SAPIENZA UNIVERSITÀ DI ROMA
DIGILAB CENTRO INTERDIPARTIMENTALE DI RICERCA E SERVIZI

ASSOCIAZIONE PER L'INFORMATICA UMANISTICA E LA CULTURA DIGITALE

DiXiT Digital Scholarly Editions Initial Training Network

AIUCD 2017 Conference
3rd EADH Day
DiXiT Workshop
"The educational impact of DSE"

23-28 January 2017 Rome

Ex Vetrerie Sciarra
Via dei Volsci, 122

http://aiucd2017.aiucd.it
Info: aiucd2017org@gmail.com

With the patronage of:

ADHO Alliance of Digital Humanities Organizations

Facoltà di Lettere e Filosofia
Sapienza Università di Roma

eadh

Indice

AIUCD 2017 Conference, 3rd EADH Day, DiXiT Workshop *"The Educational impact of DSE"*
Rome, 23-28 January 2017

AIUCD CONFERENCE

With the patronage of:

LONG PAPERS

Using Local Grammar Induction for Corpus Stylistics

Andrew Salway, Uni Research Computing, Bergen, Norway, Andrew.Salway@uni.no
Michaela Mahlberg, University of Birmingham, UK, m.a.mahlberg@bham.ac.uk

Introduction: goals of the paper

This paper proposes the use of local grammar induction – a text mining technique for data-driven content analysis – as part of corpus stylistic investigations. The contribution of this paper is twofold. Firstly we will sketch the theoretical context for our work within the field of digital humanities and secondly we will explore the use of local grammar induction by drawing on our previous work in corpus stylistics to validate this method.

Corpus stylistics in digital humanities

Because of the increasing availability of literary texts in digital form it is timely to consider how new language technologies can enhance literary and narrative analyses, not only of single works, but of authors' oeuvres, genres, and centuries' worth of literature. Within the field of digital humanities a range of methods and approaches illustrates the variety of theoretical and disciplinary concerns that guide the development and application of computational techniques. While studies in computational and statistical linguistics, for instance, might be concerned with identifying the authorial finger print of a particular author (e.g. Hoover, 2002), proponents of what has come to be known as 'distant reading' are interested in the general trends that are visible across large data sets (Moretti, 2013). To represent such trends, various types of visualization techniques are used (see Jänicke et al., 2015 for an overview).

In this paper, we focus on an area of the digital humanities that is referred to as 'corpus stylistics'. Research in this area uses corpus linguistic methods for the analysis of literature, stressing the relationship between linguistic description and literary appreciation (e.g. Toolan, 2009; Fischer-Starcke, 2010; McIntyre, 2010). The success of corpus stylistic research depends on the extent to which automatic text analysis techniques can draw the researcher's attention to linguistic patterning that relates to literary functions. Equally, computer-assisted methods are used to provide evidence for perceived literary effects. Focusing on narrative fiction in particular, work in corpus stylistics emphasizes the relationship between patterns in the text and the functions associated with them. Once such relationships are identified, it becomes possible to compare the distribution of patterns, and hence functions, within or across oeuvres, genres, and over time.

Work in corpus stylistics can draw on a repertoire that includes descriptive categories such as the lexical item (Sinclair, 2004), patterns in the Pattern Grammar sense (Hunston and Francis, 2000), or types of lexical priming (Hoey, 2005). While all of these units or descriptive categories are identified on the basis of large corpora of electronic texts, work in corpus stylistics then emphasises their applicability to the analysis of individual texts. In practical terms, work in corpus stylistics has largely focused on words, key words and word clusters (*n*-grams) as the basic units of analysis for identifying potentially interesting linguistic patterning in corpora of literary texts. This focus reflects the reliance on standard techniques and software packages in corpus linguistics. However, research in corpus stylistics has shown that it is valuable to reconsider what are useful textual units for the analysis of literature and explore alternatives to standard search and display techniques (Mahlberg et al., 2013; forthcoming).

Applying local grammar induction to Dickens's novels

Here, for the first time, we propose the use of local grammar induction (Salway and Touileb, 2014) as a technique for identifying salient lexico-grammatical patterning in corpora of literary texts. Our motivation is that the local grammar induction technique appears to simultaneously capture and integrate sequential and paradigmatic information about the co-text of a chosen word. This contrasts with existing techniques that give either sequential information (*n*-grams, also referred to as word clusters) or paradigmatic information (collocations, skip-grams, distributionally-derived word classes, topic modelling). Like existing techniques it is data-driven in the sense that it does not require manual corpus annotation, nor does it depend on resources such as lexicons, taggers or parsers. This helps to maintain portability between different languages and text types, and to alleviate the bias that can arise from a priori annotation schemes and linguistic resources.

As a first step towards using local grammar induction for corpus stylistics, this paper aims to validate the technique by analyzing the lexico-grammatical patterning it recovers from a corpus of Charles Dickens's novels. The point is to determine whether this patterning relates to known literary and narrative phenomena of interest in a corpus that has already been well studied, such that the technique could be used with some confidence in research on other material.

Local grammar induction (Salway and Touileb, 2014) is based on a grammar induction algorithm – ADIOS (Solan et al., 2005) – which identifies significant sequences and equivalence classes (cf. lexico-grammatical patterning) from partially overlapping sentences in a corpus. The algorithm is made 'local' by focusing it to analyze only text fragments around a single word at a time (cf. a concordance), by prioritizing patterning closest to the word, and optionally by presenting text snippets from a specialist/restricted corpus.

This is in part inspired by the concept of local grammar (Gross, 1997) but it is not clear exactly how the output from the technique maps to the concept of local grammar in a consistent way. However, the technique has been used successfully as a way to elucidate linguistic patterning in specialist corpora in the exploratory stages of investigations. In these cases it was effective in identifying linguistic patterning that was interpreted in terms of domain specific information structures, which then became the basis for further analyses. For example, the following pattern was induced from the set of all text fragments around the word *expressed* in a corpus containing the minutes of international climate change negotiations (Salway et al., 2014).

```
(COUNTRY expressed
((disappointment|concern) that)|
((support|appreciation) for)|
((readiness|willingness) to)|
(satisfaction (with the) (outcome |reconstitution|functioning|work)
(of the)))
```

The case study we present in this paper assesses the extent to which local grammar induction can support the corpus stylistician in selecting potentially relevant linguistic patterning that relates to narrative functions in a corpus of literary texts. We chose to work with Charles Dickens's novels because they have already been subject to extensive corpus stylistic research, e.g. Mahlberg (2013). So our focus is not on generating new insights into Dickens's writing but rather to validate the results generated by the local grammar induction against what is already known about the literary works.

