza della versione cartacea in cui le dimensioni ridotte dei caratteri non permettono un facile accesso. Nella ricerca si possono certamente perdere alcuni risultati di forme "occultate" dagli errori di OCR ma, una volta arrivati alla pagina, il consultatore può attingere appieno a tutte le preziose informazioni del dizionario. Questa rappresenta soltanto una fase iniziale del progetto: il contributo scientifico dell'ILC si inserisce a fronte dell'esigenza di fornire un accesso più articolato alle informazioni. Grazie a

storiche esperienze nella lessicografia computazionale (Calzolari et al., 1987; Calzolari et al., 1993) stiamo stati coinvolti per implementare il complesso processo di estrazione e riconoscimento della struttura delle entrate. L'analisi dei dati ha evidenziato un input costituito da oltre 23.000 pagine di testo, rappresentate in un formato Word contenente diverse tipologie di errore. Come affermato nell'introduzione, il testo cartaceo originale presenta caratteristiche stilistiche e scelte di layout che hanno condotto il sistema di OCR verso inevitabili problemi di corretta interpretazione. Gli errori di riconoscimento sono stati analizzati su ogni singola caratteristica strutturale del dizionario: lemma, varianti ortografiche, categoria grammaticale, codici d'uso, definizione, etimologia, sensi principali e sensi aggiuntivi (annidati).

N.	Originale cartaceo	Testo OCR
1	**Amminoazobenzène** (*aminoazobenzène*), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di *giallo d'anilina*: cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).	Am mi no a z ob e nz è ne (*aminoazobenzène*), sm. Chim. Composto organico classificato tra i coloranti azoici conosciuto anche col nome di *giallo d'anilina:* cristalli gialli che si sciolgono in alcole ed etere, assai meno in acqua (usato nella colorazione di prodotti alimentari e per preparare altri coloranti).
2	**Assolare**[1], tr. (*assólo*). Disus. Rendere solo. – *Assolare una carta*: tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da *solo* (v.). **Assolare**[2], tr. (*assólo*). Esporre al sole; rendere soleggiato. = Deriv. da *sole* (v.). **Assolare**[3] (*assuolare*), tr. (*assólo* o *assuólo*). Disporre a strati. = Deriv. da *suolo* (v.).	**Assolare**[1], tr. (*assólo*). Disus. Rendere solo. – *Assolare una carta*: tenere scompagnata, nel gioco, una carta di un dato segno. = Deriv. da *solo* (v.). **Assolare**[2], tr. (*assólo*). Esporre al sole; rendere soleggiato. = Deriv. da *sole* (v.). **Assolare**[3] (*assuolare*), tr. (*assólo* o *assuólo*). Disporre a strati. = Deriv. da *suolo* (v.).
3	**Ammacchiare**[1], rifl. (*m'ammàcchio, t'ammacchi*). Raro. Nascondersi nella macchia. *B. Davanzati*, I-136: Floro s'ammacchiò: vedendo poi presi i passi dell'uscita, s'uccise.	Ammacchiate[1]) rifl. (*m'ammàcchio, Vammacchi*). Raro. Nascondersi nella macchia. *B. Davanzati*, I-136: Floro s'ammacchiò: vedendo poi presi i passi dell'uscita, s'uccise.
4	**Attendista,** agg. e sm. e f. (plur. m. *-i*). Neol. Chi evita di prendere posizione (e resta in attesa degli avvenimenti, riservandosi di decidere secondo il loro svolgersi). = Fr. *attentiste* (1941), da *attendre* 'attendere'. **Attenditóre,** agg. e sm. (femm. *-trice*). Ant. Che attende, aspetta.	= Deriv. da *attendere*.¶ Attendista, agg. e sm. e f. (plur. m. *-i*). Neol. Chi evita di prendere posizione (e resta in attesa degli avvenimenti, riservandosi di decidere secondo il loro svolgersi).¶ = Fr. *attentiste* (1941), da *attendre* "attendere".¶ Attenditóre, agg. e sm. (femm. *-trice*). Ant.¶ Che attende, aspetta.¶

Tabella 2: esempi di errori del sistema di OCR

Ciascuno dei campi presenta errori di vario tipo, che vanno dalla mancata segmentazione dei paragrafi, all'interpretazione errata della punteggiatura e dell'ortografia delle parole, al mancato rispetto delle diverse sezioni della voce del dizionario: punti elenco, rientro, dimensione del carattere ecc. (vedi Tab. 2). La presenza di errori ha assunto quindi un peso decisivo nel progetto e ha mostrato come le sole procedure automatiche, per quanto raffinate e puntuali, non sarebbero state sufficienti a produrre un risultato corretto.

2.2 Le fasi di lavoro

Siamo partiti da una sommaria classificazione dei problemi relativi all'inesattezza del dato distinguendo tra errori "bloccanti" e "non bloccanti", per poi procedere con i casi più specifici. La differenza sta nell'impatto dell'errore sulla procedura di *parsing* dei dati. Gli errori bloccanti sono costituiti prevalentemente dal mancato riconoscimento di un nuovo lemma. In questo caso, non potendo chiudere correttamente la voce precedente, si inficia il successivo processo di raffinamento, impedendo la definizione dei confini e campi dell'entrata (vedi Tab. 3).

Errore	Tipo	Correttivo	Esempi
Ortografico in lemma	Non bloccante	Riferimento con indicazione di vol. e pagina nel file di report	
Segni di puntegiatura con funzione di separatori tra campi	Non bloccante	Correzione automatica dei dati e rif. puntuale nel file di report	
Omissione	Bloccante	Non risolto	
Lemma non trovato ad inizio paragrafo	Bloccante	Evento comunque individuato e riportato nel file di report	(vedi Tab. 2. n.4)

Tabella 3: alcuni tipi di errore nel lemma

Un errore "non bloccante" interviene invece quando non è possibile separare il codice grammaticale, da quello d'uso, e/o dalle varianti ortografiche e/o queste dalla definizione. Questa tipologia di errori producono

un'entrata non corretta, ma sulla quale si possono impostare le successive fasi di raffinamento progressivo, procedendo in un certo senso a 'tappe' nella strutturazione della voce. Mentre per gli errori "bloccanti" non abbiamo trovato un'efficiente soluzione alternativa alla revisione manuale post-processing, per gli altri è possibile corredare il parser di meccanismi di annotazione puntuale. Segnalare quando mancano campi obbligatori o se il loro ordine non è rispettato, e riferendo puntualmente il caso in un file di report. In alcuni casi, quando è possibile impostare un'indagine più puntuale del dato, i file di report sono finalizzati al controllo delle soluzioni già inserite in fase di parsing, così da alleggerire il lavoro di revisione manuale.

3 Prospettive

Nelle fasi successive del progetto le risorse estratte hanno assunto una valenza autonoma, per esempio abbiamo prodotto un confronto tra i lemmi del GDLI e quelli del TLIO[3]: il primo dizionario storico di tutte le varietà dell'italiano antico fino al 1375. Stiamo pensando di allargare il confronto anche ad altri dizionari, primo fra tutti il Dizionario Macchina dell'Italiano (DMI) che è patrimonio di storiche linee di ricerca dell'ILC. Nel recente passato le ricerche nel settore si sono concentrate principalmente sullo sviluppo di lessici computazionali in applicazioni di elaborazione del linguaggio naturale, ma oggi i metodi e le tecniche sviluppati per estrarre, strutturare e rappresentare dizionari, possono avere un ruolo potenziale per la progettazione e costruzione di risorse orientate all'uomo, nelle attività lessicografiche dell'editoria, soprattutto digitale. I dizionari storici sono in grado di documentare l'evoluzione diacronica della lingua, mostrando la dimensione storica del lessico. I potenziali vantaggi della digitalizzazione e strutturazione di un dizionario monumentale come il GDLI risiedono anche nell'importanza delle citazioni che vi si possono consultare. Come sostenuto da Beltrami e Fornara (2004), il vero fulcro del dizionario è la presenza massiccia di citazioni di testo, che coprono un'ampia varietà di usi linguistici, dalla lingua quotidiana e letteraria, alle lingue regionali e/o specializzate/specialistiche, ai neologismi e alle parole straniere. Le citazioni offrono preziose informazioni sulle prime attestazioni delle parole, sulle loro varianti formali/diacroniche/diatopiche; sugli autori che le citano e sulle loro etimologie. Per questo motivo stiamo implementando procedure software che da un lato estraggano le varianti dalla struttura della voce e dall'altro, attraverso l'elaborazione delle informazioni estratte dal volume dell'indice degli autori citati, consentano di predisporre filtri su autore ed epoca/data per le rispettive citazioni.

4 Conclusioni

Il nostro impegno è finalizzato a rendere una delle maggiori risorse lessicografiche dell'italiano utilizzabile per il trattamento computazionale, ma l'analisi conclusiva dell'approccio adottato è ancora prematura, soprattutto per quanto riguarda l'estrazione dei sensi annidati. A progetto in corso un'analisi conclusiva del lavoro non è possibile, tuttavia ci sembra di comune utilità descrivere la nostra esperienza, come aiuto per pianificare progetti analoghi, per i quali mancano riferimenti certi in letteratura. Questi progetti, avendo un alto grado di complessità e di incognite, si sviluppano troppo spesso senza un'adeguata divulgazione, il che significa che spesso i ricercatori e gli studiosi devono in un certo senso "reinventare la ruota". L'intento di questo articolo è proporre il nostro approccio come *caso di studio* in contesti in cui non è possibile ricorrere a strumenti e/o procedure consolidate o sperimentali già note in letteratura e magari offrire spunti per discutere delle strategie specifiche che sono state utilizzate.

References

Sassolini E., Khan A. F., Biffi M., Monachini M., Montemagni S. 2019. *Converting and structuring a Digital Historical Dictionary of Italian: a case study*. (eds.) Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o.

Khemakhem M., Herold A., Romary L. 2018. *Enhancing Usability for Automatically Structuring Digitised Dictionaries.* GLOBALEX workshop at LREC 2018. May 2018. Miyazaki. Japan.

Agostiniani L., Montemagni S., Paoli M., Picchi E. 2004. *Lessicografia dialettale e computer: questioni di rappresentazione e recupero dei dati*. Centro Interuniversitario di Studi Veneti, Venezia (Italia).

Beltrami P.G., Fornara S. 2004. *Italian historical dictionaries: from the Accademia della Crusca to the web.* In «International Journal of Lexicography». vol. 17. n. 4. pp. 357-384.

[3] http://tlio.ovi.cnr.it/TLIO/

Grande Dizionario della lingua italiana. Opera diretta da Salvatore Battaglia. Torino. UTET. 1961-2002.

Calzolari N., Hagman J., Marinai E., Montemagni S., Spanu A., Zampolli A. 1993. *Encoding Lexicographic Definitions as Typed Feature Structures*. In: F. Beckmann, G. Heyeder (eds.). Theorie und Praxis des Lexikons. Beiträge zu einem Kolloquium über theoretische Lexicologie und praktische Lexikographie. Walter de Gruyter. Berlin. pp. 274-315.

Monachini M., Picchi E. 1993. *Computational lexicography: a query system for text corpora.*

Calzolari N., Picchi E. 1988. *Acquisition of semantic information from an on-line dictionary*. (1988). Proceedings.

Calzolari N., Picchi E., Zampolli A. 1987. *The Use of Computers in Lexicography and Lexicology*. (1985). Proceedings.

Calzolari N. 1984, "*Detecting Patterns in a Lexical Database".* Proceedings of the 10th International Conference on Computational Linguistics. Stanford. California. pp. 170-173.

TEI Consortium. eds. *"9. Dictionaries."* TEI P5: Guidelines for Electronic Text Encoding and Interchange. [3.5.0]. [Last updated on 16th July 2019]. TEI Consortium. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSFLT (30/10/19)

Encoding Byzantine Seals: SigiDoc

Alessio Sopracasa
Sorbonne Université
Équipe Monde Byzantin (CNRS UMR 8167)
Institut d'histoire et civilisation de Byzance
alessio.sopracasa@paris-sorbonne.fr

Martina Filosa
University of Cologne
Department of Byzantine and Modern Greek Studies
martina.filosa@uni-koeln.de

Abstract

English. This paper will illustrate the current state of development of SigiDoc, an XML-based and TEI-compliant effort for the encoding of Byzantine seals. SigiDoc represents the first attempt to extend the digital approach — already applied to inscriptions, coins and papyri — to Byzantine seals (and bread stamps). The project — which has been discussed, albeit without results, in the framework of Byzantine studies since the early 2000s — was taken up in 2015 by Alessio Sopracasa as part of his Marie Curie fellowship, and the work is currently continuing through a collaboration between Paris, London and Cologne, particularly with Martina Filosa. SigiDoc is in a fairly advanced development phase and currently allows the display of metadata dealing with the seal as an object, such as physical description, history of the seal, typology, description of the iconography, as well as the critical edition of the legend, along with *apparatus criticus* and commentary. The transformation of XML documents into a webpage, various indexing possibilities, a search interface as well as multi-lingual features are delivered to SigiDoc by EFES (EpiDoc Front-End Services).

Italiano. Obiettivo del contributo è illustrare lo stato attuale dello sviluppo di SigiDoc, uno strumento basato su XML e conforme a TEI per la codifica dei sigilli bizantini. SigiDoc rappresenta il primo tentativo di estendere l'approccio digitale — già applicato ad epigrafi, monete e papiri — ai sigilli bizantini (facilmente adattabile alla codifica degli stampi eucaristici). Il progetto — del quale si è discusso in seno alla bizantinistica internazionale sin dall'inizio degli anni 2000, ma senza risultati — è stato intrapreso nel 2015 da Alessio Sopracasa nell'ambito di una fellowship Marie Curie ed il lavoro continua oggi in collaborazione tra Parigi, Londra e Colonia, in particolare con Martina Filosa. SigiDoc si trova in una fase di sviluppo piuttosto avanzata e permette attualmente la visualizzazione dei metadati (descrizione fisica, storia del sigillo, tipologia, descrizione dell'iconografia), l'edizione critica della legenda, insieme all'apparato critico ed al commento. SigiDoc si serve di EFES (EpiDoc Front-End Services) per la visualizzazione web provvisoria dei documenti XML, per l'indicizzazione, per le opzioni multilingua, nonché per l'interfaccia di ricerca.

1 Byzantine Sigillography and Digital Humanities

SigiDoc is an effort providing XML-based and TEI-compliant encoding standards for Byzantine seals. It represents the first attempt to extend the digital approach, which in the past decade has already been applied to inscriptions, coins and papyri[1], to Byzantine seals, and, with minor adjustments, to other coin-like objects, such as bread stamps. Already presented to the international community of Byzantine sigillographers during the 12th International Symposium of Byzantine Sigillography held in May 2019 at the Hermitage Museum, St. Petersburg, SigiDoc will be presented at the 9th Annual Conference AIUCD for the first time to a wide community of digital humanists.

Seals are the only survivors of the written documents used in daily administration of the Byzantine Empire, the former Eastern Roman Empire (4th to 15th century). These small discs, mostly made of lead, commonly have an iconographic representation on one side (a saint, the Virgin, etc.) and a legend — with name, at times surname, dignities, functions, places — on the other. Hence, what we call "seal" is both the object bearing an impression and the impression itself, left on the disk by a tool called *boulloterion*, i.e. the matrix (only six of these tools survive nowadays in comparison to the extant 80/100,000 seals' impressions). The seals greatly

[1] A comprehensive and up-to-date list of projects related to inscriptions, coins and papyri within the Digital Humanities can be found at: <https://wiki.digitalclassicist.org/Category:Projects>.

improved our knowledge of the central and provincial administrative apparatus of the State, the Church, and the Army, and thanks to them the Byzantine administrative, political, and social history is gradually being rewritten. But the impact of the seals is still limited, since access remains restricted to a very small number of researchers, the great amount of unpublished material is widely scattered across museums, libraries, and private collections, and the paper publications are fragmented and far from being widely available.

Despite this situation, unlike sister ancillary disciplines such as epigraphy, papyrology or numismatics, Byzantine sigillography has not received much attention from the Digital Humanities for a long time, although this has been a wish of at least part of the scientific community since the 2000s. Resumed and abandoned several times without significant results, the project has finally been revived by the authors. SigiDoc strives to combine traditional research methods and scholarship in the field of Byzantine sigillography with the technologies offered by the Digital Humanities. In fact, standards for scientific publications within Byzantine sigillography have been established during the 20th century with the works by Vitalien Laurent[2], George Zacos[3], Nicolas Oikonomides[4], Jean-Claude Cheynet[5], and Werner Seibt[6]: hence, one of the main tasks of SigiDoc has been to convert these standard publications into a digital and TEI-compliant form, while partially reassessing them and providing them with more consistency.

2 What SigiDoc is and for what purpose it has been developed

At the time this paper is written, SigiDoc is in an advanced *beta* version and it will hopefully be close to its 1.0 version by the time the AIUCD annual conference takes place.

SigiDoc is largely based on the ongoing experience of EpiDoc, a collaborative effort that provides guidelines and tools for the encoding of scholarly editions of ancient documents; alongside the scientific objectives, EpiDoc developed also an educational asset, with the didactic experiences of EpiDoc training weeks, showing great potential in teaching traditional epigraphy and papyrology.[7] EpiDoc uses a subset of the TEI standard for the representation of texts in digital form[8], having clearly established itself as the most robust and widely supported format for encoding editions of ancient texts on different text-bearing objects.

Far from being only a simple adaptation and implementation of existing solutions, from a technical point of view, SigiDoc is: 1) a schema, merged with EpiDoc's one; 2) a template, i.e. SigiDoc's edition structure; 3) a stylesheet for web visualisation; 4) a set of stylesheets for the critical edition of the legends on seals, derived from EpiDoc but adapted to the needs of Byzantine sigillography; 5) a highly customised version of EFES (see below); 6) a set of guidelines (for metadata, leidenisation, indexing, etc.); 7) a set of files intended to be shared among all the future SigiDoc projects and expanded through the time and the experience, in order to avoid superfluous duplications on one hand, and to ensure consistency on the other (IDs' lists, controlled vocabularies, authority lists, ontologies, etc.).[9]

SigiDoc is intended for both the creation of digital-born editions of Byzantine seals as well as for the digital conversion of paper publications, and it has been conceived in order on one hand, to provide the users with a common ground for developing their projects, on the other, also to give them the freedom to customise their approach.

During its presentation in St. Petersburg, the community of Byzantine sigillographers gave an enthusiastic feedback to SigiDoc, and several projects aimed at the creation of online *corpora* are now waiting for SigiDoc 1.0 to be released, thus ensuring both its dissemination and implementation. These projects involve leading scholars in the field of Byzantine studies — such as Prof. Jean-Claude Cheynet, Prof. Claudia Sode, Dr. Vivien Prigent — as well as important cultural institutions, such as the Bibliothèque Nationale de France (Paris), the

---

[2] See Vitalien Laurent. 1963–1981. *Le Corpus Des Sceaux de L'empire Byzantin, Voll. 1–5*, Paris.