The corpus we use is an early version of the Non-Quote corpus outlined in Mahlberg et al. (forthcoming), i.e. the text from Dickens's 15 novels that occurs outside of quotation marks; this amounts to about 2.5 million words. We focus on narratorial description separately from speech because we expect to find different kinds of linguistic patterning in each, as argued in Mahlberg et al. (forthcoming). We are

specifically interested in patterns of body language descriptions. Prior work (e.g. Mahlberg, 2013) found that such patterning around body part nouns relates to distinctive aspects of Dickens's style and his externalized techniques of characterization. Therefore, in our case study we use a set of body part nouns as the starting point for our local grammar induction algorithm and examine the resulting patterns. We select the most frequent patterns for detailed discussion in light of our previous findings on Dickens's narrative techniques.

Narrative patterns in Dickens – results of the case study

Our results show that the linguistic patterning around body parts retrieved by the algorithm include patterns with textual functions describing character interactions. For instance, local grammar patters including prepositions can be interpreted as showing that body language is a form of interaction where characters position themselves in relation to other characters or objects in their environment. Other local grammar patterns include–*ing* forms which are reflective of the fact that body language often is presented as circumstantial information taking place simultaneously with other activities. In our paper, we will also discuss patterns around characters' names with the string *CHARACTER* replacing the variety of names in the corpus. Such patterns include a range of speech verbs for instance:

```
(COUNTRY expressed
((disappointment|concern) that)|
((support|appreciation) for)|
((readiness|willingness) to)|
(satisfaction (with the) (outcome |reconstitution|functioning|work)
(of the)))
```

Overall, our case study shows that the patterns identified by the algorithm can be interpreted in line with our previous work on body language descriptions in Dickens. This case study gives us confidence that local grammar induction is effective in identifying linguistic patterning that can usefully be interpreted in terms of its functional relevance for the corpus under analysis. Thus the method, which captures and integrates sequential and paradigmatic information about the co-text of chosen words, has the potential to be a useful addition to the corpus stylistician's toolkit.

Bibliographic References

Fischer-Starcke, Bettina. 2010. *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. Continuum, London.

Gross, Maurice. 1997. The Construction of Local Grammars. In: E. Roche and Y. Schabes (eds.), *Finite-State Language Processing*. The MIT Press, Cambridge MA: 329-354.

Hoey, Michael. 2005. *Lexical Priming. A New Theory of Words and Language*. Routledge, London.

Hoover, David L. 2002. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing,* 17(2):157–180.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English*. John Benjamins, Amsterdam.

Jänicke, S., G. Franzini, M. F. Cheema, and G. Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *Eurographics Conference on Visualization (EuroVis)*.

Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction*. Routledge, New York.

Mahlberg, Michaela, Catherine Smith, and Simon Preston. 2013. Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels. *International Journal of Corpus Linguistics,* 18(1):35-56.
Mahlberg, Michaela, Peter Stockwell, Johan de Joode, Catherine Smith, and Matthew Brook O'Donnell. Forthcoming. CLiC Dickens – Novel uses of concordances for the integration of corpus stylistics and cognitive poetics, *Corpora*.
McIntyre, Dan. 2010. Dialogue and characterisation in Quentin Tarantino's Reservoir Dogs: a corpus stylistic analysis. In: *Language and Style*. Palgrave, Basingstoke, UK, pp. 162-182.
Moretti, Franco. 2013. *Distant Reading.* Verso, London.
Salway, Andrew and Samia Touileb. 2014. Applying Grammar Induction to Text Mining. *Procs. ACL 2014*.
Salway, Andrew, Samia Touileb, and Endre Tvinnereim. 2014. Inducing Information Structures for Data-driven Text Analysis. *Procs. ACL 2014 Workshop on Language Technologies and Computational Social Science*.
Sinclair, John. 2004. *Trust the Text. Language, Corpus and Discourse*. Routledge, London.
Toolan, Michael. 2009. *Narrative Progression in the Short Story. A Corpus Stylistic Approach*. John Benjamins, Amsterdam.
Solan, Zach, David Horn, Eytan Ruppin, and Shimon Edelman. 2005. Unsupervised learning of natural languages. *Procs. of the National Academy of Sciences* 102(33):11629-11634.

Educazione e Big Data: un progetto di analisi dei documenti ufficiali degli Istituti Comprensivi

Gianfranco Bandini, Università di Firenze, bandini@unifi.it
Marina Baretta, Dipartimento per l'Istruzione, MIUR, marina.baretta@istruzione.it
Federico Betti, Istituto comprensivo "G. Falcone", Cascina, federico.betti@istruzione.it
Silvano Cacciari, Università di Firenze, mcsilvan@gmail.com
Renza Campagni, Università di Firenze, renza.campagni@unifi.it
Rosa De Pasquale, capo dipartimento Istruzione e Formazione del Miur, DPIT.segreteria@istruzione.it
Paolo Ferragina, Università di Pisa – SoBigData, ferragina@di.unipi.it
Fosca Giannotti, CNR – SoBigData, fosca.giannotti@isti.cnr.it
Filomena Maggino, Università di Firenze, filomena.maggino@unifi.it
Stefano Oliviero, Università di Firenze, stefano.oliviero@unifi.it
Dino Pedreschi, Università di Pisa – SoBigData, pedre@di.unipi.it
Maria Vincelli, Liceo scientifico "U. Dini", Pisa, maria.vincelli1@istruzione.it

L'infrastruttura di ricerca europea per i Social Big data

La ricerca europea si colloca nello scenario internazionale con una spiccata propensione matematica, ingegneristica e innovativa; al tempo stesso presenta una vasta area di ricerche umanistiche che hanno sperimentato il contesto digitale fin dagli anni Settanta, in vari modi e con esiti differenziati a seconda delle comunità di ricerca. Nel campo della ricerca sociale gli approcci che vengono contrassegnati con l'etichetta "big data" rappresentano uno dei settori più promettenti e ricchi di sfide (Boyd, Crawford, 2012). La difficoltà intrinseca del lavoro di analisi di rilevanti moli di dati è unita a risultati molto incoraggianti che in alcuni casi portano perfino a riconsiderare l'epistemologia delle discipline o, comunque, a metterne in discussione alcuni capisaldi tradizionalmente considerati.

Fra i più interessanti progetti che stanno offrendo approfondimenti teorici e un'infrastruttura per la ricerca su dati di "natura sociale" – quali documenti testuali, post di social networks, news, tracce di mobilità o di interazione sociale, ecc. – occorre segnalare il progetto europeo SoBigData (http://www.sobigdata.eu), svolto da una rete di 12 centri europei, di cui 4 sono partner italiani quali: il Consiglio Nazionale delle Ricerche con gli istituti ISTI e IIT (coordinatori dell'Infrastruttura), l'Università di Pisa, la Scuola Normale Superiore e l'IMT di Lucca. I partner europei sono l'Università di Sheffield, il King's College di Londra, il Fraunhofer Institut, l'Università Leibniz di Hannover, l'Università estone di Tartu, l'Università finlandese di Aalto, l'ETH di Zurigo, e la Technische Universiteit di Delft. Quindi una rete di centri di eccellenza scientifica su alcuni temi chiave quali: data and text mining, social media analytics, analisi delle reti sociali e della mobilità, visualizzazione e analitica visuale, e-government data analytics, open data, e-health data, privacy ed etica dell'ICT.

L'infrastruttura mette a disposizione dati, strumenti e competenze di *data scientist* per condurre esperimenti di *big data analytics* da parte di ricercatori, innovatori, startuppers, policy-makers, istituzioni pubbliche. Esperimenti rivolti ad estrarre senso dalle tracce digitali delle attività umane registrate appunto nei Big Data, ed usare la conoscenza estratta per nuove scoperte scientifiche, ad esempio in campo sociale ed economico, per nuovi servizi e prodotti, o a supporto delle decisioni. SoBigData, insomma, ha l'ambizione di diventare un ecosistema di tecnologie e di persone in grado di liberare il potenziale dei Big Data come *commons*, come bene pubblico, aperto e accessibile, in un quadro etico di valori imprescindibili: trasparenza dei processi e delle finalità, rispetto della privacy, fiducia e responsabilità.

È proprio in questo contesto che si innesta il progetto descritto in questo articolo. Esso trarrà pieno beneficio dagli strumenti e competenze messe a disposizione dall'Infrastruttura SoBigData analizzando in modo innovativo (come verrà dettagliato nel seguito) l'insieme di dati prodotti dal sistema informativo e comunicativo del mondo scolastico italiano tramite le tecniche allo stato dell'arte sviluppate dai ricercatori dell'infrastruttura nell'ambito del Text Mining e della Big Data Analytics.

Big Data, educazione, ricerca sociale tra sfide e opportunità

Le scienze sociali e umanistiche, in particolare quelle che più di altre si sono avvicinate ai processi educativi, hanno sviluppato molti metodi per concettualizzare i fenomeni, analizzare le situazioni, creare delle modalità di studio - spesso a base statistica - per capire i fattori in gioco e le possibili dinamiche evolutive dei fenomeni. Ciò che è stato oggetto di lunghe e appassionate discussioni è soprattutto il contrasto tra approcci idiografici della realtà sociale e approcci nomotetici. Nell'area delle discipline pedagogiche ciò si è tradotto in una sovrabbondanza di studi basati su singoli casi e su una metodologia fortemente qualitativa. In accordo con le più aggiornate tendenze della ricerca internazionale, anche nell'area pedagogica si è fatta strada con maggior forza una tendenza sperimentale che ha fortemente rivalutato le metodiche quantitative. Di fatto non è più possibile parlare di qualitativo e quantitativo come approcci contrapposti, così come non è possibile contrapporre micro analisi e macro analisi. Nell'ambito della ricerca linguistica e letteraria, in particolare, si è giustamente affermato che *distant reading* e *close reading* sono aspetti complementari e, di fatto, largamente ibridati (Jockers, 2013).

In questo contesto i Big Data costituiscono una grande sfida a modificare non solo le metodologie della ricerca, ma anche gli assetti operativi dei *team* di studiosi al lavoro (Morabito, 2015). Il settore educativo, e in modo speciale quello dell'istruzione, ha un grande bisogno di questo tipo di contributi, volti a dare valore a masse imponenti di dati che prima giacevano nei faldoni cartacei e oggi rischiano di "prendere la polvere" nei server. L'accumulo di dati in formato digitale, che ha visto una crescita imponente soprattutto dall'inizio del presente millennio, costituisce certamente un'opportunità di grande rilievo. Si tratta di vedere se è possibile, proprio grazie agli algoritmi, ai software e alle metodologie per l'analisi dei Big Data, fornire degli strumenti che diano delle chiavi di lettura e delle leve operative non solo ai *policy makers*, ma anche a tutti i soggetti coinvolti nei processi di *governance* e trasformazione della scuola con una potenziale ed evidente ricaduta sociale, nell'ottica di quello che comunemente viene definito *public engagement* e che rientra nella terza missione dell'università (Mayer-Schönberger, Cukier, 2014).

I Big data e il sistema di istruzione

Il progetto che presentiamo in queste pagine, ancora in sviluppo, trova le sue motivazioni nella complessità della trasformazione strutturale che il sistema d'istruzione italiano è chiamato ad affrontare (Barone, Luijkx, Schizzerotto, 2010; Viteritti, Giancola, 2015). I rapporti di autovalutazione e i piani di miglioramento, l'organico potenziato o le rinnovate forme di reclutamento e formazione sono solo alcune novità da cui gli attori del sistema scolastico sono stati coinvolti e con cui si stanno misurando nelle loro attività quotidiane di *governance* e di progettualità didattico-educativa, senza tuttavia possedere tutti gli strumenti più idonei a gestire la direzione del cambiamento in corso.

Al fine di affrontare questa sfida in modo completo ed efficace il progetto in oggetto mette insieme studiosi di diversa appartenenza disciplinare e provenienti da diversi enti di ricerca, MIUR e istituti scolastici che riconoscono la necessità di unire le professionalità del *data scientist* a quelle di altri specialisti (Calders, Mykola, 2012; Gruppo di lavoro MIUR, 2016) così da formare un *team* di lavoro multi-disciplinare e multi-settoriale. Il team dell'infrastruttura europea SoBigData metterà a disposizione non solo le proprie conoscenze e competenze informatiche nell'ambito della memorizzazione, indicizzazione e

analisi di Big Data, ma anche il supporto all'uso di un insieme molto ampio di sofisticati strumenti software disponibili all'interno dell'infrastruttura (si veda dopo).

I dati che il progetto sta esaminando riguardano (si segnala che USR sta per Ufficio Scolastico Regionale):

Tipologia documenti	Fonte	Anni scolastici
RAV - Rapporto di autovalutazione	MIUR	2015-2016 / 2016-2017
PTOF - Piano Triennale dell'Offerta Formativa (e revisione)	MIUR	2015-2016 / 2016-2017
POF - Piano dell'Offerta Formativa	USR	2013-2014 / 2014-2015
PdM - Piano di Miglioramento	USR	2015-2016 / 2016-2017
PA - Piani di Ambito di formazione del personale docente	USR	2016-2017

Ogni documento viene redatto dal singolo istituto una volta all'anno: se ci riferiamo agli Istituti Comprensivi complessivamente si tratta di 38.400 documenti (che potrebbero subire una riduzione del 15-30 % in ragione della difficoltà di reperimento dei dati non conservati a livello centrale).