[3] See George Zacos and Alexander Veglery. 1972. *Byzantine Lead Seals, Vol. 1*, Basel.

[4] See Nicolas Oikonomides. 1986. *A Collection of Dated Byzantine Lead Seals*. Washington, D.C.

[5] See Jean-Claude Cheynet, Turan Gokyildirim, and Vera Bulgurlu. 2012. *Les Sceaux Byzantins du Musée Archéologique d'Istanbul, Istanbul; id. and Maria Campagnolo-Pothitou. 2016. Sceaux de la collection Georges Zacos au Musée d'art et d'histoire de Genève, Geneva.*

[6] See Werner Seibt. 1978, *Die byzantinischen Bleisiegel in Österreich I. Teil, Kaiserhof*, Vienna; *id.* and Alexandra-Kyriaki Wassiliou. 2004. *Die Byzantinischen Bleisiegel in Österreich, Vol. 2, Zentral- und Provinzialverwaltung*, Vienna.

[7] Gabriel Bodard and Simona Stoyanova. 2016. *Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy*, in Gabriel Bodard and Matteo Romanello (eds.), *Digital Classics Outside the Echo-Chamber. Teaching, Knowledge Exchange & Public Engagement*, p. 51–68, London, available: <http://dx.doi.org/10.5334/bat>.

[8] Tom Elliott, Gabriel Bodard, Hugh Cayless *et al.* 2006-2016. *EpiDoc: Epigraphic Documents in TEI XML*. Online material, available: <http://epidoc.sf.net>, delivers thorough information about history and mission of EpiDoc and offers always up-to-date versions of the EpiDoc Guidelines as well as documentation, software and tools to work with EpiDoc.

[9] In section 3, this paper will address only a selection of these topics.

Dumbarton Oaks Research Library and Collection (Harvard's research institute for Byzantine Studies in Washington D.C.), the Epigraphic and Numismatic Museum of Athens and the Geneva Museum of Art. Researchers, museums, public institutions, as well as private collectors will be able to get a definite and stable record of their Byzantine seals, thus preventing deterioration and making their collections available for research, teaching, and presentation to the general public.

Thanks to the increasing number of SigiDoc-based projects, the long-term aim is to ensure dissemination, sharing, and sustainability of the data, and to make available a very wide range of published and unpublished material, edited to a high scholarly standard. SigiDoc has not been conceived just to realise individual projects (an interesting, though limited aim): the use of the same guidelines and set of tools is intended to allow the creation of a common search interface, through which all *corpora* will be virtually unified in a higher-order catalogue, enabling actions going from the simple cross-referencing to the advanced search throughout every *corpus* published in SigiDoc standard.

Through a dedicated website, the developers will ensure a proper dissemination of SigiDoc. The website (http://sigidoc.huma-num.fr/) — which is, as for now, empty — will host SigiDoc's documentation as well as all the aforementioned materials needed to run it. Through this site the user will be informed about the life of SigiDoc: status of SigiDoc-based projects, training sessions and events, technical updates, etc.

In order to use SigiDoc, a formal training is needed. Training weeks as well as shorter training events will take place regularly (once or twice a year): they are inspired by the well-established EpiDoc training weeks, which have repeatedly shown the feasibility as well as the effectiveness of this teaching format. Through in-depth training, SigiDoc will be able to increase the dissemination and continuity of sigillography itself, while the creation of a new professional figure, i.e. the digital sigillographer, will facilitate the integration between the traditional and the digital approach to Byzantine seals. Consequently, the foundation of a more interrelated scientific community will be laid: a network of digital Byzantine sigillographers is still a *desideratum* within the larger community of Byzantine sigillography; in order to achieve this goal, a common scientific and prac-tical ground delivered by a standard like SigiDoc is much needed.

3 Main Features of SigiDoc

3.1 TEI-XML Template and Data Encoding

SigiDoc XML template organises the information in hierarchical mark-up inside three main common ele-ments of the standard TEI structure: 1) <teiHeader/> for metadata; 2) <facsimile/> for the digital reproduction of the artefact; 3) <text/> for the legend's critical edition, commentary and bibliography.

teiHeader: among the data nested inside this element, SigiDoc stresses the importance of providing each seal with a unique numerical identifier (being it the first attempt of systematic categorisation in Byzantine sigillography).To preserve consistency, a common file of ID numbers will be shared among all the SigiDoc projects.

Several thesauri and controlled vocabularies are being prepared: among them, the classification of the seal (imperial, military, etc.), the milieu of the issuer, the language(s) of the legend, the work type (original impres-sion, drawing, verbal description, etc.), the material, the layout (iconography only, text only, both, etc.), the execution (struck, cast, printed, etc.), the shape, the iconography (see below). All these lists will be provided in different languages (English, French, German, and Italian by default, but each project will be able to cus-tomise their languages).

The preservation history of the seal is a major concern, not only in establishing which is the current repos-itory of a seal, but also in being able to follow it through its different displacements, which is of the utmost importance especially when the seal enters a private collection or is sold in an auction. Byzantine seals are increasingly present in online auctions: thanks to SigiDoc it will be easier to follow them before they disappear in private collections; the leading journal in the field — *Studies in Byzantine Sigillography*[10] — includes a final section listing the seals sold through auctions, but its biannual publication (without photos of the seals) limits its effectiveness, whereas with SigiDoc the information will be updated without any delay, thus promot-ing its circulation and its scientific study.

An important part of the metadata is devoted to the findspot and the find circumstances: these data contribute significantly to the historical interpretation of the seal, helping to establish both the areas directly administrated by the Byzantine Empire and those — outside the Empire — of contact or influence. Unfortunately,

[10] See <https://www.degruyter.com/view/serial/36534>.

this kind of data is often lacking for Byzantine seals, and this makes particularly valuable the preservation of this information and its linking with similar data among different *corpora*.

The iconography deserves here a special mention: this is a key element to Byzantine sigillography, but also one of the most challenging due the numerous and specific iconographic typologies to be found on Byzantine seals. In SigiDoc 1.0 this topic will be addressed in two ways: a short and general identification of the iconographic theme (according to a shared controlled vocabulary), and a detailed description. However, the degree of details in iconographic description is perhaps the most changing criterion in Byzantine sigillographic editions: this is the reason why a standard tool for the description of images is being developed, in order to introduce consistency in sigillographic editions, and to restore the importance of this feature, too often neglected. This tool will be released after the launch of SigiDoc 1.0.

Links and relationships both within the *corpora*, and between texts and external datasets can be established and realised as hyperlinks in SigiDoc: the data is being enriched and its interoperability is being increased using online resources and authority files such as prosopographies and geographical gazetteers.[11]

Facsimile: The digital reproduction of the seals will be displayed as a digital facsimile above the edition: in case of seals in bad state of preservation, some of the ongoing projects (especially those based in Cologne and Paris) will provide images created with RTI technology (Reflectance Transformation Imaging).[12] Of course, digital reproductions won't be always available, especially for seals edited between the 19th and the 20th century and of which no trace can be found: in this case, we have often drawings or verbal descriptions.

Text: As far as the critical edition of the seals is concerned, variant readings and restorations are encoded in TEI, thus enabling the generation of *apparatus criticus*, parallel texts, and diplomatic editions; moreover, the leidenisation of the legend allows for a full editorial interpretation based on the Leiden conventions (especially Panciera), but adapted according to the sigillographic editorial standards.[13] For the diplomatic edition, SigiDoc uses AthenaRuby[14], a Unicode-compliant font based on the lettering/epigraphy of Byzantine coins and seals, and designed by Joel Kalvesmaki at the Center for Byzantine Studies at Dumbarton Oaks (Washington D.C.). AthenaRuby is currently the most accurate Greek font in terms of the lettering of Byzantine coins and seals. It is not yet widely used within the sigillographic (and numismatic) community due to the preference accorded by them to more abstract yet more approximate fonts, such as New Athena. Being the lettering of the legend a key factor in dating and contextualising a *specimen*, SigiDoc's developers promote and strongly recommend the use of AthenaRuby for the encoding of Byzantine seals: the diplomatic edition carried out with this font delivers a more accurate representation of the seal's lettering, thus facilitating its understanding even without the observation of the digital reproduction.

3.2 Web Visualisation and Data Valorisation: Contribution from (and to) EFES (EpiDoc Front-End Services)

EFES (EpiDoc Front-End Services)[15], which builds upon existing tools such as Kiln[16], is a highly customisable platform which allows expert and less expert users to get, in a relatively easy and fast way, four main critical features for a TEI-XML based *corpus*: multiple indices, multilingual options, faceted search interface, and a (raw) webpage. TEI-XML files created in SigiDoc, their XSLT stylesheets, as well as part of their tagging have been designed to be best dealt with in EFES. In 2018 SigiDoc became one of the pilot projects using and testing EFES and, in this way, actively contributing to its development, being also the only non-strictly epigraphical project.

The most notable contribution of EFES to SigiDoc is certainly the creation, based on Authority Lists, of automatic indices. The lemmatisation of words and the identification of relevant entities within the legend, as well as the encoding of key words and terms, enable the indexing of words, personal names, geographical entities, offices, titulatures, and other features of philological, epigraphical, and historical interest at large.

[11] See, for example, *Prosopography of the Byzantine World* (<http://pbw2016.kdl.kcl.ac.uk/>); *Prosopographie der mittelbyzan-tinischen Zeit* (<http://www.degruyter.com/view/db/pmbz/>); *Pleiades* (<http://pleiades.stoa.org/>).

[12] For further information regarding RTI technology and its application in sigillographic studies, see Franz Fischer and Stephan Makowski. 2017. *Digitalisierung von Siegeln mittels Reflectance Transformation Imaging (RTI)*, Paginae historiae – Sborník Národ-ního archivu, 25/1, p. 137–141, available: <http://kups.ub.uni-koeln.de/id/eprint/7882>.

[13] For example, the rendering after transformation of several kinds of <gap/> tags has been changed.

[14] AthenaRuby is an OpenType and Unicode-compliant font. For documentation, tools, and selected bibliography visit: <https://www.doaks.org/resources/athena-ruby>.

[15] See <https://github.com/EpiDoc/EFES/wiki> for the technical documentation and <https://github.com/EpiDoc/EFES/wiki/User-Guide> for detailed guidelines and user guide.

[16] See <https://github.com/kcl-ddh/kiln> for the documentation.

SigiDoc users will be able to potentially index every feature deemed relevant for their corpus: the previous listing of indexed features — featuring the most common indices in Byzantine sigillographic publications — is what the authors recommend to all future SigiDoc projects in order to harmonise their indices. Nonetheless, a higher degree of specialisation will be enabled: for example, thanks to the consistent use of AthenaRuby, it will be possible to index — and, ultimately, to search for — single variant letters within the legend.

The using made by SigiDoc of EFES is essentially based on the customisation of the solutions offered by it: this is especially true for the use of Athena Ruby instead of generic Greek capital letters in the diplomatic edition; the XSLT stylesheet organising the webpage, in order to create a list of fields appropriate for Byzantine sigillography; the EpiDoc stylesheets used for the edition of the legends; and, of course, the customisation of the indices and the search interface.

During the next months some improvements related to EFES — mainly concerning further indexing features and automatic bibliographic references — will be delivered.

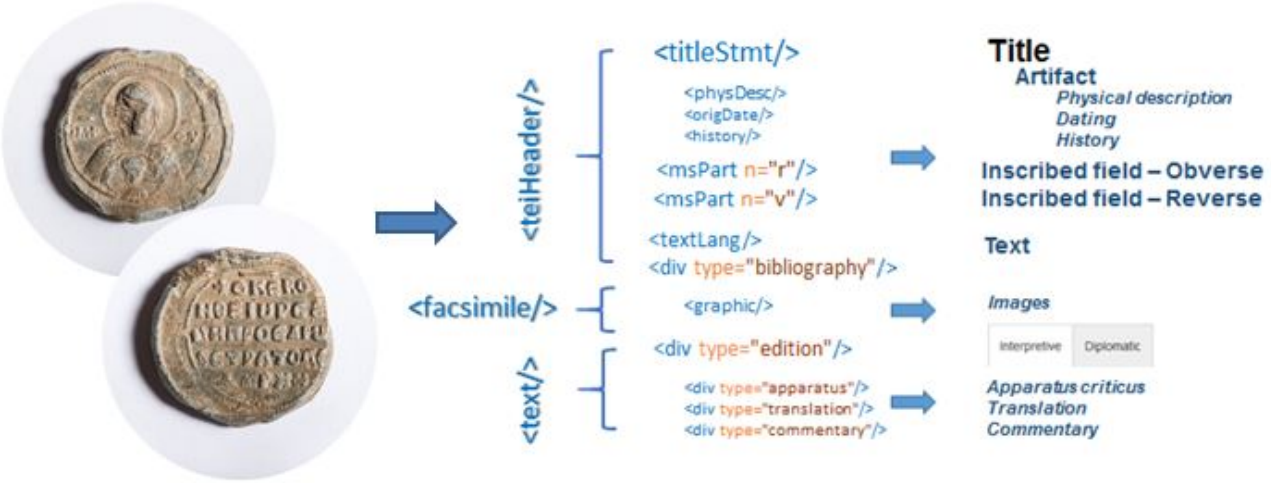


Figure 1. From the seal (left) to the EFES-generated webpage (right), through the SigiDoc XML template (middle).

Conclusions

Digital Byzantine sigillography is an entirely new discipline: SigiDoc has therefore been designed for users with no prior computer skills but with a background in Byzantine sigillography or Byzantine history at large. SigiDoc's aim is to deliver a (reasonably) easy and ready-to-use tool, with a template ready to be filled in according to the need of each project; but this also means that a more advanced user will be able to go further and customise it to a larger extent.

The visibility and availability of data coming from an increasingly number of seals, jointly with the possibility of establishing relations among them, will push further our knowledge of several aspects of Byzantine history, allowing the specialists to carry out analysis in several directions (i.e. the structure of Byzantine society and the relations among individuals and families, the organisation of the administration, the role of personal piety in the choice of the iconography, art history, but also sigillographic epigraphy or lettering, the "writing uses", the Greek language, etc., including the evolution of all these aspects throughout the centuries).

Albeit most of these research directions (and many others) already existed before SigiDoc, thanks to it their analysis will be greatly enhanced, in a way and to an extent extremely difficult to reach without this tool.

Acknowledgments

The authors would like to thank the anonymous reviewers for their suggestions, which have been taken into account where appropriate.

References

Gabriel Bodard and Simona Stoyanova. 2016. *Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy*, in Gabriel Bodard and Matteo Romanello (eds.), *Digital Classics Outside the Echo-Chamber. Teaching, Knowledge Exchange & Public Engagement*, p. 51–68, London, available: <http://dx.doi.org/10.5334/bat>.

Jean-Claude Cheynet, Turan Gokyildirim, and Vera Bulgurlu. 2012. *Les Sceaux Byzantins du Musée Archéologique d'Istanbul*, Istanbul.

Jean-Claude Cheynet and Maria Campagnolo-Pothitou. 2016. *Sceaux de la collection Georges Zacos au Musée d'art et d'histoire de Genève*, Geneva.

Tom Elliott, Gabriel Bodard, Hugh Cayless *et al*. 2006–2016. *EpiDoc: Epigraphic Documents in TEI XML*. Online material, available: <http://epidoc.sf.net/>

Franz Fischer and Stephan Makowski. 2017. *Digitalisierung von Siegeln mittels Reflectance Transformation Imaging (RTI)*, Paginae historiae – Sborník Národního archivu, 25/1, p. 137–141, available: <http://kups.ub.uni-ko-eln.de/id/eprint/7882>.

Vitalien Laurent. 1963–1981. *Le Corpus Des Sceaux de L'empire Byzantin, Voll. 1–5*, Paris.

Nicolas Oikonomides. 1986. *A Collection of Dated Byzantine Lead Seals*. Washington, D.C.

Werner Seibt. 1978, *Die byzantinischen Bleisiegel in Österreich I. Teil, Kaiserhof*, Vienna.

Werner Seibt and Alexandra-Kyriaki Wassiliou. 2004. *Die Byzantinischen Bleisiegel in Österreich, Vol. 2, Zentral- und Provinzialverwaltung*, Vienna.

George Zacos and Alexander Veglery. 1972. *Byzantine Lead Seals, Vol. 1*, Basel.

Preliminary Results on Mapping Digital Humanities Research

Gianmarco Spinaci
DHDK - FICLIT University of Bologna
gianmarco.spinaci@studio.unibo.it

Giovanni Colavizza
University of Amsterdam
g.colavizza@uva.nl

Silvio Peroni
DHARC - FICLIT University of Bologna
silvio.peroni@unibo.it

Abstract

English. Maps of research could provide key insights in the current and future development of the digital humanities. Two obstacles to this end are the definition of what can be considered digital humanities research in the first place, and the mostly unknown coverage of digital humanities publications in citation indexes. In view of addressing these challenges, we release a list of digital humanities journals developed via an iterative approach combining manual curation with citation data. Based on this list, we further assess the journal article coverage of Web of Science, Scopus, Crossref and Dimensions. We find that Crossref has the best coverage of exclusively digital humanities journals, while Dimensions has the best coverage of related journals. Furthermore, we use Dimensions data to map digital humanities research via a directed citation network, finding connections with fields such as computational linguistics and digital libraries, and a strong correlation between journals and citation clusters.

Italiano. Le mappe della ricerca scientifica possono aiutare a comprendere gli sviluppi attuali e futori delle *digital humanities*. I maggiori ostacoli alla loro realizzazione sono la definizione di cosa includere nelle pubblicazioni *digital humanities*, e la copertura degli indici citazionali al riguardo. In questo contributo, pubblichiamo una prima lista di riviste digital humanities, curata tramite un metodo iterativo che include revisione manuale e uso di dati citazionali. Sulla base della lista, verifichiamo la copertura di pubblicazioni *digital humanities* in Web of Science, Scopus, Crossref e Dimensions. Crossref risulta avere la migliore copertura rispetto alle riviste esclusivamente focalizzate alle *digital humanities*, Dimensions rispetto alle riviste adiacenti. Utilizzando i dati citazionali di Dimensions, costruiamo infine una mappa/rete delle *digital humanities* mostrando una relazione tra queste ultime e aree limitrofe quali la linguistica computazionale e le biblioteche digitali, e una forte correlazione tra gruppi di articoli individuati tramite *community detection* e le riviste in cui sono pubblicati.