Si tratta evidentemente di BigData non tanto rispetto alla loro mole (volume), ma piuttosto rispetto al loro valore, alla varietà (testo e numeri) e, in un futuro oramai prossimo, anche alla velocità, quando presumibilmente la redazione dei questi e altri documenti scolastici avverrà con una frequenza temporale maggiore rispetto a quella odierna.

Il progetto può essere così sintetizzato:

- selezione di un panel di istituti comprensivi (4.800 in Italia) rappresentativi della situazione nazionale, con un eventuale ulteriore campione di scuole secondarie;
- utilizzazione di tutti i dati ufficiali degli istituti (in particolare: Rapporto di Autovalutazione, Piano Triennale dell'Offerta Formativa);
- sviluppo di un modello di analisi dei dati che metta in primo piano gli aspetti linguistici e comunicativi. Tale modello terrà conto dei risultati della ricerca storica sulla storia della scuola e delle tendenze di lungo periodo che sono state poste in evidenza, anche in sede internazionale (Guldi, Armitage, 2014; Graham, Milligan, Weingart, 2015).
- implementazione di una fase esplorativa per una prima classificazione delle informazioni, ovvero dei dati/documenti su cui si esplica l'analisi, che permetta di studiare il fenomeno in termini statistici quantitativi (Campagni et al., 2015).
- operazione di *data* e *text mining* (Azzalini, Scarpa, 2004; Bolasco, 2005; Ferragina, Scaiella, 2012; Giannotti et al., 2013) con particolare attenzione a:
    - analisi delle tematiche (e sviluppo di mappe semantiche) con identificazione di quelle che sono caratteristiche di specifiche realtà scolastiche, in relazione alla caratterizzazione geografica e/o alla tipologia di istituto/scuola;
    - analisi della correlazione tra l'evoluzione sociale e le sue problematiche emergenti con il linguaggio/tematiche affrontate nei documenti scolastici, eventualmente con una loro caratterizzazione geografica e di ordine scolastico;

- confronto tra POF e PTOF, e modifiche annuali dei singoli PTOF, per monitorare l'andamento della politica e progettualità scolastica al fine di rendere immediatamente evidenti le differenze sintattiche tra essi, con un impatto sia a uso interno della singola scuola/istituto sia a uso esterno per facilitare, per esempio, il processo di valutazione della stessa. Tale analisi consentirà inoltre di mappare a livello temporale le tematiche di interesse per la singola scuola; si prevede anche che il clustering delle modifiche ai documenti suddetti e delle loro tematiche possa portare a identificare trend nelle politiche scolastiche da confrontare, a livello macro, con eventuali azioni legislative o progettuali lanciate dal MIUR, a livello micro, con iniziative di singole scuole o reti di scuole;
- identificazione di operazioni di copia-incolla tra documenti disponibili sul Web e i documenti scolastici oggetto della presente analisi al fine di costruire un grafo in cui i nodi sono questi documenti e gli archi denotano operazioni di copia-incolla tra essi con opportuni attributi che le descrivono. L'analisi di questo grafo permetterà di valutare l'entità e il contenuto delle copie, indentificare le sorgenti di queste e possibilmente la loro marca temporale, e quindi valutare il percorso seguito nella generazione di questi documenti, ed eventualmente se ci sono Istituti (documenti) guida o l'impatto (linguistico) della normativa sulla redazione dei documenti scolastici. Si potrebbe anche immaginare di valutare se la scelta del documento sorgente, da cui copiare, segue i suggerimenti dei motori di ricerca o è influenzato da altri canali di comunicazione;
- identificazione di gruppi di documenti simili per tematiche o per variazioni di queste tra documenti successivi (come precedentemente accennato), con l'obiettivo di raggruppare scuole in base alla loro progettualità e alle loro "esigenze" così da suggerire reti di collaborazione e scambio esperienze. Solo per fare un esempio, questo progetto potrebbe riuscire a semplificare l'organizzazione del Piano per la Formazione dei Docenti poiché, a partire dai singoli PTOF e RAV, potrebbe arrivare a stabilire le azioni formative comuni e replicabili non solo negli Istituti della singola Rete, ma anche all'esterno, creando un database di buone pratiche, sempre agognato e mai realizzato negli Istituti.

Il principale risultato del lavoro qui sinteticamente descritto può essere individuato nell'offrire agli istituti scolastici una visione panoramica e comparativa del loro lavoro, che consenta di monitorare l'effettiva coerenza strategica fra le azioni intraprese e gli obiettivi prefissati, di individuarne gli aspetti originali e innovativi, di identificare reti di scuole "simili" per obiettivi progettuali e necessità formative. La rappresentazione dei fenomeni consentirà di approfondire le varie componenti che li descrivono e consentirà di capire che tipo di interventi adottare per rispondere al meglio agli obiettivi prefissati.

Sarà possibile più in generale catturare eventuali fenomeni legati alla verticalità e trasversalità delle tematiche tra le diverse scuole, evidenziando possibilmente fenomeni di particolari aree geografiche o interrelazioni tra la tipologia e i gradi degli istituti scolastici. Lo sviluppo di questo strumento di analisi della vita scolastica potrebbe essere esteso per operare sui dati provenienti da tutte le scuole italiane di ogni ordine e grado.

Nel raggiungere questi risultati il progetto si servirà di strumenti software e tecniche algoritmiche allo stato dell'arte sviluppate dal team SoBigData nel corso di questi anni e, in parte, disponibili all'interno dell'Infrastruttura di Ricerca europea su menzionata. Tra gli altri preme ricordare gli algoritmi per l'analisi dei *big data* di natura testuale che permettono di superare le limitazioni del paradigma classico dell'Information Retrieval, denominato *bag-of-words,* che analizza il testo come un insieme di parole senza tentare di comprenderne il significato ignorando così i problemi di polisemia, sinonimia, e i riferimenti alla conoscenza umana insiti nel testo che inducono tale approccio a ottenere prestazioni non soddisfacenti. Per ovviare a queste limitazioni, il team del SoBigData è stato tra i primi gruppi di ricerca al mondo a sviluppare nel 2010 una nuova tecnica di "annotazione semantica" dei testi che identifica, efficientemente ed efficacemente, sequenze significative di termini nel testo in input e le collega alle pagine di Wikipedia che

sono pertinenti a descriverne il significato. Il software che implementa questi algoritmi di annotazione prende il nome di TagMe. Esso è in grado di annotare testi in lingua italiana, inglese e tedesca (potenzialmente estendibile ad altre lingue). TagMe risulta correntemente uno dei tool di riferimento internazionale sia a livello accademico che industriale nel contesto dell'annotazione testuale ed è disponibile pubblicamente all'indirizzo: http://tagme.di.unipi.it/.

I risultati scientifici alla base di TagMe sono pubblicati nei Proceedings delle più prestigiose conferenze internazionali del settore dell'Information Retrieval (si vedano ad esempio Ferragina, Scaiella 2012; Scaiella et al. 2012; Cornolti et al. 2016), e hanno ricevuto diversi premi tra cui due Google Faculty Award, negli anni 2010 e 2012. TagMe è stato utilizzato con successo in varie applicazioni quali la classificazione e il *clustering* di diversi tipi di testi. L'elemento di successo in tutte queste applicazioni consiste nel fatto che TagMe consente di rappresentare un testo non solo come un insieme di parole quanto piuttosto come un "grafo di concetti" derivati dall'annotazione semantica del testo in input. I risultati scientifici di questi ultimi anni, ottenuti da vari gruppi di ricerca internazionali, hanno dimostrato che questa rappresentazione è molto potente perché fornisce ai testi una "contestualizzazione" particolarmente significativa la quale permette di superare le limitazioni del classico paradigma del *Bag-of-Words* menzionate precedentemente. Infatti, non solo i concetti identificati hanno un significato univoco (a differenza delle parole) ma in più, essendo essi parte di una base di conoscenza (Knowledge Base) rappresentata in forma di grafo (come Wikipedia o Wikidata), essi sono associati ad altri dati e *linkati* tra loro.

Si ritiene che TagMe e l'insieme degli altri algoritmi e software per l'analisi testuale disponibili nell'infrastruttura SoBigData possano costituire un potente strumento a supporto delle indagini su specificate.

Riferimenti Bibliografici

Azzalini, Adelchi, Scarpa, Bruno. 2004. *Analisi dei dati e datamining*. Milano: Springer-Verlag Italia.

Barone, Carlo, Luijkx, Ruud, Schizzerotto, Antonio. "Elogio dei grandi numeri: il lento declino delle disuguaglianze nelle opportunità di istruzione in Italia." *Polis* XXIV.1 (2010): 5-34.

Bolasco, Sergio. "Statistica testuale e text mining: alcuni paradigmi applicativi." *Quaderni di Statistica* 7 (2005): 1-36.

Boyd, Danah, Crawford, Kate. "Critical Questions for Big Data." *Information, Communication &* Society 15.5 (2012): 662-679.

Calders, Toon, Pechenizkiy, Mykola. "Introduction to the special section on educational data mining." *ACM SIGKDD Explorations Newsletter* 13.2 (2012): 3-6.

Campagni, R.; Merlini, D.; Verri, M. C. 2015. "An Analysis of Courses Evaluation Through Clustering." In: Zvacek S., Restivo M., Uhomoibhi J., Helfert M.. *Computer Supported Education*, 211-224. Switzerland: Springer International Publishing.

Cornolti, Marco, Ferragina, Paolo, Ciaramita, Massimiliano, Rued, Stefan, Schutze, Hinrich. "A piggyback system for joint entity mention detection and linking in web queries". *International World Wide Web Conference (WWW)*, Vancouver (Canada), 567-578, 2016.

Cristóbal, Romero, Ventura, Sebastián 2014. "A survey on pre-processing educational data." In *Educational Data Mining. Studies in Computational Intelligence,* edited by Alejandro Peña-Ayala, 29–64. Switzerland: Springer International Publishing.

Cohen, Daniel J., Rosenzweig, Roy 2005. *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadelphia: University of Pennsylvania Press.

Daniel, Ben Kei, ed. 2016. *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*. Switzerland: Springer International Publishing.

Ferragina, Paolo, Scaiella, Ugo. “Fast and accurate annotation of short texts with Wikipedia pages”. *IEEE Software*, 29(1): 70-75, 2012.
Giannotti, Fosca, Lakshmanan, Laks V. S., Monreale, Anna, Pedreschi Dino, Wang, Hui "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases," *IEEE Systems Journal* 7.3 (2013): 385-395.
Graham, Shawn, Milligan, Ian, Weingart, Scott. 2015. *Exploring Big Historical Data: The Historian's Macroscope.* London: Imperial College Press.
Greengrass, Mark, Hughues, Lorna M., eds. 2008. *The Virtual Representation of the Past*, Aldershot: Ashgate Publishing.
Gruppo di lavoro Miur 2016. *Big Data@Miur, Rapporto del gruppo di lavoro* (costituito con D.M. 2710112016), 28 luglio 2016 (http://www.istruzione.it/allegati/2016/bigdata.pdf).
Guldi, Jo, Armitage, David 2014. *The history manifesto.* Cambridge: Cambridge University Press.
Jockers, Matthew 2013. *Macroanalysis: Digital Methods & Literary History.* Champaign, IL: University of Illinois Press.
Manning, Patrick. 2013. *Big Data in History.* London: Palgrave MacMillan.
Mayer-Schönberger, Viktor, and Cukier, Kenneth 2014. *Learning with Big Data: The Future of Education*. New York: Houghton Mifflin Harcourt.
Morabito, Vincenzo 2015. *Big Data and Analytics: Strategic and Organizational Impacts*. Switzerland: Springer International Publishing.
Pollo, Simone. “Progresso scientifico e progresso morale. Sentimentalismo, oggettività e scienza.” *Rivista di filosofia* 107.2 (2016): 219-240.
Scaiella, Ugo, Ferragina, Paolo, Marino, Andrea, Ciaramita, Massimiliano. “Topical Clustering of Search Results”. *ACM International Conference on Web Search and Data Mining (WSDM)*, Seattle (USA), 223-232, 2012.
Van Eijnatten, Joris, Toine Pieters, and Jaap Verheul. "Big data for global history: The transformative promise of digital humanities." *BMGN-Low Countries Historical Review* 128-4 (2013): 55-77.
Viteritti, Assunta, Giancola, Orazio. “Il ruolo delle grandi survey in campo educativo. L'indagine PISA e il governo dell'educazione tramite i numeri.” *Rassegna Italiana di Sociologia* 56.3-4 (2015): 555-580.

Dalla Digital Culture al Digital Cultural Heritage: l'evoluzione impossibile?

Nicola Barbuti, Dipartimento di Studi Umanistici (DiSUM) Università degli Studi di Bari,
nicola.barbuti@uniba.it

Premessa

Nel dibattito attuale sul futuro dell'attuale Evo digitale, la questione della sostenibilità e conservazione dei dati digitali costituisce forse il principale nodo irrisolto.