1 Introduction

It is a scientometrics trope to consider humanities research as poorly indexed in citation databases, and thus poorly understood in terms of research outputs (Hammarfelt, 2016). Several studies have pointed out to the limitations of indexes such as Web of Science and Scopus with respect to the humanities, both in terms of quantity and quality (e.g., lack of books) (Nederhof, 2006). Nevertheless, in recent years more indexes have become available, such as Dimensions and Microsoft Academic, while coverage has been improving (Harzing and Alakangas, 2016). In view of these developments, a comprehensive and cross-index map of research in the (digital) humanities is still pending.

Some previous work has considered the intellectual and social organization of the digital humanities using bibliometrics. Nyhan and Duke-Williams (2014) focused on collaboration patterns in the journals *Computers and the Humanities* and *Literary and Linguistic Computing* (up to 2011), finding a propensity to collaborate within small, tight groups and a persisting tendency for single-author publishing. It is worth noting that more recent work on the humanities as a whole found an increasing propensity for

collaboration, albeit with high variation among different disciplines/departments (Burroughs, 2017). Citation analyses based on *Computers and the Humanities*, *Digital Scholarship in the Humanities* and *Digital Humanities Quarterly* highlighted instead a sparser organization, around thematic areas such as information studies, historical literature, linguistics, natural language processing and statistical text analysis (Gao et al., 2017, 2018). Further work based on the *Journal of Digital Humanities*, *Digital Humanities Quarterly*, *International Journal of Humanities and Arts Computing*, *Digital Medievalist*, *Digital Studies*, *Literary and Linguistics Computing* assessed co-authorship, co-citation and bibliographic coupling networks. The authors found a sustained growth in digital humanities publications, coupled with increasing integration with respect to citation networks, and persisting fragmentation with respect to collaborations as mapped by co-authorship relations (Tang et al., 2017).

A recent bibliometric comparison considered the annual conference of the Italian Association of Digital Humanities and Digital Culture (AIUCD) and the annual Italian conference on computational linguistics (CLiC-it). Results show how collaborations are sparser in digital humanities, how research methodologies usually are introduced in the computational linguistics conference and then readopted in the digital humanities one, and how the citation behaviour in the latter one closely resembles that of humanities scholarship (e.g., higher ratio of references to books) (Sprugnoli et al., 2019). Altmetrics data has also been used to map the digital humanities community worldwide. In particular, Twitter follower and co-retweet networks were used to show how the community is organized around few "influencers" and according to language and geographical region (Grandjean, 2016; Gao et al., 2018). Finally, some scholars attempted to position the digital humanities within the broader context of humanities scholarship (Leydesdorff and Akdag Salah, 2010; Salah et al., 2015).[1]

In this paper, we address the following research questions: *a) what qualifies as digital humanities research, from a bibliometric point of view? b) What is the coverage of citation indexes with respect to digital humanities research? c) What is the organization of the resulting map of research?* We propose an iterative method to individuate digital humanities publications by combining manual journal classification and automatic citation clustering. One of the outcomes of our work is the first version of a list which includes digital humanities journals. We use this list to assess the number of digital humanities journal publications indexed by Web of Science (WoS), Scopus, Crossref and Dimensions. Finally, we use the citation data included in the index with most digital humanities publications, i.e., Dimensions (Hook et al., 2018), to present a map of digital humanities research based on journal articles. It is worth noticing that our results are still preliminary and stem from ongoing work to create a comprehensive map of humanities research.

2 Data and methods

Database coverage limitations notwithstanding, individuating digital humanities (DH) publications is problematic in itself. First of all, there is little agreement on what constitutes DH research among practitioners. Secondly, DH research tends to be highly interdisciplinary, so much so that clear-cut classifications would be intrinsically arbitrary. We adopt here a combination of top-down journal level classification, in view of expanding the ERIH-Plus journal list [2], and a bottom-up clustering approach, where we use citation clusters to find candidate journals to be added to the list.

More in detail, we perform the following steps:

1. create a seed list of known journals in DH, by disseminating a survey to the participants to DH 2019 and in the Humanist mailing list, which resulted in obtaining 14 replies;

2. consider a fine-grained clustering of all publications within each citation index, obtained by using the Leiden algorithm (Traag et al., 2019) and following the heuristics proposed in Waltman and van Eck (2012);

[1] A reasoned review of quantitative analyses of the digital humanities is maintained by Scott Weingart at http://scottbot.net/dh-quantified

[2] See https://dbh.nsd.uib.no/publiseringskanaler/erihplus

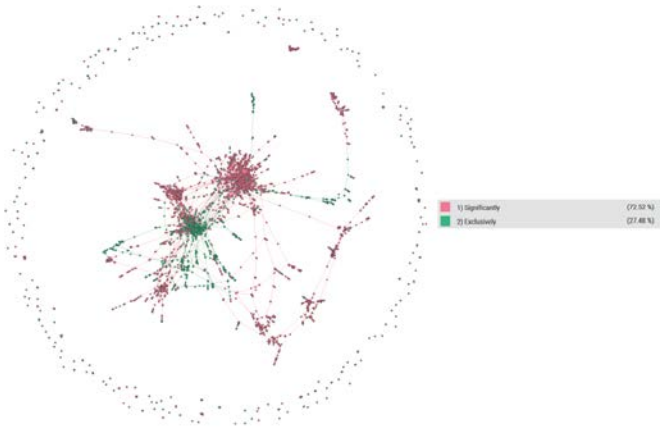


Figure 1: DH articles in Dimensions according to their journal category: "exclusively" (green) and "significantly" (pink). Only articles with with one citation or more (34.4% of the total) were considered. The network was visualized using Gephi 0.9.2 and the Force Atlas 2 layout. The size of the nodes (i.e., the articles) is proportional to number of citations they receive.

3. detect which clusters contain a relatively high proportion of publications from the journals in the list, in order to identify those which are also highly represented in the detected clusters but that are not part of the list obtained in point 1. We do this by considering the top 5 journals (by number of articles) per cluster with more than 5% of articles already from DH journals within the list;

4. manually assess each journal in the set identified in the previous point so as to add it in the original list; and

5. iterate again from point 2 until a convergence criterion is met.[3]

As convergence criterion, we iterate the proposed method twice (i.e., seed plus first iteration) and focus exclusively on research articles or review articles as publication typologies, published from the year 2000. This approach follows previous work in citation indexing for the humanities (Colavizza et al., 2018).

2.1 Journal classification

Given the highly interdisciplinary character of most publications in DH, we classify journals in the list using three categories: *exclusively*, if we deem a journal to be solely devoted to DH; *significantly*, if we deem that at least 50% of publications in the journal can be considered DH; *marginally*, if the journal contains an estimated 5% to 50% publications in DH. Categories were assigned by survey participants (iteration 1) or the authors (iteration 2) independently, and disagreements solved by majority. We acknowledge as a limitation the subjective perspective and biases this approach might have introduced in the resulting list of journals.

3 Results

The first outcome of our study is a list of DH journals (Spinaci et al., 2019), arranged according to the proposed categories, and containing 19 "exclusively", 17 "significantly" and 64 "marginally" classified journals.[4]

[3]Due to data access constraints, we worked with the following versions of the citation indexes under consideration: Web of Science: December 2018, Scopus: May 2019, Dimensions: December 2018, Crossref: August 2018. Coverage results might be affected accordingly.

[4]The list has been also made available in a Google sheet (https://tinyurl.com/y6rfrsuw) which can be commented for further feedback.

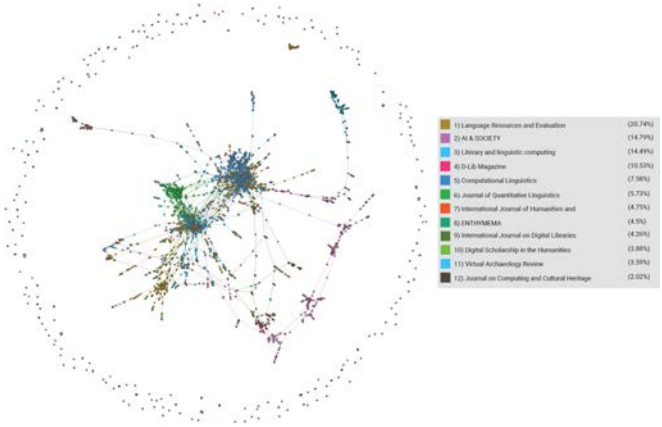


Figure 2: DH articles in Dimensions according to their related journal, if it contained more than 2% of the visible articles. The underlying network and layout is created as in Figure 1. The journal names are: *1) Language Resources and Evaluation; 2) AI & Society; 3) Literary and Linguistic Computing; 4) D-Lib Magazine; 5) Computational Linguistics; 6) Journal of Quantitative Linguistics; 7) International Journal of Humanities and Arts Computing; 8) Enthynema; 9) International Journal on Digital Libraries; 10) Digital Scholarship in the Humanities; 11) Virtual Archaeology Review; 12) Journal on Computing and Cultural Heritage*.

Table 1: Database coverage for journals in the three categories.

Journal	**WoS**	**Scopus**	**Crossref**	**Dimensions**
Exclusively	1243	1858	3989	2751
Significantly	1096	4421	5395	7259
Marginally	39,655	40,439	55,126	117,782
Total	41,994	46,718	64,510	127,792

3.1 Coverage

The overall database coverage, expressed as the number of indexed articles per category, is shown in Table 1. Crossref has the best coverage with respect to "exclusively" journals, while Dimensions has better coverage in the "significantly" and "marginally" categories.

The publication coverage of journals in the "exclusively" category is shown in Table 2. Only few journals show a good coverage, including some non-active ones: *Computers and the Humanities*, *Digital Scholarship in the Humanities*, *International Journal of Humanities and Arts Computing*, *Journal on Computing and Cultural Heritage*, and *Literary and Linguistic Computing*. Many other DH journals we collected in (Spinaci et al., 2019) were either poorly represented or even not present in the indexes we used for the analysis. Crossref appears to be the most comprehensive database in this respect.

3.2 Map of research

We further present a preliminary map of DH research, focusing on journal articles from the "exclusively" and "significantly" categories. We chose Dimensions for this analysis in order to better explore its apparently complementary coverage with respect to both categories and with respect to previous work (Gao et al., 2017; Tang et al., 2017; Gao et al., 2018). Coverage in the "significantly" category is mostly due to work in computational linguistics and digital libraries: *Journal of Quantitative Linguistics* (574 articles), *Computational Linguistics* (759), *D-Lib Magazine* (1054), *Language Resources and Evaluation* (2076). We also highlight the presence of almost 1500 articles from the journal *AI & Society*, a topic of increasing interest in DH. The map shown in Figures 1, 2, 3 considers all articles with at least one (given or received) citation, that is to say articles with a degree of one or more. The network initially contains 10,010 articles and 5,283 citation edges, while the number of articles with citations is 3,446 (34.4% of the

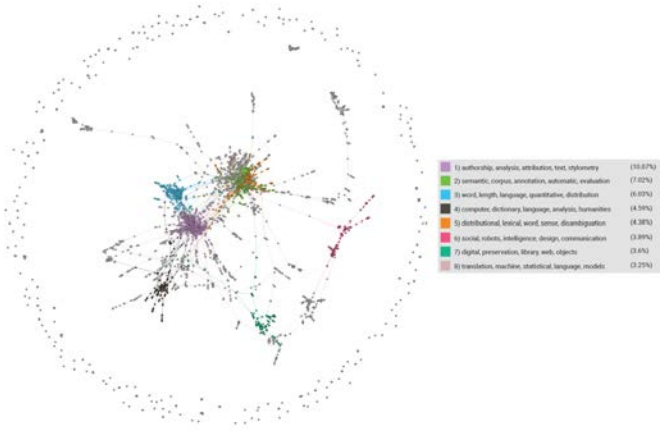


Figure 3: DH citation clusters calculated using the citations available in Dimensions, considering only the clusters containing more than 2% of the visible articles in the dataset. The legend contains the five most frequently occurring words in the titles of the articles within each cluster, after filtering out uninteresting ones. The underlying network and layout is created as in Figure 1. The journal coverage for each cluster is as follows (we only include journals accounting for more than 10% of the articles within a cluster): *1) Literary and Linguistics Computing (39.7%), Language Resources and Evaluation (19.3%), Digital Scholarship in the Humanities (17.6%), Journal of Quantitative Linguistics (17%), 2) Computational Linguistics (55.4%), Language Resources and Evaluation (34.3%), 3) Journal of Quantitative Linguistics (91.2%), 4) Language Resources and Evaluation (91.1%), 5) Language Resources and Evaluation (55%), Computational Linguistics (32.5%), 6) AI Society (99%), 7) International Journal on Digital Libraries (50.8%), D-Lib Magazine (37.1%), 8) Computational Linguistics (53.6%), Language Resources and Evaluation (34.8%)*

total). Two thirds of the articles are not connected to any other article through citations. The network's layout was created using Force Atlas 2 (Jacomy et al., 2014).

Figure 1 shows the articles assigned to the categories "exclusively" and "significantly". The bulk of the articles in "exclusively" journals include, somewhat predictably, the articles in *Literary and Linguistic Computing* or its successor, *Digital Scholarship in the Humanities*. However, we noticed that, when considering the most represented journals in our dataset (i.e., those with most articles) in Figure 2, the articles are visually arranged by journal and tend to follow field-specific patterns: digital libraries (*IJDL*, *D-Lib*), computational linguistics (*Computational Linguistics*, the *Journal of Quantitative Linguistics*), artificial intelligence and society (*AI & Society*), and DH (*Literary and Linguistic Computing*, *Digital Scholarship in the Humanities*). Instead, the articles in *Language Resources and Evaluation* are more evenly spread across different field clusters.

When we consider citation clusters detected using a modularity-maximizing method (Blondel et al., 2008), in Figure 3, we observe a modular structure with a high correlation with respect to the publication venue. The main focus of the DH cluster (number 1 in Figure 3) are quantitative literary studies, e.g., stylometry and authorship attribution. Other clusters cover the DH-related areas of computational linguistics and natural language processing (2,3,4,5,8), digital libraries (7) and AI and society (6). As far as we could notice from this graph, DH publications tend to connect to related disciplinary areas, even if each area maintained its distinctiveness. The publication venue remains a key trait of the intellectual structure of the DH.

4 Conclusion

In this article, we proposed an approach to find digital humanities publications by iterating between a list of journals (top-down) and its expansion using citation clustering (bottom-up). In this way, we were able to propose a first version of a list of digital humanities journals split in three categories: those that are "exclusively", "significantly" and "marginally" related to the digital humanities. We assessed the

Table 2: Database coverage for journals exclusively devoted to digital humanities scholarship.

Journal	**WoS**	**Scopus**	**Crossref**	**Dimensions**
Computers and the Humanities	663	806	1465	
Digital Humanities Quarterly (DHQ)		38		
Digital Medievalist			67	66
Digital Scholarship in the Humanities (DSH)	254	237	327	388
Digital Studies / Le champ numérique			230	
Digitális Bölcsészet / Digital Humanities				19
Frontiers in Digital Humanities			58	66
International Journal of Digital Humanities				
International Journal of Humanities and Arts Computing		15	253	475
Journal of Cultural Analytics			13	27
Journal of Data Mining and Digital Humanities				
Journal of Digital Archives and Digital Humanities				
Journal of Digital Humanities				
Journal of the Japanese Association for Digital Humanities			15	21
Journal of the Text Encoding Initiative			73	
Journal on Computing and Cultural Heritage (JOCCH)	75	178		202
Literary and Linguistics Computing	251	584	1465	1450
Revista de humanidades digitales			23	37
Umanistica Digitale				
Total	1243	1858	3989	2751

coverage of Web of Science, Scopus, Crossref and Dimensions in this respect, finding that Crossref has the best coverage of "exclusively" digital humanities journals, while Dimensions has the best coverage of the number digital humanities-related articles overall. We discussed a first map of research using citation data from Dimensions, highlighting how just one third of the articles in Dimensions are connected with each other via citations. We further found that digital humanities articles are connected via citations to computational linguistics and natural language processing, digital libraries and other developing areas such as AI and society. Nevertheless, we also found that the venues (i.e., the journals) strongly overlap with citation clusters, and are a key trait of the intellectual organization of digital humanities research.

We acknowledge that our work is still in progress, and thus it has a set of limitations which we plan to address in the future. In particular, we plan to include additional bibliographic entity types in addition to journal articles (e.g., books), and to also include the COCI (Heibi et al., 2019) and Microsoft Academic citation indexes to the comparison. Coverage will also be assessed at the article level (i.e., which citation index contains which articles) and chronologically. Lastly, we will elaborate on the map of research by including a comparison across all indexes.

Acknowledgements

The authors would like to thank the Centre for Science and Technology Studies (CWTS), Leiden University, for providing access to their databases and computing facilities. This work was in part conducted when Spinaci was visiting CWTS.

References

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Jennie M. Burroughs. 2017. No Uniform Culture: Patterns of Collaborative Research in the Humanities. *portal: Libraries and the Academy* 17(3):507–527. https://doi.org/10.1353/pla.2017.0032

Giovanni Colavizza, Matteo Romanello, and Frédéric Kaplan. 2018. The references of references: a method to enrich humanities library catalogs with citation data. *International Journal on Digital Libraries* 19(2-3):151–161. https://doi.org/10.1007/s00799-017-0210-1

Jin Gao, Oliver Duke-Williams, and Simon Mahony. 2017. The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis. In *Digital Humanities Conference Proceedings*.

Jin Gao, Julianne Nyhan, Oliver Duke-Williams, and Simon Mahony. 2018. Visualising the Digital Humanities Community: A Comparison Study Between Citation Network and Social Network. In *Digital Humanities Conference Proceedings*.

Martin Grandjean. 2016. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities* 3(1). https://doi.org/10.1080/23311983.2016.1171458

Björn Hammarfelt. 2016. Beyond Coverage: Toward a Bibliometrics for the Humanities. In Michael Ochsner, Sven E. Hug, and Hans-Dieter Daniel, editors, Research Assessment in the Humanities, Springer International Publishing, Cham, pages 115–131.

Anne-Wil Harzing and Satu Alakangas. 2016. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics* 106(2):787–804. https://doi.org/10.1007/s11192-015-1798-9

Ivan Heibi, Silvio Peroni, and David Shotton. 2019. COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics* https://doi.org/10.1007/s11192-019-03217-6

Daniel W. Hook, Simon J. Porter, and Christian Herzog. 2018. Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics* 3. https://doi.org/10.3389/frma.2018.00023

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 9(6):e98679. https://doi.org/10.1371/journal.pone.0098679

Loet Leydesdorff and Alkim Almila Akdag Salah. 2010. Maps on the basis of the Arts & Humanities Citation Index: The journals Leonardo and Art Journal versus "digital humanities" as a topic. *Journal of the American Society for Information Science and Technology* 61(4):787–801. https://doi.org/10.1002/asi.21303

Anton J. Nederhof. 2006. Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics* 66(1):81–100. https://doi.org/10.1007/s11192-006-0007-2

Julianne Nyhan and Oliver Duke-Williams. 2014. Joint and multi-authored publication patterns in the Digital Humanities. *Literary and Linguistic Computing* 29(3):387–399. https://doi.org/10.1093/llc/fqu018

Alkim Almila Akdag Salah, Andrea Scharnhorst, and Sally Wyatt. 2015. Analysing an Academic Field through the Lenses of Internet Science: Digital Humanities as a Virtual Community. In Thanassis Tiropanis, Athena Vakali, Laura Sartori, and Pete Burnap, editors, Internet Science, Springer International Publishing, Cham, volume 9089, pages 78–89. https://doi.org/10.1007/978-3-319-18609-2_6

Gianmarco Spinaci, Giovanni Colavizza, and Silvio Peroni. 2019. List of digital humanities journals. https://doi.org/10.5281/zenodo.3406564

Rachele Sprugnoli, Gabriella Pardelli, Federico Boschetti, and Riccardo Del Gratta. 2019. Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale. *Umanistica Digitale* 5. https://doi.org/10.6092/issn.2532-8816/8581

Muh-Chyun Tang, Yun Jen Cheng, and Kuang Hua Chen. 2017. A longitudinal study of intellectual cohesion in dig-ital humanities using bibliometric analyses. *Scientometrics* 113(2):985–1008. https://doi.org/10.1007/s11192-017-2496-6

Vincent A. Traag, Ludo Waltman, and Nees J. van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* 9(1). https://doi.org/10.1038/s41598-019-41695-z

Ludo Waltman and Nees Jan van Eck. 2012. A new methodology for constructing a publication-level classification system of science: A New Methodology for Constructing a Publication-Level Classification System of Science. *Journal of the American Society for Information Science and Technology* 63(12):2378–2392. https://doi.org/10.1002/asi.22748

Epistolario De Gasperi: National Edition of De Gasperi's Letters in Digital Format

Sara Tonelli[†], Rachele Sprugnoli[‡], Giovanni Moretti[†‡], Stefano Malfatti[◇], Marco Odorizzi[◇]

[†]Fondazione Bruno Kessler, Trento, Italy
[‡]Università Cattolica, Milan, Italy
[◇]Fondazione Trentina Alcide De Gasperi, Trento, Italy
{satonelli,moretti}@fbk.eu
rachele.sprugnoli@unicatt.it
s.malfatti@epistolariodegasperi.it
modorizzi@degasperitn.it

Abstract

English. We present an ongoing project aimed at creating the National Edition of Alcide De Gasperi's letters in digital format. Our main goal is to systematically collect and transcribe a large number of private and public letters, present in different archives, written or received by De Gasperi throughout his life, and to shed light into all the critical steps of his biography and of the major events of the time. Thirty-eight transcribers have already acquired and annotated 1,549 letters, using an ad-hoc tool specifically developed for the project. All the transcribed material is available through a constantly updated free access web platform.

Italiano. In questo contributo presentiamo un progetto tuttora in corso volto a creare l'edizione nazionale delle lettere di Alcide De Gasperi in formato digitale. Il nostro obiettivo principale è quello di raccogliere e trascrivere sistematicamente un gran numero di lettere private e pubbliche, presenti in diversi archivi, scritte o ricevute da De Gasperi nel corso della sua vita, e di far luce su tutti i passaggi critici della sua biografia e dei principali eventi del tempo. Trentotto trascrittori hanno già acquisito e annotato 1.549 lettere, utilizzando uno strumento appositamente sviluppato per il progetto. Tutto il materiale trascritto è disponibile attraverso una piattaforma web ad accesso libero constantemente aggiornata.

1 Introduction

Alcide De Gasperi (1881-1954) was one of the most important statesmen in European history, recognized as the father of the Italian Republic and one of the founding fathers of Europe. Despite the rich historiography about him, many aspects of his biography are still waiting to be enlightened. Indeed, up until now, the historical analysis on De Gasperi has been based almost exclusively on institutional sources, on his published works and on his public speeches (De Gasperi, 2006, 2008a,b, 2009), also available in digital format[1] (Sprugnoli et al., 2016). What is still missing is his extraordinary correspondence, which covers all the critical steps of his biography and often offers a different point of view on many events of the time. So far, only very few letters written or received by De Gasperi have been made available through anthological publications, see (De Gasperi, 1955, 1974) among others, or through the publication of the correspondence with specific people, e.g. (Antonazzi, 1999). Nevertheless, a large number of letters has not been published yet. This has paved the way to this ambitious project, coordinated by a scientific committee that brings together all the major scholars of De Gasperi, whose goal is to provide an exhaustive collection of De Gasperi's correspondence in digital format, covering both his public and private life. A preliminary search of the letters to be digitized and edited lists about 5,000 documents written in multiple languages, currently stored in 114 archives all over the world. More than 1,500 of these letters are already available online at the time of writing, all searchable and accessible through an online platform: https://epistolariodegasperi.it/#/archivio_digitale/lettere.

The Italian Ministry of Culture supports the project, which has been recognized as a "National Edition". Significantly, this is the first Italian National Edition exclusively conceived through digital tools.

[1] http://alcidedigitale.fbk.eu/

2 Related Work

Editorial projects involving the digitization of documents are always increasing and the standards that characterize them are constantly evolving (Pierazzo, 2014). The literature reports a rich and varied panorama of digital editions of works by writers, intellectuals and historical figures (Franzini et al., 2016). In recent years, also the digitization of letters has received much attention (Hankins, 2015) with projects dedicated to individual characters, such as Darwin[2] and Van Gogh[3], or that aim to reconstruct large epistolary networks (Ravenek et al., 2017; Baillot and Seifert, 2013). A centralized web service to search within the metadata of diverse scholarly editions of letters has been also developed (Dumont, 2016). If we look at the Italian landscape, we can see that the development of digital editions of letters is rather limited and that current initiatives have mostly focused on the collection and digitization of literary documents created in the Late Medieval and modern era, such as in the Progetto Datini[4] and in the digital edition of the letters by Vespasiano da Bisticci (Tomasi, 2012).

In our opinion, the project presented in this paper has at least four significant aspects to be highlighted if compared with other works and that constitute our contribution: (i) the strong interdisciplinary synergy that led to define a rich set of metadata and editorial choices taking into consideration different aspects (i.e., linguistic, historical, philological); (ii) the development of a transcription infrastructure, characterized by three access levels and a graphical interface that is both intuitive and compliant with existing editing standard and that we plan to release for the research community so to be adopted in other projects; (iii) the subject of the edition, that is the exhaustive collection of correspondence related to contemporary history until now closed in archives and difficult to reach; (iv) the attention to different types of final users so that mechanisms have been implemented in order to allow a customization of the reading complexity.

3 Editorial Practice

Several transcribers, mostly history scholars, have been involved in the project to digitally acquire De Gasperi's correspondence. They have been provided with an ad-hoc tool (Moretti et al., 2018), through which they have inserted various metadata for each document, including: sender, recipient, chrono-topical data, type of document (letter, telegram etc.)[5] (Malfatti, 2019). They also specify if the document is an original, a copy or a draft and if the topic is of a personal or institutional nature. Presence of envelope, letterhead, autograph signatures, attachments have to be reported as well. Transcribers must also indicate for each document the number of cards, writing technique (manuscript/typewritten), conservation status, as well as the reference code for the correct identification of the document. Transcriptions must accurately reflect the text and the layout of the document, including possible mistakes, illegible words and cancellations, punctuation and paragraph division, etc. Margin notes have to be reported in a specific field, possibly with the corresponding position and author. Historical (or commentaries) notes clarify and integrate the content of the text. The transcript must also be preceded by an abstract to contextualise and summarize its content.

4 Transcription Tool

For the acquisition of De Gasperi's letters we developed a new infrastructure that allows different transcribers to work in parallel and to smoothly publish the transcribed documents online as soon as they have been revised by a small pool of experts. Specifically, the infrastructure functionalities are the following:

- possibility to work both offline and online, so that the transcriptions can be performed also in archives without an Internet connection;
- use of a wide set of metadata defined by the Scientific Committee and compliant with the Dublin Core standard and the standard for the cataloging of manuscripts in Italian libraries;

[2] https://www.darwinproject.ac.uk/
[3] http://vangoghletters.org/
[4] http://datini.archiviodistato.prato.it/
[5] The editorial guidelines are available online: https://epistolariodegasperi.it/#/risorse

- clear division of roles into three types of contributions: (i) the transcriber, who creates a digital picture of the original document and uploads it in the transcription tool, inserts metadata, transcribes and annotates the text, (ii) the supervisor, who adds the critical apparatus, and (iii) the editor, who is in charge of validating the correctness of the transcription and publishes the letter in the final online platform;

- automatic upload and update of each letter on a central database; (v) easy-to-use transcription interface not requiring explicit knowledge of tag annotation and coding schemes;

- easy-to-use transcription interface not requiring explicit knowledge of tag annotation and coding schemes;

- presence of autocomplete features to reduce the risk of mistakes in the insertion of metadata.

From the technical point of view, the infrastructure includes a MySQL database and three software applications (one for each role) written in Javascript/ECMAScript 6. The interface is implemented using the ReactJS framework.

5 Digital Archive

Currently, the digital archive allows to search by title, sender, recipient, type (i.e., letter, telegram, postcard, note, other), theme (i.e. private life, national politics, international politics, local politics, religion, culture, economy, society, political party), free text. A slider can be used to select a specific time period from 1902 to the year of De Gasperi's death in 1954. The interface responds in real time to what the user types, modifying the list of letters that correspond to the search key. Searches can also be combined: for example, it is possible to search for all the notes sent by Giulio Andreotti after 1950 or for all the postcards containing the word "augurio". When clicking on a letter, all the metadata compiled by the transcriber are displayed together with an abstract written by a historian that summarizes the content of the document. The photograph and the transcription are placed next to each other and the transcription respects the annotation made by the transcribers showing, among others, words that are underlined, erased parts, spelling errors. Footnotes explain content that could be unclear to the reader adding contextualization. In addition, 380 proper names of persons are accompanied by a biographical note written by historians. Given that the aim of the project is to attract both general public and experts, readers can access the National Edition from different perspectives. More specifically, with a simple click the user goes from a complete transcription of the letter containing all the annotations to a simplified one that is more suitable for the general public: in this reader-friendly view of the letter the deleted parts are not displayed, the typos are corrected and the abbreviations are extended.

6 Current State of the Project

The transcription process started in August 2018 and it currently involves 38 transcribers and 10 supervisors. At the moment of writing (September 2019), 1,549 letters from 106 different private and public archives in Italy and abroad have been transcribed, annotated, revised and published online. The number of letters written in each year currently present in the digital archive is shown in Figure 1. As expected, most letters are exchanged between 1945 and 1954, when De Gasperi was a very prominent political figure with key roles in the Italian government. However, we observe an interesting element for the years between 1926 and 1942, which historians describe as De Gasperi's *internal exile*, because he did not cover official roles under fascism: while the archive of public documents of Alcide De Gasperi[6] contains only few documents for that period (Moretti et al., 2016), 132 letters have been found for the same years, and 24 of them have been tagged as being about *Politics*. This shows that, even if De Gasperi did not have any official role, he was still involved in political discussions and was expressing his opinion on the current situation.

[6]Available at http://alcidedigitale.fbk.eu/

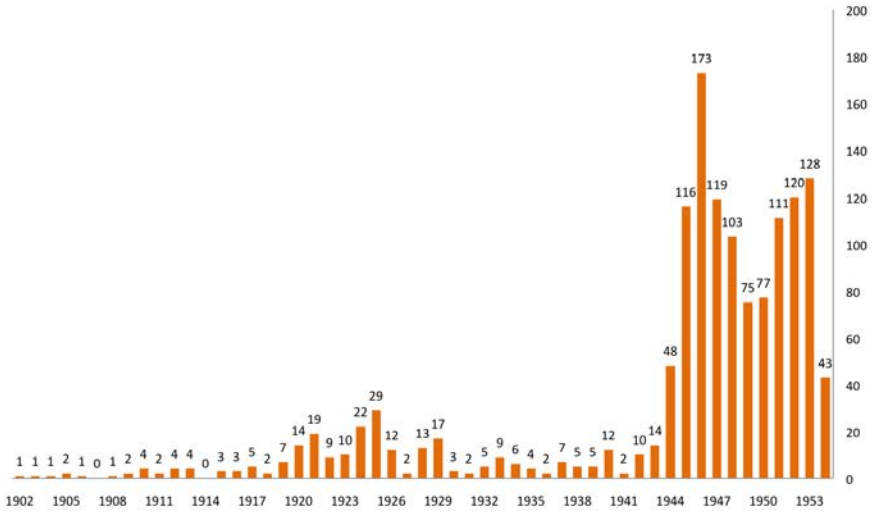


Figure 1: Letter distribution over time

Additional statistics are reported in Table 1, where we list the five senders and receivers that are currently most present in the correspondence. De Gasperi appears in most letters either as sender and as receiver. Only in few examples he is not explicitly mentioned, for instance when the receiver is a collective entity (e.g. 'Governo italiano', 'Soci della Tridentum') where De Gasperi was included. Some persons appear in the table because a version of their correspondence had already been curated and published before, see for example Sturzo (Antonazzi, 1999), so that the Epistolario project could take advantage from already transcribed blocks of letters.

Senders	**#**
Alcide De Gasperi	816
Luigi Sturzo	143
Piero Malvestiti	71
Luigi Granello	39
Guido de Gentili	28
Agostino Gemelli	22

Receivers	**#**
Alcide De Gasperi	704
Luigi Sturzo	107
Giulio Delugan	55
Piero Malvestiti	41
Giuseppe Micheli	34
Amintore Fanfani	31

Table 1: Top-5 senders (left) and receivers (right) in the digital collection at the moment of writing.

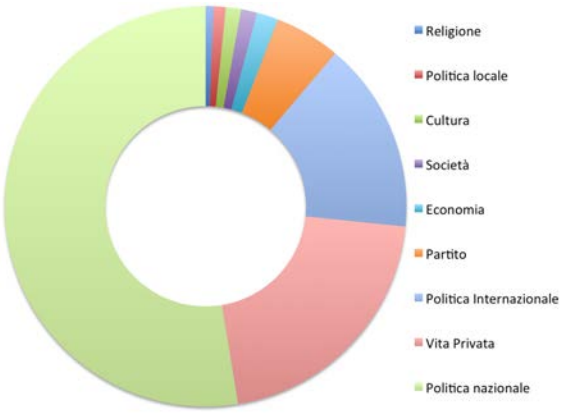


Figure 2: Topic distribution in the letters

Another interesting analysis is related to the topic(s) of the letters. Indeed, each transcriber has to assign one or more topics to each letter chosen among the following categories: Private Life, National Politics, International Politics, Local Politics, Political Party, Economy, Society, Culture, Religion. The choice to allow multiple annotations was guided by domain experts, who identified in each letter multiple

topics and opted not to select only the main one. We report in Figure 2 the distribution of topics in the letters transcribed so far. As expected, national politics is the most present topic. The fact that private life is the second most frequent topic, instead, is rather surprising, since the majority of De Gasperi's letter exchanged with other family members are still kept private by his descendants, and have not been included in the Epistolario. This means that De Gasperi tended to mention personal information and tell about private aspects of his life in letters to members of the christian-democratic party, politicians and other public figures, mixing political, cultural and private topics.

Figure 3: Keyphrases extracted from eighteen English letters sent by/to De Gasperi.

Another intriguing aspect of the collection is the fact that, given the international role played by De Gasperi, especially after WWII, several letters are either in English or in German (resp. 33 and 10 at the moment of writing). Figure 2 shows a preliminary content analysis of the English transcribed letters performed with the keyword extraction tool KD (Moretti et al., 2015). The letters, dated 1945-1948, were exchanged between De Gasperi and important US personalities, such as President Truman and the Secretary of Defense Robert A. Lovett. The focus of these letters is on the international cooperation between the two nations, the treaty after WW2, the necessity of peace for Italy reconstruction and the role of supranational communities.

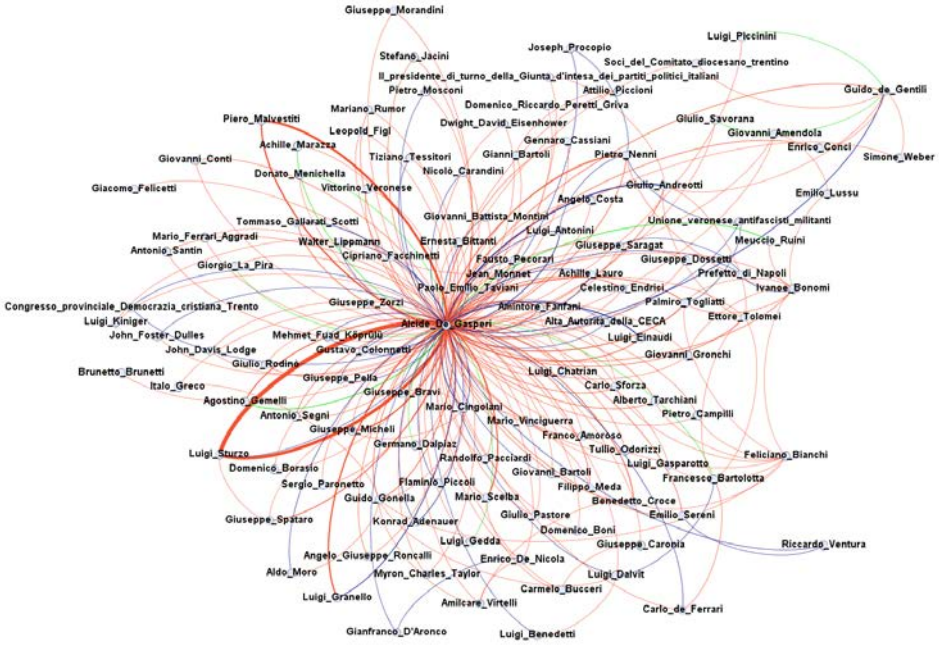


Figure 4: Network of letters

Figure 4 presents the correspondence network of the collection at the time of writing, considering only exchanges of at least two letters. The network is a direct graph in which node and edge size is related to the amount of exchanged letters and edge colour represents different types of document. We can notice

that the strong majority are letters (75%, in red) but there are also telegrams (20%, in blue) and short notes (4%).