Nonostante già da diversi anni autorevoli voci scientifiche si siano dedicate e si dedichino a discutere il problema[1], solo recentemente l'argomento è diventato oggetto di attenzione diffusa essendone stata riconosciuta la condizione di emergenza, in particolare in seguito all'allarme lanciato da Vinton Cerf agli inizi del 2015 circa il rischio che la contemporaneità diventi per i posteri il buco nero nella storia evolutiva dell'umanità.[2]

Forse anche in seguito all'allarme lanciato da Cerf, l'AgID si è soffermato ampiamente sulla questione nelle recenti *Linee Guida sulla conservazione dei documenti informatici* del dicembre 2015, sottolineandone l'estrema importanza.[3]

Tuttavia, tranne rarissime eccezioni, nel dibattito sul tema difficilmente si prende in considerazione l'aspetto che, invece, da secoli costituisce la sola condizione perché un dato, di qualunque natura esso sia (analogica materiale o immateriale, intangibile, magnetico, digitale, etc.), sia conservato e trasferito nello spazio e nel tempo: la sua identificazione quale potenziale fonte di conoscenza e informazione e, quindi, quale retaggio culturale destinato a essere per le generazioni future testimonianza del presente/corrente e veicolatore di conoscenza e memoria storica cui accedere per studiarla, conoscerla, comprenderle, riusarla.

A spasso nel tempo: non c'è conservazione digitale se non è retaggio culturale

Applicando l'assunto sopra esposto all'evo digitale contemporaneo, perché nella progettazione e pianificazione di un ecosistema di conservazione digitale si presti la massima attenzione alla sua sostenibilità non si può più prescindere dal conferire alla cultura digitale contemporanea (digital culture) l'identità di potenziale retaggio culturale digitale del futuro (digital cultural heritage), in quanto costituita da processi/contenuti digitali cui è connaturata la funzione di potenziali fonti cognitive/informative deputate a favorire la conoscenza della facies culturale contemporanea per le generazioni future, quindi a esserne memoria storica.

1 Valga qui ricordare il sempre attuale lavoro di M. Guercio, *Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali*, Roma-Bari, Laterza, ed. 2013. Insieme a diversi altre autorevoli voci, dopo la Guercio si è occupato del tema dal punto di vista della memoria V. Gambetta, *La conservazione della memoria digitale*, [Rubano], Siav, 2009. In alcuni passaggi del testo si ritrovano sostanziali similitudini con le parole di E. Casanova, *Archivistica*, Siena 1928 a proposito della fondamentale funzione degli archivi per la conservazione della memoria storica delle società.

2 Per ascoltare l'intervista a Cerf: http://www.bbc.com/news/science-environment-31450389. Un interessante articolo a commento su: http://www.wired.it/attualita/2015/02/16/vint-cerf-futuro-medievale-bit-putrefatti/

3 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).

Ne consegue che, fin dal momento dell'analisi e progettazione di un sistema digitale, deve esserne riconosciuto il potenziale culturale connaturato sia a esso che ai dati che dovrà contenere e disseminare: assunto che obbliga a definire regole e procedure certe e ineludibili da seguire nella strutturazione dell'insieme per svincolarlo dalla preminente funzione di valorizzazione del presente/corrente e conferirgli definitivamente l'identità di potenziale componente del retaggio culturale da trasferire alle generazioni future.

L'orientamento in base al quale definire gli obiettivi cui un ecosistema di conservazione digitale realmente efficace ed efficiente dovrebbe pervenire non possono perciò, a nostro parere, prescindere dalla riflessione su alcune essenziali considerazioni di seguito enucleate.

1. Il digitale e la digitalizzazione sono i nuovi processi/contenuti culturali rappresentativi dell'evo contemporaneo. In quanto tali, non si può più eludere la necessità di conferire sia ai contenitori che ai contenuti digitali il ruolo di potenziali testimoni culturali del presente/corrente, componenti essenziali della memoria storica dell'evoluzione culturale in atto. Questa memoria ha valore solo se debitamente ed efficacemente organizzata, gestita, conservata e trasferita, in modo da renderla fruibile secondo criteri di omogeneità, accessibilità, intellegibilità, interpretabilità e riusabilità per le generazioni future.
2. Si rende perciò indispensabile ripensare i processi/contenuti digitali, emancipandoli dall'identificazione attuale di strumenti di riproduzione/rappresentazione della realtà ad accesso globale utili quasi esclusivamente a valorizzare il presente, per evolverli in nuove tipologie di documentazione il cui insieme concorre a comporre il retaggio culturale da trasferire alle generazioni future quale fonte per la conoscenza del presente. Tuttavia, perché diventino tali, è indispensabile garantirne sostenibilità, permanenza, accessibilità e fruibilità di lungo termine.
3. Ne consegue che la sopravvivenza di quanto prodotto dal digitale e dalla digitalizzazione – da identificare nell'insieme di processi metodologici e tecnologici, sistemi, informazioni, strutture dei dati e dei metadati, strutture dei contenuti e loro rappresentazione iconica, descrizione dei dati, degli insiemi di dati e dei contenuti di ciascun dato e dei dati organizzati in sistemi complessi – di diritto e di fatto è non da oggi, ma già da l'altro ieri (e da allora un'infinità di retaggio digitale culturale è già andato irreversibilmente distrutto) la principale emergenza su cui è necessario concentrarsi per la ricerca di soluzioni da adottare. In tale direzione, gli open data costituiscono senz'altro una componente essenziale e un fondamentale punto di partenza, ma non sono da considerare e non potranno essere la soluzione al problema.
4. Come indicato anche nelle *Linee guida* dell'AgID, la riflessione deve oggi concentrarsi soprattutto sui contenuti da rappresentare. Diventa perciò fondamentale focalizzare l'attenzione sui contenuti sia descrittivi (metadati) che iconici (formati di rappresentazione dei dati analogici), e particolarmente sulla loro configurazione sia quantitativa (rapporto tra esaustività dell'informazione/conoscenza da fornire e quantità degli elementi necessari a recuperarla, fruirla e conservarla) che qualitativa (proporzione tra scelta da parte del produttore/gestore del livello informativo/cognitivo da dare a ciascun descrittore e all'insieme dei descrittori e variabili del bisogno informativo/cognitivo degli utenti, a seconda che siano correnti o futuri).