7 Conclusion and Future Work

A digital edition presents, with respect to a print edition, new opportunities both for curators and users. First of all, there are virtually no limits to the amount of material that can be included in a digital collection; in addition, archiving on a web platform makes the organisation and management of documentary material more flexible, and provides users with interactive tools that are easier and more attractive for non-experts. At the same time, however, a digital project also requires an interdisciplinary approach. In particular, it is necessary to involve different competences already in the planning phase to create effective synergies between historians, DH experts, computer scientists, archivists and publishers. The National Edition of De Gasperi's letters in digital format is an example of this interdisciplinary approach: all the choices about the documentary editing are the results of discussions among the project members that have guided the choices implemented in the platform, for example the possibility to annotate the letters from different perspectives (linguistic, historical, philological).

The development of the digital archive is still ongoing thus several improvements and additions are planned for the future. Search options will be extended to take advantage of the many metadata provided by the transcribers through the transcription tool: e.g. language used in the letter, presence or absence of headed paper, presence or absence of a handwritten signature. Thematic paths will also be designed to guide the user in discovering the collection of letters, for example by navigating the various biographical phases of De Gasperi's life or the relationship between De Gasperi and his correspondents. Moreover, it will be possible to browse the letters stored in a specific archive or library using an interactive map. Users will also be allowed to download the letters in different format, including pdf and TEI-XML. In addition, we plan to release to the research community the transcription infrastructure with a Creative Commons 4.0 license.

Acknowledgments

Marco Odorizzi wrote Section 1 of this paper, Stefano Malfatti wrote Section 3, while Rachele Sprugnoli, Sara Tonelli and Giovanni Moretti equally contributed to the analyses and the writing of the other sections.

References

Giovanni Antonazzi. 1999. *Luigi Sturzo-Alcide De Gasperi. Carteggio*. Istituto Luigi Sturzo, Roma.

Anne Baillot and Sabine Seifert. 2013. The project" berlin intellectuals 1800–1830" between research and teaching. *Journal of the Text Encoding Initiative* (4).

Alcide De Gasperi. 1955. *Lettere dalla prigione*. Milano, Mondadori.

Alcide De Gasperi. 1974. *De Gasperi scrive: Corrispondenza con capi di Stato, cardinali, uomini politici, giornalisti, diplomatici*, volume 1. Morcelliana.

Alcide De Gasperi. 2006. Alcide De Gasperi nel Trentino asburgico. In *Scritti e discorsi politici di Alcide De Gasperi*, Il Mulino, volume 1.

Alcide De Gasperi. 2008a. Alcide De Gasperi dal Partito popolare italiano all'esilio interno 1919-1942. In *Scritti e discorsi politici di Alcide De Gasperi*, Il Mulino, volume 2.

Alcide De Gasperi. 2008b. Alcide De Gasperi e la fondazione della Democrazia cristiana, 1943-1948. In *Scritti e discorsi politici di Alcide De Gasperi*, Il Mulino, volume 3.

Alcide De Gasperi. 2009. Alcide de Gasperi e la stabilizzazione della Repubblica 1948-1954. In *Scritti e discorsi politici di Alcide De Gasperi*, Il Mulino, volume 4.

Stefan Dumont. 2016. correspsearch–connecting scholarly editions of letters. *Journal of the Text Encoding Initiative* (10).

Greta Franzini, Melissa Terras, and Simon Mahony. 2016. A catalogue of digital editions. *Digital scholarly editing: Theories and practices* pages 161–182.

Gabriel Hankins. 2015. Correspondence: Theory, practice, and horizons. *Literary Studies in the Digital Age* .

Stefano Malfatti. 2019. Per un'edizione online dell'epistolario di alcide de gasperi. criteri di digitalizzazione, schedatura, regestazione ed edizione di lettere del novecento. In *Book of Abstract of AIUCD 2019. Teaching and research in Digital Humanities' era*. AIUCD, pages 236–240.

Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems* 111:100–112.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the Dirt: Extracting Keyphrases from Texts with KD. In *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2018. Lettere: Letters transcription environment for research. In *Book of abstract of AIUCD 2018*.

Elena Pierazzo. 2014. Digital documentary editions and the others. *Scholarly Editing* 35:1–23.

Walter Ravenek, Charles van den Heuvel, and Guido Gerritsen. 2017. The epistolarium: Origins and techniques. *CLARIN in the low countries* pages 317–323.

Rachele Sprugnoli, Giovanni Moretti, Sara Tonelli, and Stefano Menini. 2016. Fifty years of european history through the lens of computational linguistics: the de gasperi project. *IJCol-Italian journal of computational linguistics* 2(2):89–100.

Francesca Tomasi. 2012. L'edizione digitale e la rappresentazione della conoscenza. un esempio: Vespasiano da bisticci e le sue lettere. *Ecdotica* 9.

Visualizing *Romanesco*; or, Old Data, New Insights

Gianluca Valenti
Université de Liège
gianluca.valenti@uliege.be

Abstract

English. The evolution of the Roman language over time is a problematic topic, which has been analysed by many scholars and different points of view. Nonetheless, we do not know enough about the process of its Tuscanisation. With the help of some data visualisations, I show that from the sole analysis of the language of all known sources, it is not possible to clarify this issue. Conversely, new attention given to the physical supports that transmit the Roman documents can provide new, interesting insights.

Italiano. L'evoluzione dell'antico romanesco è un argomento analizzato da numerosi studiosi ma, ad oggi, ancora non compreso in ogni suo aspetto. In particolare, non si è giunti a un accordo su tempistiche e modalità con cui avvenne il processo della sua toscanizzazione. Uno sguardo d'insieme ai testi con una o più caratteristiche di romanesco antico chiarisce come la questione non possa essere risolta se l'attuale corpus è indagato esclusivamente nei suoi aspetti linguistici; per ottenere nuove informazioni bisognerà dunque volgersi allo studio dei supporti materiali che trasmettono tali documenti.

1 The sociolinguistic background

In the past decades, dialectology studies have exponentially increased. All linguists nowadays acknowledge the importance of the dialects in the social and cultural history of humanity, and put them on an equal footing with standard languages. In particular, in the context of Italian sociolinguistics, the study of the dialect of Rome is considered as an extremely relevant topic.

Since the beginnings, Italian dialects are divided into three main groups: Northern, Tuscan, and Southern varieties. The medieval and renaissance periods marked an irreversible revolution in the Roman social background, and consequently, in the Roman language—the *Romanesco*.

Although *Romanesco* formed part of the Southern dialects, before the second half of the 16th century at the latest, it came to resemble the Tuscan varieties—a process called 'Tuscanisation' (Ernst, 1970). This change, whose dynamics remain largely unclear, represents a unique episode in the history of the Italian language. The uniqueness of the 'Rome case' has been stressed on many occasions, and several explanations have been proposed for it, but no agreement has been reached yet.[1] Indeed, the mutation of *Romanesco* was so deep (and without similar precedents) that scholars tend to refer to it as 'disintegration, decay' (Migliorini, 1932), instead of 'evolution' (as is the usual case for most of the languages).

More than fifty years after Migliorini's work, Mancini (1987:41) argued that, until then, scholars 'placed too much emphasis on some demographically macroscopic events,' such as the Sack of Rome of 1527. Instead, according to him, the mutation of *Romanesco* was a slow event, already in place, at least in its germinal stages, in the *Trecento* and the *Quattrocento*. Trifone (1990:92) quickly reacted with a detailed linguistic analysis of new documents, and concluded that the demographic de-southernisation of Rome (caused by the sack) and the ensuing repopulation of the city post-1527 were the main reasons for the de-southernisation of the spoken language of its lowest social classes. The debate has continued for several years (cf. also Trifone 1992 and Mancini 1993) without a consensus.

From one side, it is unquestionable that, as De Caprio (1988:453) states, 'the sack of Rome of 1527 is a traumatic caesura in Roman cultural history,' but, from the other side, its role in the context of the Tuscanisation is admittedly unclear. Even at the present time, instead of positioning themselves on one side or

[1] Cf. e.g., Vignuzzi (1988, 1995), De Mauro (1989), Palermo (1991), and, recently, Coluccia (2011), who argues that the exceptionality of the phenomenon lies in the untimely Tuscanisation at a spoken level.

the other, scholars tend to report both opinions, at most trying to harmonise them into a single whole.[2] Currently, and maybe because no one has established a definitive answer to the issue, researchers seem to prefer to focus on the currently spoken *Romanesco*, a language that has still many points of contact with its renaissance variety.[3] Interest in the old *Romanesco* has not waned, though, as is clear by the recent organisation of the roundtable 'Rome in history, in linguistics and in literature' (Rome, 23th July 2016) and the international conference 'Il romanesco tra ieri e oggi', which I organised in Liège the 9th September 2019.

2 Old data, new insights

The old epistemological framework—where single scholars tried to explain the evolution of the Roman language by studying analytically one or few texts—has proved to be unfit for the task of understanding the dynamics of the Tuscanisation of *Romanesco*. At the present time, it is necessary to look at a broader picture, and consider the Roman texts as if they were a single whole. Indeed, up to now, scholars essentially based their findings on qualitative research, but did not exploit the potentialities of databases and digital tools. Most of the current papers focused on *Romanesco* analyse the language of a specific text (or a bunch of texts), while a general overview that takes into account the huge amount of data pertaining to the Roman sources collected by the scientific community during the past hundred years is still missing.

I recently made the first step in order to fill this gap, by building and putting online a database that allows users to make queries into the whole corpus of texts written in *Romanesco* from the Origins to 1550. It is available online, free of charges, at the address http://www.romanesco.uliege.be/. Regularly updated, the database makes available metadata concerning not only the texts, but also the physical supports (printed books, manuscripts, and places, such as churches or catacombs) that transmit them.

Working on digital data leads scholars to new findings, and allows them to answer old research questions, which could not be solved with the traditional approach. In this paper, with the help of some visualisations, I show that—due to the scarcity of the sources—we do not have sufficient data to get new insights about the language of Rome and its evolution through time if we only look at the languages of the texts. Therefore, to have a sharp understanding of the Tuscanisation process, we need to reanalyse the linguistic features of the epigraphs, whose language has been often defined, maybe too quickly, as generically 'Vernacular'.

2.1 Corpus and tool

To conduct this study, I put in plain text files the metadata of 372 texts, written from 800 to 1550, with at least some features of *Romanesco* in them. I took the data from D'Achille and Giovanardi (1984). With regard to the languages, notice that a text can contain some features that do not belong to its original linguistic system. In consequence, for each text, I identified its primary language and all the potential secondary languages (i.e., the languages that occur to a lesser extent). The condition for a text to be included in this corpus is to have *Romanesco* as its primary or, at least, its secondary language. The corpus is thus composed of texts written in *Romanesco*—which may contain some pieces of Tuscan or Latin—, but also of texts that contain only a small amount of features of *Romanesco*, while their primary language is Tuscan, Latin, or Vernacular.

Each text is transmitted by one physical support: a) 'places' transmit epigraphic texts; b) 'manuscripts' transmit handwritten texts; and c) 'printed books' transmit printed texts.

All the visualizations are made with the software *Tableau*.

2.2 Results

Figures 1–4 show the total number of occurrences of primary and secondary languages, and their evolution over time. The figures provide some interesting insights that, in a way, strengthen both hypotheses of Mario Mancini and Pietro Trifone.

[2] Cf. e.g., Vignuzzi (1994), Giovanardi (1998:61), D'Achille and Petrocchi (2004:122–123).
[3] Cf. e.g. the projects *VRC. Vocabolario del Romanesco Contemporaneo* and *ERC. Etimologie del romanesco contemporaneo*.

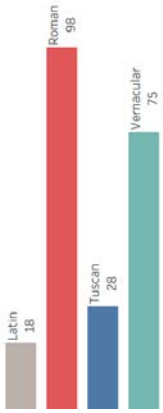


Figure 1. Primary languages

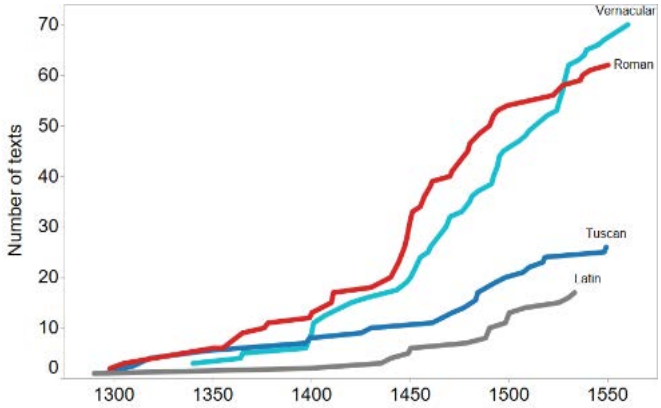


Figure 2. Primary languages over time (I)

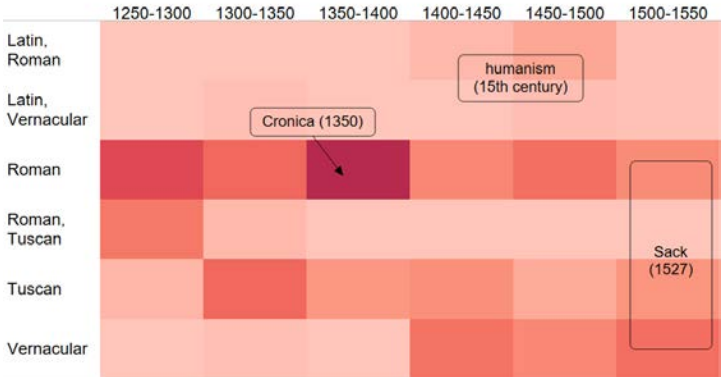


Figure 3. Primary languages over time (II)

Figure 4. Secondary languages

We notice that the total amount of occurrences of texts written in *Romanesco* and Tuscan starts to diverge (in favour of the former) in the 2nd half of the 14th century, maybe due to the success of Anonimo Romano's *Cronica* (1357–1358), the most important Roman text of its century.

However, the Tuscan language is attested all over the centuries, from the 14th to the 16th century. I do not register any dramatic increase right after 1527 (cf. figure 2). This observation seems to endorse Mancini's view of the Tuscanisation as a slow process, already in place in the *Quattrocento*.

On the other hand, if we look at figure 3 we notice, in the first half of the 16th century, a slight decrease of texts written in *Romanesco*, and a parallel growth of texts written in the Tuscan language. Admittedly, this outcome may be related—as Trifone states—to the de-southernisation of the Roman population after the Sack, and the subsequent increase in the number of Tuscan people moving to the town.

These visualisations improve significantly our perception of the evolution of *Romanesco* through time. Nonetheless, they do not provide any irrefutable evidence that would end the debate on the timing and modalities of its Tuscanisation. Therefore, we need to look at the problem from another perspective.

Indeed, an aspect has escaped the scrutiny of most of the past scholars: observing the physical supports that transmit Roman texts over the years, we notice some interesting insights.

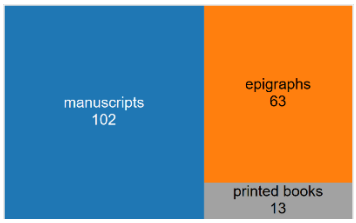


Figure 5. Physical supports

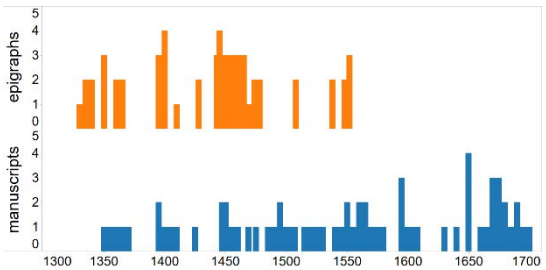


Figure 6. Physical supports over time

The prevalent supports of literary texts are, up until and including the 15th century, manuscripts, and afterwards, printed books. In the present corpus, though, the total number of epigraphs is significantly high (cf. figure 5). Furthermore, if we consider only the sources ranging from 800 to 1550, texts transmitted by epigraphs are even more than texts transmitted by manuscripts (cf. figure 6).[4] This is due to the fact that most of the texts of this corpus are practical documents (such as receipts, letters, and private notes), and have no literary value. Indeed, texts that are transmitted by perishable material—such as pieces of paper—get easily lost, while texts that are carved on the column of a church are more likely to be preserved.

By their very nature, epigraphs are dramatically short, and in consequence, linguistic features that are typical of a given area are less likely to be detected in epigraphs than in other textual typologies. The high number of epigraphs included in this corpus may be related to the high number of occurrences of texts written in a language that has been defined, generically, as 'Vernacular'. Therefore—maybe because of their apparently low linguistic value—the past surveys on the Tuscanisation of *Romanesco* did not take enough into account epigraphic texts.

However, the low number of handwritten documents of *Romanesco* and, in contrast, the high number of epigraphic texts, make the latter a critical source to understand the linguistic mutations in the medieval and renaissance Rome. Moreover, we should not forget that, as a starting point for the research, we have at disposal a solid documentary basis, the volume on Vernacular texts found in churches, edited by Sabatini et al. (1987). There, the authors provide detailed linguistic analyses, which could serve as a model for further studies focused on the language of the newly discovered epigraphic texts of the past thirty years.

Once we have collected a significant amount of new data, it will be possible to look again at the linguistic features of the Roman sources—including but not limited to manuscripts and printed books—thus refining our theories and reaching new conclusions about the Tuscanisation process of *Romanesco*.

3 Conclusions and new perspectives

Within the traditional approach, scholars tried to explain timing and causes of the Tuscanisation of *Romanesco* by analysing the linguistic features of a small selection of texts. The results of this approach—albeit essential in many respects—did not lead to a sharp understanding of this particular linguistic process. I have shown that the reason for this failure is not entirely due to the little number of texts analysed. Indeed, even though we consider all texts at our disposal, we are not able to recognise a clear pattern in favour of one or the other theory. In order to resolve this issue, we need more data, i.e., we need to look back at those texts that—until now—have been catalogued as written in 'Vernacular' language.

The high number of texts that we did not assign to any specific linguistic system is probably related to the high number of epigraphs that transmit them. Nowadays, scholars should approach the epigraphic texts with renewed attention, looking for pieces of evidence of their linguistic features. While awaiting additional archival findings, this is the only way to increase the number of texts whose language is known, which is our only chance to make new assumptions that could explain the Tuscanisation of *Romanesco*.

References

Alberto Asor Rosa. 2009. *Storia europea della letteratura italiana. I. Le origini e il Rinascimento*. Einaudi, Turin.

Rosario Coluccia. 2011. Scripta. In: R. Simone (ed.), *Enciclopedia dell'italiano*, volume 2, Istituto della Enciclopedia Italiana, Rome: 1287–1290.

Paolo D'Achille and Claudio Giovanardi. 1984. *La letteratura volgare e i dialetti di Roma e del Lazio. Bibliografia dei testi e degli studi. Vol. 1: Dalle origini al 1550*. Bonacci, Rome.

Paolo D'Achille and Stefano Petrocchi. 2004. *Limes* linguistico e *limes* artistico nella Roma del Rinascimento. In: V. Casale and P. D'Achille (eds.), *Storia della lingua e storia dell'arte in Italia. Dissimmetrie e intersezioni. Atti del III Convegno ASLI (Roma, 30–31 maggio 2002)*. Cesati, Florence: 99–137.

Vincenzo De Caprio. 1988. Roma. In: A. Asor Rosa (ed.), *Letteratura italiana. Storia e geografia*, volume 2/1, Einaudi, Turin: 327–472.

[4] I did not put in the plot the printed books, because only three of them are attested before the end of the 16th century. Similarly, I did not put in the plot the sources from the 9th to the 13th century, because they are extremely rare (only four epigraphs in total). Notice also that, even though I only consider texts ranging from 800 to 1550, the dating of the sources can be later.

Tullio De Mauro. 1989. Per una storia linguistica della città di Roma. In: Id. (ed.). *Il romanesco ieri e oggi. Atti del Convegno del centro romanesco Trilussa e del Dipartimento di Scienze del linguaggio dell'Università 'La Sapienza' (Roma, 12–13 ottobre 1984)*, Bonacci, Rome: XIII–XXXVII.

Gerhard Ernst. 1970. *Die Toskanisierung des romischen Dialekts im 15. und 16. Jahrhundert*. Niemeyer, Tübingen.

Claudio Giovanardi. 1998. *La teoria cortigiana e il dibattito linguistico nel primo Cinquecento*. Bulzoni, Rome.

Mario Mancini. 1987. Aspetti sociolinguistici del romanesco nel Quattrocento. *RR. Roma nel Rinascimento. Bibliografia e note* 3:38–75.

Mario Mancini. 1993. Nuove prospettive sulla storia del romanesco. In: Istituto Nazionale di Studi Romani (ed.), *«Effetto Roma». Romababilonia*, Bulzoni, Rome: 9–40.

Bruno Migliorini. 1932. Dialetto e lingua nazionale a Roma. Reprint in: Id. 1948. *Lingua e cultura*. Tumminelli, Rome: 109–123.

Massimo Palermo. 1991. Fenomeni di standardizzazione a Roma nel primo Cinquecento. *Contributi di Filologia dell'Italia mediana* 5:23–52.

Francesco Sabatini, Sergio Raffaelli and Paolo D'Achille. 1987. *Il volgare nelle chiese di Roma. Messaggi graffiti, dipinti e incisi dal IX al XVI secolo*. Bonacci, Rome.

Pietro Trifone. 1990. La svolta del romanesco tra Quattro e Cinquecento. Reprint in: Id. 2006. *Rinascimento dal basso. Il nuovo spazio del volgare tra Quattrocento e Cinquecento*. Bulzoni, Rome: 61–94.

Pietro Trifone. 1992. *L'italiano nelle regioni. Roma e il Lazio*. Utet, Turin.

Ugo Vignuzzi. 1988. Italienisch: Areallinguistik VII. Marche, Umbrien, Lazio / *Aree linguistiche VII. Marche, Umbria, Lazio*. In: G. Holtus, M. Metzeltin and C. Schmitt (eds.), *Lexikon der Romanistischen Linguistik*, volume 4, Niemeyer, Tübingen: 606–642.

Ugo Vignuzzi. 1994. Il volgare nell'Italia mediana. In L. Serianni and P. Trifone (eds.), *Storia della lingua italiana*, volume 2, Einaudi, Turin: 329–372.

Ugo Vignuzzi. 1995. Marche, Umbrien, Lazio / *Marche, Umbria, Lazio*. In: G. Holtus, M. Metzeltin and C. Schmitt (eds.), *Lexikon der Romanistischen Linguistik*, volume 2/2, Niemeyer, Tübingen: 151–169.

What is a Last Letter?
A Linguistic/Preventive Analysis of Prisoner Letters from the Two World Wars

Giovanni Pietro Vitali
University College Cork
giovannipietrovitali@gmail.com

Abstract

English. This paper aims to draw some preliminary analysis from the analyses carried out so far on the corpora collected during the first period of the Marie Curie project *Last Letters from the World Wars: Forming Italian Language, Identity and Memory in Texts of Conflict*. The project explores the linguistic and thematic features of the last letters of people who were sentenced to death during the two world wars. Following the creation of a corpus of letters, written by prisoners in the two world wars, substantial differences in the language and contents of these documents were revealed. These differences are due to the nature of the texts themselves and allow us to make some interesting hypotheses about a possible definition of a 'last letter' genre.

Italiano. Questo paper ha l'obiettivo di trarre alcune analisi preventive in merito alle attività di ricerca sinora svolte sui corpora raccolti all'interno del progetto Marie Curie *Last Letters from the World Wars: Forming Italian Language, Identity and Memory in Texts of Conflict*. Tale progetto vuole approfondire gli aspetti linguistico-tematici delle ultime lettere dei condannati a morte delle due guerre mondiali. In seguito alla creazione di un corpus di lettere di prigionieri di queste due guerre, si sono palesate delle differenze sostanziali nella lingua e nei contenuti di questi documenti. Tali diversità sono legate alla natura degli stessi testi e permettono di porsi alcuni interessanti interrogativi rispetto ad una possibile definizione del genere 'ultima lettera'.

1 Introduction

This paper aims to report some of the first results of the Marie Curie project entitled *Last Letters from the World Wars: Forming Italian Language, Identity and Memory in Texts of Conflict*, which started in September 2018. The project analyses the linguistic and thematic features of the last letters of people sentenced to death during the two World Wars, and is conducted with digital humanities tools. The documents concerning the First World War have been collected mainly in the Central Archives of the Italian State in Rome and thanks to a kind donation by Professor Giovanna Procacci, who offered her letters of Italian prisoners (Procacci, 2000), published and unpublished, for analysis in the Last Letters project. The letters from the Second World War were collected in close collaboration with the Ferruccio Parri National Institute (ex INSMLI) and the Centre for Contemporary Jewish Documentation (CDEC), both in Milan. The majority of the Second World War texts were collected through these two organisations, although some also come from other Italian institutes of resistance connected to the Ferruccio Parri National Institute, which is the central organisation of the Italian Network of Institutes for the History of the Resistance and the Contemporary Age. Other letters were found thanks to the National Association of Ex-Deportees in the Nazi Camps (ANED), again in Milan. In other words, we were able to collect letters from Italian prisoners captured by the Austrian or German armies as far as the First World War is concerned, whereas we composed a corpus of letters from partisans and Jewish deportees for the Second World War. The total number of letters is 1203 for letters from WWII and 960 for WWI. I selected those documents, which were analysed for this paper, from a total of approximately 3500 letters collected in the first six months of my archival research.

2 Objectives

The objective of this paper is to display the main differences between these two corpora (WWI-WWII). In fact, as far as the Second World War is concerned, the prisoners who wrote those letters were mainly partisans who knew for certain that they were going to be executed, whereas for the First World War the writers did not have a precise notion of what their fate was going to be, despite the precariousness of their situation. This dichotomy between letters from prisoners who knew they were sentenced to death and prisoners who still believed in a chance of survival, is underlined by textual and extra-textual elements. These elements were made evident by a digital analysis that allowed a greater understanding of these letters. The purpose of this paper is not to provide a complete interpretation of the subject but to propose a framework for the genre 'last letter' in order to understand if it is possible to give a preliminary answer to such a question. The texts analysed were submitted to a first NLP analysis carried out with TreeTagger (Schmid, 1994, 1995). To conduct this

preliminary analysis, I used, in TreeTagger, the Italian parameter file of Professor Achim Stein (University of Stuttgart). I then scanned the text for potential errors. Several writers for example, especially during World War I, often write a group of words as a single word. An example of this is '*saperesezicarlo*' which means 'to know if uncle Carlo…' The transcripts of these texts normally respect the language and spelling of the original documents. This way of writing, combined with dialectal forms, could not be read by TreeTagger. In order to rectify the pos-tagging procedure, I decided to correct manually the repertories of tagged word so that the work could be more precise, considering that I had to divide those groups of words. Regarding the stop words, I decided to include them in the analysis because they represent another characteristic habit of the writers. As I explained, these texts often display groups of united words, and in some cases, this happens because the writers do not know the correct spelling of these phrases, nor the concept of collocating, for instance, a preposition and a noun. An example of this tendency is the recurring *intrincea* [in the trenches], which should be written in two separate words, or *aggorizia*, which is a case of syntactic gemination (Repetti, 1991), a typical phenomenon that occurs in spoken Italian. In three different cases, the writers wrote *aggorizia* instead of *a Gorizia* [to Gorizia], imitating the sound they reproduce orally. This example is typical of the main category of writers who are part of the two *corpora*, that is, partially literate people. They often write groups of words attached together, like *alacamba*, literally *alla gamba* [to the leg], with the palatalization of the velar consonant (Pellegrini, 1985: 272). Another interesting example is the assimilation of the verb 'have' to the following past participle. In 1,2% of the *passato prossimo* [present perfect] forms it is possible to read expressions like *oreclamato* [I reclaimed] or *oscrito* [I wrote]. Finally, a third kind of word grouping, which occurs quite rarely (0,02% of the tokens), consists in the writer attaching entire phrases in a unique form. Some examples of this tendency are to be found in cases like *nosischersa* [don't joke about it] that display the low level of education of the writer, who obviously does not know the correct spelling of the verb *scherzare*, in which there is no 's' before the final suffix. The preliminary analysis that I am proposing in this paper presents the very first results of my ongoing research, namely, the first comparison between WWI and WWII corpora.

3 Letters

So what is a 'last letter'? Is it the last text written by someone before his or her disappearance or should it have some precise characteristics in terms of language and contents? Can a prisoner who was ignorant of his/her fate really write a last letter? Can we consider a 'last letter' one written by a prisoner who is then pardoned? Traditionally in the history of epistolary memoirs, the last letter has always been vaguely described as the last message that remains to us from someone who died. However, there are several types of documents that fit this description, and yet also have other, distinct features. For example, the Jewish partisan Emanuele Artom (Aosta, 23/06/1915 – Torino, 7/04/1944) kept a diary (Artom, 1966) during his imprisonment by the Nazis. His spiritual testament is contained within this diary, but not at the end of the text. Could this message be considered Artom's last letter even though it is a diary page, simply because it contains the last message he wrote? Considering that the World War II corpus is mainly composed by attested last letters, and the World War I corpus comprises letters written by prisoners who were, in most cases, unaware of their possible execution, I will compare them in order to see what the main differences between these two corpora are. Then I will determine whether, among these differences, there are shared, distinctive features that can generally be attributed to a last letter genre, as people's final messages obviously present some recurring peculiarities. One of the main characteristics of last letters is the request to the family for forgiveness. An example of this, taken from the WWI corpus, is the famous letter by Fabio Filzi (Pisino, 20/11/1884 – Trento, 12/11/1916), an Italian volunteer and irredentist executed by the Austrians. In his last message to his parents, one can immediately recognise a request for forgiveness which is one of the common traits of letters in both WWI and WWII corpora.

> Cari genitori, prima di morire non posso fare a meno di esprimere il mio profondo rincrescimento, per il fatto che mi sovrasta, invero non per la mia esistenza, ma per voi che avete fatto tanto per me e che non approvaste i miei sentimenti italiani. Io ho sempre adempito il mio dovere con scrupolosità seguendo sempre l'impulso della mia coscienza. Prima di morire rivolgo il pensiero a voi e alla mia cara Emma, che si trova a Padova, e contro i cui consigli ho agito arruolandomi. Addio per sempre, baci ai miei fratelli. Fabio Filzi.[1]

[1] Dear parents, before I die I cannot help but express my deep regret, for the fact that it overwhelms me, indeed not for my existence, but for you who have done so much for me and who did not approve of my Italian feelings. I have always fulfilled my duty with scrupulousness, always following the impulse of my conscience. Before dying I turn my thoughts to you and to my dear Emma, who is in Padua, and against whose advice I acted by enlisting. Goodbye forever, kiss my brothers. Fabio Filzi.

This letter stands out as one of the few examples of surviving letters written by Italian WWI soldiers who were executed. The red part shows the request for forgiveness to his parents for the pain that his own death will cause them. This sentence could also be easily contained in the letter of a partisan sentenced to death. On the contrary, the green part is something totally different compared to the letters of the Second World War. In the letters of this period there are no ideological clashes between the people sentenced to death and their families. The contrast is always linked to the desire of the condemned person's loved ones not to lose them. Another feature that differentiates the two corpora is that First World War letters were written only by adult men while in the WWII corpus there are several other categories of writers, like teenagers and women. An example is Renato Mantovani (Treviso, 16/12/1928 – Pieve di Teco (IM), 26/01/1945), one of the youngest partisans in the corpus, who was 16 years old.

> Notizia ai genitori. 'Sono accusato di appartenere alle bande comuniste, vi domando perdono, ora mi fucilano'. Renato.[2]

Again, as you can see in green, Mantovani apologizes to his family for the pain that his shooting will cause them, as in the previous case of Fabio Filzi. These two texts show that there are characteristics in common among the last letters, although these two texts were produced 20 years apart and were written by different profiles of writers. Another common feature of all these letters, even those not written with full awareness of certain death, is to entrust one's family to the care of the recipient of the message. This is because the writer thinks that they will never be able to see their loved ones again. An example from the First World War is the following excerpt from the letter of a soldier who is writing from a trench shortly before an assault, with an obvious fear of dying.

> […] e bacia i banbini ch'io sara dificile a poterli rivedere ancora una volta; mediante il mio scritto la lagrime cadono dalli occhi che una volta ti rimirava. adio. […]. (Procacci, 2000: 414).[3]

We only know the name of this infantryman, Beppe, and we cannot say with certainty that this is his last letter even if it is plausible to think so. In red, it is evident that the letter refers to never seeing one's loved ones again with the probable intention of exorcising the fear that this soldier had that this eventuality would actually occur. An example of a letter from World War II with the same kind of tone is that of Vanda Abenaim (Pisa, 6/05/1907 - Auschwitz, unknown), a Tuscan Jew who did not survive Auschwitz. The case of Abenaim is very interesting because it is one of the few that presents coded messages in a last letter. Based on her son's accounts, we know that the family had devised a coded means of communicating in case they should find themselves in a dangerous situation (Pacifici, 1993: 129).

Firenze 30/11/1943

> Gent.ma Signora, Mi farebbe tanto la gentilezza di consegnare a mio fratello la presente perché purtroppo sono ferita gravemente e non so quale destino mi sono destinata. Sono molto avvilita perché non so se potrò essere salva e rivedere più i miei cari. Già sono in camerata. Pregata tanto per me. I bimbi sono stati salvati. Per ora sono sempre a Firenze. Mando tanti baci al mio caro Carlo e mando baci alla mia mamma e chissà quando la rivedrò. Saluto tanto anche lei e pure la sua signorina. Sua aff.m a nipote Vanda.[4]

The letter is theoretically addressed to a woman, but in truth it is addressed to her brother. In red you can see the parts where the woman asks her brother in code to take care of the children because she fears she will never see them again. The Abenaim family had established a secret code if they were captured by the Germans (Abenaim, 2015). The examples in the text are in green. With the sentence *sono gravemente ferita e non so quale destino mi sono destinata* [I am seriously injured and do not know what fate I am destined for], she warns her family that she has been taken prisoner by the Germans. Moreover, with the expression *Già sono in camerata* [I'm already in my dormitory], she informs her brother that she is already on the train to the concentration camp. Considering that these texts have some common traits despite their many differences, I tried to keep the characteristic elements of the original documents to better underline the differences between

[2] Notify the parents. I am accused of belonging to communist gangs, I ask your forgiveness, now they shoot me 'Renato'.
[3] And kiss the children as it will be difficult for me to see them again; through my writing the tears fall from the eyes that once gazed at you. Farewell.
[4] Dear Madam, could you be so kind as to bring this letter over to my brother because unfortunately I am seriously injured and do not know what fate I am destined for. I am very discouraged because I do not know if I will be saved and I will be able to see my loved ones again. I'm already in my dormitory. Please pray a lot for me. The children have been saved. For now I'm still in Florence. I send many kisses to my dear Carlo and I send kisses to my mother and who knows when I will see her again. I also greet you and your lady. Your affectionate nephew Vanda.

one corpus and the other, through a semantic and morpho-syntactic analysis of the two. I mainly used TreeTagger and I displayed some results through the use of the Links tool of Voyant-Tools. The next step of the project will consist in the comparative analysis of another group of texts carried out with TreeTagger and other lemmatizers such as Tint (tint.fbk.eu), UDPipe (http://ufal.mff.cuni.cz/udpipe) and T2K (http://www.italianlp.it/demo/t2k-text-to-knowledge/).

4 Part-of-Speech analysis

It is the first time that such an analysis is applied to these texts, despite the fact that in some cases these letters have already been studied and analysed, both for the WWI corpus (Spitzer, 2016) as well as the WWII one (Bozzola, 2013), but in any case, this is the very first *digital* analysis that has been applied to these texts. The possibilities given by digital tools have enabled us to clearly see the differences between World War I and World War II, and to establish some of the characteristics of the last letter genre. The result of the pos-tagging and the following manual corrections enabled me to gain a better understanding of the language of these letters. The following table summarises all the characteristics of these texts.

	World War I	World War II
Letters	960	1203
Tokens	63.637	134.103
Types	9.953	12.912
Lemmas	6122	8031
Type-Token Ratio (TTR)	0,156	0,096
Lemma-Token Ratio (LTR)	0,96	0,6
Average Words Per Sentence	19,2	32,3
Accuracy	89,8%	95,5%
Corrected Tokens	6505 (10,2% of the total)	6044 (4,5% of the total)

The data concerning the accuracy are very interesting, and are linked to the linguistic nature of the two corpora. As a matter of fact, the World War I corpus is linguistically more problematic, and Treetagger had a harder time analysing it. As you can see, on World War I texts, it had an index of accuracy of 89,8%, which is 5.7 percentage points less than the accuracy score it had on World War II letters. A manual correction confirmed these data, but I also noticed, thanks to a close reading approach to World War I letters, some linguistic peculiarities that I did not think were canonical peculiarities, such as local and dialectal traits. As I explained in the introduction, World War I writers were not confident about the spelling of Italian, owing to their education and their being dialect speakers. These characteristics, which TreeTagger cannot tag, are extremely representative of the World War I corpus, but totally absent from the other one. Moving from these general comments about parts of speech to a more detailed analysis, it is possible to verify the differences between the two corpora in terms of what grammatical categories are used, and where TreeTagger struggled the most.