Breve conclusione

Quanto detto in precedenza non ha ragione di esistere, se non si presuppone il seguente assunto di partenza: l'insieme di processi, ricerche, studi, confronti, errori, revisioni, correzioni, evoluzioni, espunzioni e cancellazioni già oggi prodotti deve essere da subito considerato quale potenziale retaggio culturale e, a seconda delle tipologie digitali, opportunamente collocato se identificato quale materiale, o intangibile/immateriale. Se continuiamo a considerare ogni elemento pregresso dell'evoluzione digitale

inutile ed eliminabile in quanto obsoleto e non più funzionale, forniremo alle generazioni future il primo vero caso di incognita evolutiva nella storia dell'umanità.

Riferimenti Bibliografici

Gambetta, V. 2009. *La conservazione della memoria digitale*. [Rubano] : Siav.
Guercio, M. 2013 *Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali*. Roma-Bari : Laterza

Ripensare i formati, ripensare i metadati: prove "tecniche" di conservazione digitale

Nicola Barbuti, Dipartimento di Studi Umanistici (DiSUM) Università degli Studi di Bari Aldo Moro, nicola.barbuti@uniba.it
Mauro Paolo Bruno, Regione Puglia, mp.bruno@regione.puglia.it
Maria Assunta Apollonio, Innova Puglia Spa, m.apollonio@innova.puglia.it
Giuseppe Cultrera, Innova Puglia Spa, g.cultrera@innova.puglia.it
Maria Paola Marzocca, Innova Puglia Spa, m.paolamarzocca@gmail.com

Introduzione

La conservazione digitale di lungo termine, discussa negli aspetti relativi alla questione delicata e ancora irrisolta dei formati di dati digitali in grado di garantirne conservazione, stabilità, sostenibilità, fruibilità e riusabilità nello spazio e nel tempo, è da qualche tempo oggetto del dibattito scientifico e divulgativo sul futuro del digitale e della digitalizzazione.

La riflessione trae origine dalla constatazione che, nell'evoluzione delle società umane, la sopravvivenza e il trasferimento nello spazio/tempo di qualsiasi entità materiale e intangibile/immateriale è stata sempre strettamente vincolata all'identità di retaggio culturale che esse hanno rivestito in quanto testimonianze storiche portatrici di conoscenza per le generazioni future. Tale processo di evoluzione da dato empirico corrente a dato storico/retaggio culturale è da sempre esito di selezione in parte connaturata alle stesse entità, in parte derivate da scelte socio culturali generazionali.

L'evo contemporaneo, invece, pur permeato pervasivamente dalla rivoluzione digitale in atto già da tempo, sembra per nulla preoccupato dalla problematica della conservazione delle entità già prodotte e in fieri, e ancora meno dalla necessità di riconoscere al digitale il rango di facies culturale della contemporaneità.

Pure, si tratta ormai di vera e propria emergenza, se negli ultimi due anni diverse voci autorevoli hanno lanciato l'allarme sul rischio serissimo che l'evo digitale contemporaneo si caratterizzi per essere il primo a non lasciare alcuna traccia recuperabile nella storia evolutiva dell'uomo[1].

Quale approccio per il ripensamento?

Un primo, indispensabile assunto da cui partire per ripensare le politiche di conservazione delle entità digitali risiede nella necessità di evolvere l'approccio alla percezione corrente del digitale e della digitalizzazione, oggi intesi esclusivamente quali strumenti atti a valorizzare i contenuti nel presente e a semplificare la qualità della vita delle comunità: gli stessi contenuti digitali sono considerati rappresentazioni o surroghe dell'analogico, e in quanto tali destinati a esaurire la loro utilità e valorizzazione nella fruizione corrente.

1 Valga qui ricordare l'allarme lanciato da Vinton Cerf nei primi mesi del 2015, http://www.bbc.com/news/science-environment-31450389. Anche l'AgID ha dedicato una corposa sezione alla problematica questione della conservazione digitale nelle ultime linee guida, v. 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).

È evidente come questo approccio condizioni ab origine l'idea stessa di necessità di conservare il digitale. Ma è altrettanto evidente come proprio la realtà quotidiana in cui viviamo e ci muoviamo renda inevitabile e urgente ripensare l'attuale evo digitale in termini di facies culturale. Dal che deriverebbe l'evoluzione di cui sopra, da indirizzare verso un approccio consapevole che non possa più prescindere dal conferire alla cultura digitale contemporanea (digital culture) l'identità di potenziale retaggio culturale digitale per il futuro (digital cultural heritage), in quanto costituita da processi/contenuti digitali cui è connaturata la funzione di potenziali fonti cognitive/informative e memoria storica, portatrici di conoscenza dell'evo contemporaneo per le generazioni future.

Digitalizzare è co-creare cultura: i metadati come storytelling per la conservazione della cultura digitale contemporanea

Le problematiche della conservazione digitale sono diverse e tutte complesse. In questa sede ci si soffermerà sulla complessa questione relativa a quale possa essere la forma e la sostanza da conferire ai dati digitali e alle loro rappresentazioni per garantirne la sopravvivenza nel tempo come testimoni dell'evo digitale contemporaneo, conferendo sia alla struttura che ai contenuti descrittivi la funzione di fonti cui le generazioni future potranno accedere per leggere, studiare, condividere, riusare la memoria storica dell'evo digitale contemporaneo.

Supporto all'esposizione sul tema sarà il caso di studio relativo allo schema di metadati utilizzato nel progetto di digitalizzazione "Archivio storico della Casa Editrice G. Laterza & Figli", intrapreso sul finire del 2015 e in fase di ultimazione e pubblicazione nella "Puglia Digital Library" della Regione Puglia.[2]

Lo schema è stato co-creato avendo quale riferimento base per la struttura gestionale, amministrativa e descrittiva delle risorse digitali lo standard METS-SAN, poi integrato con metadati presi da altri standard basati su ontologie e linguaggi sia semantici che concettuali a completare lo schema definitivo usato nella Puglia DL. Ci soffermeremo in particolare sulla struttura co-creata per lo schema/base utilizzando lo standard METS-SAN.

Nell'analisi preliminare di progetto, si è definito quale obiettivo prioritario prevedere la conservazione sia dei contenuti del progetto Laterza nell'insieme, che di ciascuna parte componente (collezioni, raccolte, etc.) e delle singole risorse digitali.

L'assunto di partenza è stata la consapevolezza che, allo stato dell'arte, le rappresentazioni iconiche contenute in ciascun oggetto digitale prodotto dall'acquisizione ottica presentano la maggiore difficoltà a essere conservate e accessibili nel tempo. Conseguentemente, l'attenzione si è concentrata sui metadati e sul potenziale conservativo che avrebbero, qualora fossero progettati e strutturati tenendo conto non solo delle esigenze di chi li produce, ma soprattutto dei bisogni cognitivi e informativi di chi, un domani, li avrà come fonti principali (se non uniche) alle quali attingere conoscenza.