World War I	Corrected		TreeTagger		Accuracy
	Occurrences	Percentage	Percentage	Occurrences	
Abbreviations	308	0,48%	0,07%	45	0,413%
Adjectives	4843	7,61%	8%	5093	0,392%
Adverbs	5842	9,18%	8,65%	5504	0,531%
Conjunctions	4268	6,7%	6,5%	4141	0,199%
Articles	4283	6,73%	7,84%	4988	0,568%
Nouns	12038	18,91%	21,95%	13970	3,03%
Proper Names	1550	2,43%	1,91%	1217	0,523%
Punctuation	5263	8,27%	1,99%	1269	6,276%
Prepositions	9674	15,20%	13,12%	8349	2,082%
Pronouns	9078	14,26%	13,11%	8347	1,148%
Verbs	15890	24,97%	20,10%	12792	4,868%

Accuracy	TreeTagger		Corrected		**World War II**
	Occurrences	Percentage	Percentage	Occurrences	
0,055%	84	0,063%	0,12%	159	**Abbreviations**
0,118%	10809	8,06%	8,18%	10968	**Adjectives**

0,428%	11213	8,36%	8,79%	11787	**Adverbs**
0,051%	8629	6,43%	6,49%	8698	**Conjunctions**
0,040%	8806	6,57%	6,6%	8860	**Articles**
2,427%	26940	20,09%	17,66%	23685	**Nouns**
1,191%	3237	2,41%	3,61%	4835	**Proper Names**
0	3515	2,62%	2,62%	3515	**Punctuation**
0,255%	16990	12,67%	12,41%	16647	**Prepositions**
0,548%	19935	14,86%	15,41%	20670	**Pronouns**
9,994%	27154	20,25	30,19%	40485	**Verbs**

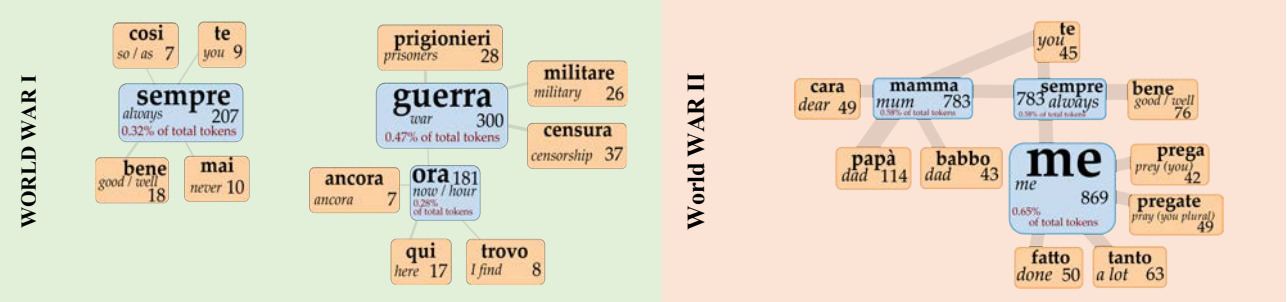
The biggest differences between the two *corpora* lie in the use of verbs. If the present, the most common tense of the indicative, appears in both corpora with almost the same frequency (WWI: 7,82 – WWII: 8%), the same cannot be said of the past tenses. The use of the past – simple past, imperfect, and present perfect – in World War II letters is more than twice as frequent (11,41%) as in WWI texts (4,9%). Combining these data with a close reading approach, it is possible to affirm that this linguistic trait is one of the peculiarities of the last letter genre. As I said, WWI letters were written by prisoners who wanted to tell their families about their everyday lives. On the other hand, WWII letters were written by people sentenced to death, who often used the memories of their past experiences as a way of exorcising the fear of capital punishment and entrusting their loved ones with the memories of happy times when they were together. The use of parts of speech being so meaningful, I decided to highlight the correlation between these grammatical elements by using a graphic tool. Then I submitted these tagged texts to Voyant-tool. The application then showed which parts of speech are the most common (blue rectangles) and which combinations they form (orange rectangles).

It is immediately evident that the texts from the First World War give greater attention to textual construction. For instance, they display a greater use of punctuation. As Spitzer (2016: 108) noticed, this use is often made incorrectly (Cortelazzo, 1972: 119-123). Nevertheless, the writers demonstrate awareness of the fact that punctuation must be there as an indispensable element of the text, (Restivo: 2018, 249), while in the Second World War letters, due to the strong emotionality of the moment, language becomes mimetic of speech. First World War morpho-syntactic chains highlight a higher number of nouns, prepositions and adverbs with an indirect relationship between substantives and adjectives. Analysing the letters with a close reading approach, I can suppose that sentences are more complex and present a higher number of indirect objects. On the other hand, the Second World War corpus, with a higher number of nouns and adjectives, displays a more prominent use of direct objects or nominal sentences.[5] Indeed texts of the First World War show a greater hypotaxis, therefore complexity, in comparison to those of the Second, precisely because of the different emotional conditions of writing but not only. In fact, the writing of the last letters of deportees and condemned to death of the resistance is very often clandestine and can literally be visualised on the page as a stream of consciousness, as the writers were trying to make the most of every moment in which they could find time to write. In a possible definition of the last letter, it will therefore be necessary to consider the morphological construction of the discourse as one of the discriminating factors in the reflection on genre.

5 Most Frequent/ Characteristic words

In order to better display the use of the lexicon, I used Voyant as a tool to visualize collocations and links within the texts. The representation of the connections of the most frequent words (blue rectangles) with the others (orange rectangles) is the following:

[5] In any case, in the next phases of the project I will go further in the syntactic analysis using tool like Coh-Metrix (http://terence.fbk.eu/services/api/computeReadability/v2/).



On the green background is the corpus of the First World War while on the red one is the corpus of the Second.[6] This analysis is surprising because it clearly shows what the nature of the two corpora is. The words *sempre*, *ora* and *guerra*, on the left, show a descriptive lexical approach to war writing, which aims to tell the stories of the front to the families of the writers. On the contrary, the most frequent words in the last letters corpus of the Second World War illustrate a familiar lexicon that reveals the emotional character of this writing. All this information allows us to say more about the generic features of the last letter. In fact, both corpora, WWI and WWII, are texts written in a tragic moment and, if we bracket the substantial differences between a person condemned to death and a soldier in prison, the condition of deportees of the Second World War can in fact be compared to that of prisoners of the First World War. The substantial difference lies in the codification that the subject makes of the reality he or she is living. A soldier learns to experience the daily realities of war as part of a group of like-minded people; imprisonment and death become a codified consequence of a tragic but commonly accepted situation. Those sentenced to death, on the other hand, find themselves alone before death, in some cases feeling incredulity, such as the case of fourteen or fifteen-year-old boys who do not expect to be shot or tortured; in other instances, they cling onto the hope that their comrades can make an exchange for a Fascist or a Nazi prisoner. On the other hand, deportees, especially for racial reasons, are faced with the unknown while in chains, considering that their imprisonment is not the result of their actions but of their personal identity, and they are kept in the dark as to their fate. The letters of many deportees also contain appeals to hope. A very interesting fact is that this concerns the lexicon in its entirety. If we observe in fact the lemma/token ratio (Jurafsky, Bell, Girand, 2002) it is evident that the WWI corpus is more lexically varied (0.96) than the WWII corpus (0,6). This is due to the fact that the letters written a few hours before execution with the certainty of having to die, are characterized by a basic lexicon that often returns. I preferred lemma-token to type-token because it is better suited to treat inflected forms of a word as the same type (McCarthy, 1990: 73). To give an example of how the lexicon of these texts works, we can for example cite the use of the word *dolore* [pain] which is often used in phrases in which the writer apologizes for the pain that death will cause his/her loved ones as in the previous case of Filzi and Mantovani. In the corpus of the First World War there are 43 occurrences of the word *dolore* and in no case is it collocated with adjectives. By contrast, in the corpus of the Second World War, the word 'pain' appears 185 times, 57 of which are accompanied by a demonstrative adjective such as: *grande* [big], *immenso* or *immane* [immense], *tremendo* [tremendous], *profondo* [deep], *accorato* [heartfelt], *straziante* [heartbreaking] or *ultimo* [last]. The language in these last letters is therefore more descriptive, especially when it describes the feelings and therefore distinguishes the story of imprisonment or of life in the trenches from an inner narrative that must condense a greater communicative intent into a few lines. It should also be noted that of the 57 occurrences of the noun *pain* with these adjectives, 7 have the adjective post-placed to the noun while 50 have it placed before the noun (Serianni, 1989: 199-205). This is typical of the syntactic structures commonly found in literary texts (Scarano, 2000: 5). It is no wonder that there should be a similar lexicon as well as sentence construction in the letters, given that their authors learnt how to write in Italian through the example of literature, for instance Dante. These letters are as diverse as the materials on which they were written. During my research in the archives, I never found a single letter from a soldier on the front of the First World War that was written on a precarious medium. In contrast, the partisans and deportees wrote their last messages really wherever they could (Bozzola, 2013: 26). There are also, for instance, 'letters' composed of three words on the edge of a book or even a list of names engraved on a loaf of dry bread. These letters represent in essence what the two wars were, and testify to their differences. Thanks to the function 'oppose' of the R package Stylo, I identified the most characteristic words of each of the *corpora*. For World War I it found *austriaci* [Austrians] (34), *macello* [slaughterhouse] (21),

[6] In the rectangles the numbers are the occurrences of the given word. An English translation is provided in italics. The blue rectangles contain the most frequent words, or parts of speech, and the orange the ones those that appear most frequently in connection with them.

licenza [license] (19), *francesi* [French] (19), *stanchi* [tired] (18). On the other hand, for World War II, *muoio* [I die] (212), *sii* [be] (117), *chiedo* [ask] (105), *perdonatemi* [forgive me] (51), *non piangete* [don't cry] (45), *ricordatevi* [you remember] (43). It is interesting to note how, even in these few cases, the most characteristic words of World War I are related to the conflict, describing it as a slaughterhouse, its protagonists – the Austrian enemies and the French allies – the authors' desires to escape the war while on leave and to one of the most common feelings of the soldiers: tiredness.

6 Conclusion

These 'last' letters focus on the content of the message, whereas World War II letters are most concentrated on the emotive and conative functions because the language focuses on the sender and the addressee (Jakobson, 1960). The sentenced to death ask to be remembered by the people they love. They want to be forgiven and for their families to be happy. We could therefore assume that one of the most salient peculiarities of a last letter is when the message mainly focuses on the sender himself/herself and on the addressee. The last letters aim to describe emotions rather than facts, and to tell about the past more than about the present, because there is no future for the sentenced to death. In the next future, I will include other letters in the corpus and I will cross the analysis done until now with TreeTagger with the use of other lemmatizers and tools. The analysis I have conducted revealed some problems in comparing World War I and World War II letters but it also highlighted changes in the writing of Italian. Most importantly, this phase of my research proved that a 'last letter' was thematically, linguistically and pragmatically definable.

Acknowledgements

I would like to thank Dr Rachele Sprugnoli for having shared with me her scientific expertise, and Professor Matthew Reynolds and Dr Anita Jorge for their precious help, advice and patience.

References

Antonietta Scarano. 1999. Storia grammaticale dell'aggettivo. Da sottoclasse di parole a parte del discorso, *Studi di grammatica italiana*, 18: 57-90. IT

Daniel Jurafsky, Alan Bell, and Cynthia Girand. 2002.The Role of the Lemma in Form Variation, *Laboratory of Phonology*, 7: 3-34, UK.

Emanuele Artom. 1966. *Diari. gennaio 1940 - febbraio 1944*, Milano, Centro di Documentazione Ebraica Contemporanea, IT.

Emanuele Pacifici. 1993. Non ti voltare: autobiografia di un ebreo, Firenze, Casa Editrice Giuntina, IT.

Giovan Battista Pellegrini. 1985. Appunti sulla 'Romania continua': la palatalizzazione di CA, in R. Ambrosini (ed.), *Tra linguistica storica e linguistica generale. Scritti in onore di Tristano Bolelli*, Pisa, Pacini: 257-273.

Giovanna Procacci. 2000. *Soldati e prigionieri italiani nella Grande guerra: con una raccolta di lettere inedite*, Torino, Bollati Boringhieri, IT.

Hellmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, vol. 12: 44–49, UK.

Hellmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop*, UK.

Leo Spitzer. 2016 [1976[1], 2014[2]]. *Lettere di prigionieri di guerra italiani. 1915- 1918*, Milano, Il Saggiatore, IT.

Luca Serianni. 1989. *Grammatica italiana*, UTET, Torino, IT.

Luca Serianni. 2010. *Prima lezione di grammatica*, Laterza, Torino, IT.

Manlio Cortelazzo. 1972. Avviamento critico allo studio della dialettologia italiana. vol. III: Lineamenti di italiano popolare, Pisa, Pacini, IT.

Maria Laura Restivo. 2018. La punteggiatura nelle scritture di italiani semicolti: le 'lettere' di Leo Spitzer, *Italiano LinguaDue*, vol. 10 n. 1: 233-250, IT.

Michael McCarthy. 1990, *Vocabulary*. Oxford, Oxford University Press, UK.

Repetti, Lori. 1991. A moraic analysis of raddoppiamento fonosintattico, *Rivista di linguistica* 3: 307-330, IT.

Roman Jakobson. 1960. Linguistics and Poetics, in T. Sebeok (ed.) *Style in Language*, Cambridge, M.I.T. Press, 1960, pp. 350-377, UK.

Sergio Bozzola. 2013. *Tra un'ora la nostra sorte*. Carocci, Roma, IT.
Umberto Abenaim. 2015, Abenaim: una famiglia ebrea e le leggi razziali, Scritture Edizioni, Piacenza, IT.

Archival Collocations

Emanuele Artom's text:
Archivio Fondazione CDEC, Fondo Emanuele Artom, b. 1, fasc. 10. [link: http://digital-library.cdec.it/cdec-web/storico/detail/IT-CDEC-ST0002-000014/artom-emanuele.html]

Fabio Filzi's letter to his parents:
Archivio Fondazione Museo Storico del Trentino, b. 11 /66. [link: https://www.cultura.trentino.it/Fotografia-Storica/ULTIMA-LETTERA-DI-FABIO-FILZI-AI-GENITORI].

Vanda Abenaim's last letter:
Archivio Fondazione CDEC, Fondo Vicissitudini dei singoli, s. I, b. 1, cc. 14. [link: http://digital-library.cdec.it/cdec-web/storico/detail/IT-CDEC-ST0005-000095/abenaim-wanda.html].

Renato Mantovani's last letter:
Archivio Istituto storico della Resistenza e dell'età contemporanea, Sezione I, cartella 77. [link: http://www.ultimelettere.it/?page_id=52&ricerca=514&doc=766].

L'organizzazione e la descrizione di un fondo nativo digitale: PAD e l'archivio Franco Buffoni

Paul Gabriele Weston **Primo Baldini** **Laura Pusterla**
Università degli studi di Pavia
{name.surname}@unipv.it

Abstract

English: Within PAD-Pavia Archivi Digitali, a project aimed at the medium and long-term preservation of born digital archives belonging to Italian writers and humanists, several procedures based on Franco Buffoni's files have been tested in order to provide a better and more sustainable description of an archive of such nature. Through IT capabilities, PAD has sought to provide the end user with the greatest possible number of access points to the resource. PAD has also experimented implementing by computer-assisted treatment, based on comparison algorithms, the description carried out manually. This test took place on textual material from the author's website which had been previously saved.

Italiano: PAD-Pavia Archivi Digitali, progetto volto alla conservazione a medio e lungo termine di archivi d'autore, ha sperimentato le procedure occorrenti alla descrizione di un archivio nativo digitale sul Fondo Franco Buffoni (scrittore e poeta, traduttore, anglista). Attraverso le potenzialità informatiche, PAD ha cercato di fornire all'utente finale la maggior quantità possibile di punti di accesso alla risorsa. Dopo averne effettuato la descrizione in modo tradizionale, PAD ha sperimentato una forma di descrizione assistita dal computer, basata su un algoritmo di comparazione. Questa prova è avvenuta sul materiale salvato dal sito web dello stesso autore.

1 Introduzione

Il progetto PAD-Pavia Archivi Digitali dell'Università di Pavia è nato nel 2009 con lo scopo di preservare dalla scomparsa gli archivi delle memorie digitali di autori contemporanei. L'Università, che attraverso il Centro per la tradizione manoscritta di autori moderni e contemporanei dal 1969 salvaguarda i documenti cartacei di scrittori e giornalisti italiani, ha voluto estendere questa esperienza ai documenti nativi digitali. Fu il giornalista e scrittore Beppe Severgnini, ex alunno dell'Università, che, partendo dalla constatazione che una parte maggioritaria della produzione culturale letteraria si basa ormai sull'utilizzo di supporti informatici, sollecitò questo ampliamento di prospettive. Per rendere tali documenti ricercabili e leggibili è necessario servirsi di infrastrutture hardware e software in continua evoluzione, un ostacolo che rende le procedure della conservazione progressivamente più impegnative con il passare degli anni. La volontà di arginare questa perdita di testimonianze della nostra storia culturale è stata alla base della creazione del progetto PAD.
PAD conserva diverse tipologie di materiali digitali, garantisce la tutela a lungo termine dei fondi ed eventualmente può essere accessibile agli studiosi, nel pieno rispetto, come è ovvio, delle disposizioni ricevute dagli autori.

Fino ad ora il progetto si è focalizzato soltanto sulla conservazione a lungo termine degli archivi digitali in locale, ospitati cioè sui dispositivi di scrittura correntemente utilizzati dagli scrittori o su apparecchiature non più utilizzate, ma da essi conservate, nonché sui supporti di archiviazione utilizzati dagli stessi nel corso degli anni (nastri magnetici, floppy di diverse dimensioni e densità di archiviazione, cd, dvd, unità compatte di archiviazione massiva). La crescente tendenza ad avvalersi della rete per comunicare ed archiviare dati ha reso necessario mettere a punto una strategia e dei dispositivi finalizzati alla salvaguardia di risorse digitali, siti web e contenuti sui social media. A questo modulo del sistema è stato dato nome PAD Web Archiving. L'intento di PAD non è, ovviamente, quello di competere con analoghi progetti internazionali di ben altro respiro, ma mantenersi come un progetto sostenibile, sia tecnologicamente, sia finanziariamente, che garantisca, ad onta delle sue dimensioni contenute, dei risultati di qualità. Spetta agli autori stessi o alle istituzioni culturali alle quali fanno capo i siti interessati richiedere espressamente che anche questa componente venga inserita nel piano complessivo di preservazione dell'archivio. L'accordo è indispensabile al fine di interagire direttamente con il committente per stabilire tempi e metodi per il salvataggio e la consultazione. Tutto il materiale resta ovviamente di proprietà dell'autore, che può in ogni momento decidere di rimuoverli dall'archivio e di rinunciare al prosieguo del progetto.

2 Il fondo Franco Buffoni

Franco Buffoni, anglista, poeta, prosatore e traduttore, il cui archivio cartaceo si trova già in deposito presso il Centro Manoscritti della stessa Università, avendolo lui conferito in anni precedenti, è uno degli autori che, nel corso degli anni, hanno conferito a PAD i propri archivi digitali. Nel 2016, nel rispetto dell'iter messo a punto allo scopo da PAD, una copia del suo ampio archivio è stata riversata in PAD. L'iter a cui si fa qui riferimento prevede che, dopo la firma di un contratto legale, un operatore di PAD si rechi presso la residenza dell'autore per prelevare una copia dei file che lui stesso ha selezionato per la conservazione. Il riversamento dell'archivio Buffoni ha riguardato 1065 elementi, per complessivi 758 MB, comprendenti tipologie di file di diversa natura: documenti di testo, immagini, video, audio, link. Nel 2019, con l'intenzione di sperimentare lo strumento messo a punto per la salvaguardia a lungo termine delle risorse web, PAD ha concordato con l'autore di utilizzare il suo sito personale (www.francobuffoni.it), ritenendolo particolarmente idoneo allo scopo a motivo della ricchezza di contenuti e della varietà di formati e tipologie.

Per prima cosa, su richiesta esplicita del proprietario del sito, PAD ne ha prelevato una copia. Dato che i siti web possono essere modificati o aggiornati anche molto di frequente, si è concordato con l'autore di procedere con l'effettuazione di salvataggi a cadenza prestabilita. In questo modo si possono conservare le varie versioni del sito, che possono essere messe a disposizione dell'utenza secondo la volontà del proprietario. Attraverso un software per il web scraping, il sito dell'autore è stato riprodotto in locale, in modo da garantirne il browsing offline. Così l'utente futuro potrà navigare liberamente nella copia dell'intero sito. Per progettare questa implementazione, si è dovuto tenere conto della struttura anche molto complessa che i siti possono talvolta presentare, comprendente riferimenti numerosi ad altre pagine, interne o esterne nel web. Per questo di ogni pagina che compone il sito web, PAD memorizza, oltre alla pagina stessa, i link anche alle pagine esterne, con un'immagine della pagina a cui il link conduce, nonché i documenti allegati. In questo modo si può tenere meglio traccia dei path che il creatore del sito ha voluto valorizzare. Se, ad esempio, un link a una pagina esterna non fosse più funzionante o se la pagina non risultasse più esistente, una parte di ciò che l'autore intendeva comunicare, una componente probabilmente significativa del suo pensiero, andrebbe perduta.

Quando il progetto ha avuto inizio è stata presa in considerazione l'idea di utilizzare principalmente il formato di archiviazione WARC (Web ARChive). Sebbene questo formato sia stato standardizzato nel 2009 (ISO 28500:2017) il suo utilizzo da parte delle grandi aziende informatiche (Microsoft, Apple, Google ecc.) non ha mai goduto negli anni della diffusione che sarebbe stata auspicabile. Un'accurata serie di verifiche ha permesso di accertare che i browser più diffusi non lo riconoscono. Allo stesso tempo il sistema di memorizzazione sembra essere stato realizzato per essere installato e usato solamente da un sistemista esperto, ciò che rischia di creare notevoli problemi agli utenti comuni. Si è preferito quindi adottare un prodotto di più facile utilizzo per l'elaborazione del sito. Il formato WARC è, invece, stato mantenuto per la parte del progetto che si occupa di preservazione a lungo termine. Quindi in PAD per il Web Archiving vengono gestiti due sistemi diversi. Il primo utilizza il software Heritrix, sviluppato da Internet Archive, mentre il risultato dei processi di crawling viene memorizzato in file con formato WARC.

Per l'elaborazione delle pagine il software che PAD utilizza prevalentemente si chiama HTTrack, un Web crawler open-source. Consente di scaricare un sito web da internet in una directory locale, ottenendo HTML, immagini e altri file dal server al computer. HTTrack mantiene la struttura originale del sito, compresi i link, permettendo all'utente di navigare da una pagina all'altra, come se la stesse visualizzando online. Esso preleva anche tutte le altre tipologie di documenti e immagini che si trovano allegate alle pagine web. Pur non basandosi su uno standard, HTTrack presenta il vantaggio della semplicità nell'aprire le pagine web o nell'estrarre il testo per elaborarlo. Alcuni autori, come ad esempio Francesco Pecoraro, hanno richiesto di creare una copia offline del proprio sito, con l'intenzione di rimuoverlo successivamente dalla rete. Questa copia resta ovviamente a loro disposizione per l'accesso, qualora ne facciano richiesta, anche a distanza. Si è visto come la versione 'mirror' del sito è stata considerata come quella più semplice da inviare e da consultare da parte di utenti non particolarmente esperti. La possibilità di navigare all'interno del sito locale, senza l'obbligo di installare preventivamente specifici software, ha reso questo servizio estremamente semplice da gestire. Al contrario, per le questioni ricordate in precedenza, l'utilizzo di un archivio WARC avrebbe comportato la necessità di assistere l'utente nel corso della procedura di consultazione.

Tutto il materiale così raccolto è stato sottoposto alle procedure di conservazione sperimentate da PAD nel corso degli anni. La prima operazione è creare più copie dell'archivio, ubicate su diversi server. Oltre che sul server interno di PAD infatti, esso viene replicato sui server dell'Università di Pavia e su quello della sede

distaccata dell'Università a Cremona, città distante da Pavia circa 70 chilometri, in modo da salvaguardare la sicurezza delle informazioni in caso di disastro ambientale. Un'ulteriore copia viene memorizzata su supporto hardware esterno. Una volta assicurata la preservazione dell'originale, l'archivio passa in un'area di working.

Vengono estratti i metadati, fondamentali per poter poi svolgere l'operazione di normalizzazione, che comporta il salvataggio di ogni documento in diversi formati, a seconda della tipologia. Terminate le operazioni preliminari, si procede a descrivere i file.

3 La descrizione

Rispetto al trattamento di un archivio cartaceo, un archivio nativo digitale presenta peculiarità, come ad esempio il numero dei file che lo costituiscono, che talvolta possono assommare a molte migliaia, che rendono la descrizione effettuata seguendo consuetudini e procedure tradizionali inadeguata e persino non sostenibile. È stato, perciò, necessario individuare strategie che potessero consentire di sfruttare al massimo le potenzialità offerte dall'informatica.

Al contempo, si registrano problematiche simili, come la questione dell'accessibilità e della riservatezza. Trattandosi di documentazione prodotta molto di recente, si è dovuto tener conto del fatto che, probabilmente, una parte anche significativa dei file non possano essere resi disponibili all'utenza senza che ciò comporti la violazione di disposizioni legislative, come quelle sulla privacy o sulla proprietà intellettuale. Anche il fatto che i documenti conferiti o salvati dal web siano stati indicati espressamente dal conferente, non è sufficiente a garantirne la libera consultazione da parte degli studiosi. Si rende perciò necessario, durante le fasi preliminari della descrizione, sottoporre ogni singolo file ad una attenta disamina volta ad escludere che contenga dati sensibili o creazioni intellettuali la cui responsabilità non sia in capo al conferente, al soggetto produttore dell'archivio o al titolare del sito. Anche lo scrittore che ha conferito il proprio archivio potrebbe richiedere, per ragioni personali, che alcuni documenti siano secretati e di conseguenza esclusi dalla consultazione, anche per motivi di studio, per un determinato periodo di tempo, il cosiddetto embargo. Per tener conto di queste evenienze, l'operatore di PAD, nel vagliare ogni file dell'archivio digitale, gli assegna una categoria di rischio, in base alla quale esso viene automaticamente reso o meno consultabile da parte degli utenti.

Si passa poi alla fase del riordino. Come per gli archivi cartacei, viene creata una struttura ad albero rovesciato che comprende le diverse serie, alle quali vengono poi assegnati i singoli file. Già in questa procedura si manifestano quelle potenzialità dell'informatica, prima ricordate, che offrono un significativo contributo agli archivisti e nuove opportunità agli utenti. In primo luogo, il sistema offre la possibilità di assegnare un singolo file a più di una sezione dell'archivio. In secondo luogo, se nell'archivio cartaceo il riordino comporta la modifica della sistemazione pensata dallo scrittore, l'archivio digitale può consentire di mantenere ad un tempo l'aspetto originale e contemporaneamente collocare i documenti in un ordinamento che segua altri criteri. Questi criteri possono anche essere più di uno, quando le esigenze lo richiedano. Durante la descrizione, infatti, l'archivista assegna a ogni documento dei tag, che hanno lo scopo di aiutare l'utente a comprendere meglio la tipologia del materiale in questione (se, ad esempio, si tratta di un testo, di una recensione, di un'immagine, di un video e così via). In ogni archivio, ovviamente, non tutti i file sono prodotti da colui o colei che conferisce l'archivio stesso. Sono molto frequenti i casi di documenti frutto del lavoro intellettuale di terze persone. La possibilità di collegare a ogni file uno o più nomi di persona che abbiano in qualche modo contribuito alla sua produzione è funzionale anche a questo scopo. Al tempo stesso, il nome della persona o dell'ente viene associato ad una tipologia di responsabilità intellettuale, espressa attraverso un vocabolario controllato, implementabile a seconda delle esigenze mediante l'inserimento di nuovi termini in una tabella. Ricorre, poi, il caso di documenti - il termine viene qui utilizzato in senso generale, senza cioè fare riferimento a funzioni di natura amministrativa - che siano stati estratti o ricavati da pubblicazioni più ampie (ad esempio, un capitolo da un libro, una poesia da una raccolta, un brano da un'intervista o da una recensione e così via). Qualora il collegamento tra i due documenti sia riconosciuto dall'archivista, il sistema consente di esplicitare la relazione anche a beneficio dell'utente, dal momento che è possibile stipulare un collegamento tra l'oggetto e il titolo della risorsa che lo contiene o del quale è parte. L'inserimento del codice ISBN o DOI nel caso di un libro o dell'ISSN per una rivista è funzionale a rendere più riconoscibile e in modo inequivoco la fonte e, di conseguenza, a consentirne più facilmente il recupero.

La presenza di tutte queste informazioni inserite dall'archivista (tag, nomi, responsabilità, identificativi univoci), unitamente ai metadati tecnici estratti direttamente dalla macchina, consente a PAD di mettere a disposizione dell'utente percorsi di ricerca alternativi, o, per meglio dire, complementari e trasversali, rispetto a quelli consueti. Infatti, vi è la possibilità di estrarre i dati dei documenti secondo l'ordinamento per serie

archivistiche, oppure secondo la disposizione originale dello scrittore, o ancora ordinati per soggetto produttore o per provenienza. In questo modo PAD cerca di venire incontro alle variegate necessità dell'utente che si trova a consultare l'archivio.

4 La descrizione del sito web

Analizzando i conferimenti effettuati dai diversi autori ci si è resi conto che spesso l'autore ha conservato i testi che poi vengono riversati sul web sotto forma di file allegati alla pagina o come link. In altri casi, i testi della pagina web si ispirano, prendono spunto oppure sono almeno in parte i medesimi di quelli presenti nell'archivio. Operando sui metadati estratti, avendo selezionato quelli più importanti per identificare in modo puntuale la risorsa web, tali correlazioni tra nativo digitale e web possono venire ricercate ed individuate.
Le procedure occorrenti non si discostano molto da quelle normalmente messe in atto in PAD al momento dei conferimenti per analizzare la struttura e la consistenza dell'archivio. Come primo passo, si procede a creare una mappa del sito; in secondo luogo si estraggono i metadati, il cui numero viene ridotto in seguito alla scrematura di quelli di scarsa utilità; infine, i documenti vengono sottoposti ad operazioni di normalizzazione. Conclusa questa fase, i documenti del web possono essere descritti.

Il sistema di descrizione è in parte automatizzato, dato che la grande quantità di materiale disponibile sul web richiederebbe tempi eccessivamente lunghi e un'attività dispendiosa, se ad occuparsene fosse un operatore. È evidente, infatti, che, una volta ricevuto il comando, il computer possa elaborare per diverse ore il materiale, fino ad arrivare alla conclusione. Per ottenere un risultato molto simile, un operatore dovrebbe lavorare per giorni. Il software PAD Web Analyzer confronta ogni pagina web - prendendo in considerazione sia il testo vero e proprio della pagina, sia eventuali file allegati - con i documenti presenti nell'archivio. Procede, quindi, a mostrare, affiancati, i testi del web e i corrispondenti testi del nativo digitale, sulla base di indicatori di similitudine. Per ciascuna corrispondenza viene stabilito un indice di similitudine, che indica in percentuale quanta parte del testo sul web sia uguale a quella di un documento presente nell'archivio nativo digitale. Un risultato molto basso, indicativamente inferiore al 50 %, non viene tenuto in considerazione dall'operatore. Al contrario, come hanno permesso di accertare le prove effettuate, la certezza di aver individuato il medesimo documento si ha quando l'indice presenta un valore intorno al 98%. In presenza di valori intermedi sarà compito dell'operatore effettuare i riscontri necessari, ma anche in questo caso il sistema è in grado di offrire assistenza.

Se il valore è relativamente elevato, il computer, assumendo di aver individuato due file "probabilmente" corrispondenti propone l'assegnazione ad una delle serie inserite con la descrizione manuale dell'archivio. Se, viceversa, il valore è relativamente baso, la casella delle assegnazioni resta vuota e deve essere quindi l'operatore a riempirla sulla base di una ricognizione autoptica. Per effettuare queste valutazioni si è tenuto conto del fatto che talvolta i documenti in archivio possono essere in formati differenti rispetto a quelli pubblicati o allegati sul web (ad esempio un file Word solitamente viene convertito in PDF per essere messo sul sito) e questo comporta un ragionevole abbassamento dell'indice. Poiché la macchina non ha le stesse capacità di discernimento di un operatore, è opportuno che il controllo finale sia effettuato esaminando in parallelo i due testi. Una specifica funzione del software di PAD mostra i due testi affiancati in modo che la ricognizione possa procedere speditamente. È ovviamente è possibile che il testo sia stato prodotto direttamente sulla rete, oppure che il file originale sia stato eliminato o non conferito: in tal caso non si evidenziano corrispondenze.

A questo scopo PAD ha implementato all'interno del software l'algoritmo Levenshtein distance, adattandolo alle proprie esigenze. La procedura appena descritta risulta attualmente vantaggiosa solo se la parte dell'archivio digitale nativo sia stata già trattata e viene quindi utilizzata in questo momento unicamente per la descrizione dei siti web o in caso di secondi (o comunque successivi) conferimenti.

Le informazioni ottenute attraverso questo procedimento serviranno all'archivista per assegnare il documento a una particolare serie inventariale, riguardante una specifica opera dell'autore o una specifica tipologia documentale. Integrando i documenti provenienti dai siti web con quelli nativi digitali si ottiene un archivio il più completo possibile. Se all'interno di una serie viene inserito del materiale proveniente dal web, viene creata una sottoserie apposita, in modo che l'utente possa trovare aggregato tutto il materiale avente lo stesso argomento e al contempo conoscerne la provenienza.

5 Prospettive per il futuro

Come nella tradizione delle realizzazioni informatiche, il sistema PAD viene costantemente implementato per fornire nuove funzionalità e migliorare quelle attuali. La scelta di configurarsi come un progetto di piccola

scala, limitato a autori o a istituti culturali selezionati, permette di dedicare grande cura nel perfezionare le soluzioni tecniche, secondo le necessità dell'archiviazione e della descrizione. L'architettura di PAD è stata pensata per la conservazione, ma negli anni si è evoluta con la finalità di consentire lo studio del materiale conferito. La stretta collaborazione con gli scrittori conferenti garantisce il rispetto delle loro decisioni sul trattamento e la gestione del loro archivio e permette di andare incontro alle esigenze che essi manifestano. Seguendo le tendenze dei cambiamenti sociali nell'uso di internet e delle sue risorse, PAD sta sperimentando una funzionalità che permette la conservazione a lungo termine e la consultazione delle pagine personali di social network, come Facebook, Twitter o Instagram, dei canali YouTube e delle e-mail. Questa tipologia di documenti informatici, che potrebbero anche essere di rilevante importanza per studi futuri, sono per ora solo pensati in funzione della conservazione. Su richiesta esplicita di un autore, verranno raccolti i dati direttamente dai social e preservati. Con lo stesso criterio verrebbero trattate le e-mail, utilizzando criteri analoghi a quelli che negli archivi tradizionali si applicano ai carteggi e agli epistolari.

Attualmente è al vaglio la possibilità di affidare un'ulteriore copia degli archivi al progetto nazionale Magazzini Digitali, avviato nel 2006 dalla Fondazione Rinascimento Digitale, dalla Biblioteca nazionale centrale di Firenze e dalla Biblioteca nazionale centrale di Roma. La conservazione digitale assicurata dai depositi digitali affidabili o fidati (trusted or trustworthy digital repositories) di un servizio pubblico è una ulteriore garanzia che archivi digitali di autore così preziosi e che rischiano l'oblio vengano adeguatamente conservati nel lungo termine.

Bibliografia

Bergamin, G., and Messina, M. 2010. Magazzini digitali: dal prototipo al servizio. *DigItalia* 1, 115-122.

Black, Paul E. 2008. Levenshtein distance. *Dictionary of Algorithms and Data Structures [online]*. U.S. National Institute of Standards and Technology, retrieved 12/09/2019

Costa, M., Gomes, D., and Silva, M. J. 2017. The evolution of web archiving. *International Journal on Digital Libraries*, *18*(3), 191-205.

Klein, M., Shankar, H., Balakireva, L., and Van de Sompel, H. 2019. The Memento Tracer Framework: Balancing Quality and Scalability for Web Archiving. *International Conference on Theory and Practice of Digital Libraries* 163-176.

Masanès, J. 2006. Web archiving: issues and methods. *Web Archiving*. Springer, Berlin, Heidelberg.

Weston, P. G., Carbé E. and Baldini P. 2017. Hold it All Together: a Case Study in Quality Control for Born-Digital Archiving. *Qualitative and Quantitative Methods in Libraries* 5.3, 695-710.

Weston, P. G., Carbé E. and Baldini P. 2017. If bits are not enough: preservation practices of the original contest for born digital literary archives. *Bibliothecae. it* 6.1, 154-177.