La scelta, dunque, è stata di considerare i metadati come i supporti per la disseminazione di conoscenza e informazione su genesi e storia dei contenuti progettuali. Passaggio delicato, in quanto ha significato reinterpretarne la normale funzione di strumenti di indicizzazione e recupero di informazioni sul web evolvendola a quella di potenziali fonti, facendo ricorso a un insieme attento e proporzionato di nodi di codice e descrizioni modulate secondo metodi e tecniche dello "storytelling" storico/narrativo.

2 La "Puglia Digital Library" è stata realizzata dalla Regione Puglia in collaborazione con Innova Puglia Spa ed è consultabile all'url: www.pugliadigitallibrary.it. Il progetto di creazione dell'archivio digitale "Archivio storico della Casa Editrice G. Laterza & Figli" rientra nell'ambito delle attività di implementazione della Digital Library, è coordinato dalla stessa Regione e realizzato da Innova Puglia Spa con la collaborazione della spin off dell'Università degli Studi di Bari Aldo Moro D.A.BI.MUS. Srl e il supporto scientifico del Dipartimento di Studi Umanistici (DiSUm) dell'Università di Bari.

Si è quindi immaginato il complesso dei singoli metadati e del loro insieme come un'organica trama testuale composta da elementi formulari (elementi e attributi) e contenuti storico/narrativi (descrizioni), di modo da produrre una manifestazione organizzata, coerente, accessibile, intellegibile e riusabile dei diversi livelli gestionali, amministrativi, giuridici e descrittivi del progetto e, contestualmente, di ciascuna risorsa.

Massima attenzione è stata prestata alla definizione dei parametri quantitativi e qualitativi necessari a mediare contenuti cognitivi/informativi esaustivi di ciascuna componente progettuale, di modo da rileggerli quali fonti destinate a far conoscere una produzione culturale che, quando tra non molti anni le rappresentazioni iconiche che ciascun metadato traduce in storytelling non saranno più fruibili, potrà essere definita dalle future generazioni un'entità culturale testimone del passato digitale, e assunto quale retaggio culturale digitale nel futuro.

Riferimenti Bibliografici

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

Joint Steering Committee for Development of RDA, *Resource Description and Access (RDA)* (http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2015/RDA_Traduzione_ICCU_5_Novembre_REV.pdf)

Stylometric classification and clustering of Spanish texts: some experiments carried out within the TRACE project

Alejandro Bia, Universidad Miguel Hernández – CIO (Spain), abia@umh.es

Abstract

In this paper we will describe the preliminary results of the TRACE project concerning stylometric classification and clustering. We performed several experiments on a corpus of Spanish texts to test classification by time period and author's gender, and also authorship recognition by clustering. Since we found no relevant literature on stylometry applied to Spanish texts, the main purpose of these experiments was to prove that computational stylometry applied to texts in Spanish language works in the same way as when applied to English texts.

These experiments have been performed using the Stylo package, based on the statistical analysis program "R". The tests performed have shown that by making the necessary adjustments and using the correct parameters, the tool is able to perform classification and clustering in a reliable way. We also learned from our own mistakes, particularly concerning text preparation. A proposed workflow for performing stylometric tests in a methodical way is included.

These experiment results will be briefly showcased during the conference presentation.

Acknowledgements: This work has been developed within the TRACE project: Software Tools for Contrastive Analysis of Texts in Parallel Bilingual Corpora, and has been financed with aid FFI2012-39012-C04-02 from the VI National Plan for Scientific Research, Development and Technological Innovation of the MINECO (Ministry of Economy and Competitiveness of Spain).

Introduction

The now popular "meta-field" or "umbrella-field" of Data Science[1], covers a wide range of disciplines, techniques and theories, as it is clearly depicted in Chandrasekaran's Metromap visualization for data science (Chandrasekaran, 2013).

Our work line covers three subfields related to the analysis of texts: information extraction from texts (text mining), computational stylometry and data visualization techniques. These fields, either use similar methods, or cooperate with each other.

Both text mining and computational stylometry use classification and clustering techniques (Jockers and Witten, 2010) (Oakes, 2004), common also to data mining, as well as NLP methods (Nieto, 2004) and in some cases neural networks (Merriam and Matthews, 1994) (Ramyaa et al., 2004). In fact, most of the hypotheses that we want to prove can be tested using clustering and classification methods, by first using training samples and then verifying the approach with test samples, the same as in data mining. Visualization techniques then facilitate a good understanding of the problem and a better analysis of the experimental results.

1 *Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, operations research, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing.* https://en.wikipedia.org/wiki/Data_science

Stylometric classification and clustering

There are several problems that we want to solve using Computational Stylometry, as for instance:

- To identify writers by their writing style
- To separate Basque writers that write both in Basque and in Spanish, from Basque writers that only right is that bilingualism affects a writer's style.
- To classify texts by time period. in the Basque language. The hypothesis here
- To separate translations affected by censorship from uncensored translations.
- To classify texts by the writer's gender
- To classify texts by author (authorship recognition)

In the presentation we're going to show some application examples on a corpus of Spanish texts:

- classification by time period (oppose method, Craigs's Zeta) (see figure 1)
- classification by author's gender (oppose method, Craig's Zeta) (see figure 2)
- author clustering (bootstrap method, consensus tree and cluster analysis)

There are several tools to be used for stylometric analysis. Among the most recent we distinguish the Java Graphical Authorship Attribution Program (JGAAP) (Juola, 2009) and the Stylo R package (Eder et al., 2016). The experiments described here have been performed using the Stylo package, based on the statistical analysis program "R". Stylo provides functions for stylometric analysis, a graphic user interface and print quality diagrams. Stylo is being developed and maintained by the Computational Stylistics group[2].

In some of the experiments mentioned in the list above, like clustering by time period and by author's gender, the method applied was Craig's Zeta (Craig and Kinney, 2009) for a binary classification according to discriminative stylometric features. Craig's Zeta is an extension of the Zeta metric originally proposed by Burrows (Burrows, 2007).

In other experiments, we applied the Bootstrap Consensus Tree (BCT) method and a dendrogram diagram for visualization for authorship recognition. According to Eder (Eder et al., 2016), the BCT method is built on the idea that the results become stable when we divide the list of most frequent words (MFWs) in non-identical, yet potentially overlapping frequency bands and then analyze these bands independently from each other (Eder, 2012). The BCT method was originally borrowed by Eder from the field of Language Evolution and Genetics; since a number of successful applications of the technique have been reported in the literature (Rybicki and Heydel, 2013; van Dalen-Oskam, 2014; Stover et al., 2016). For authorship recognition, we also used the cluster analysis feature of the Stylo package which groups the samples by branches of a hierarchical tree structure.

The corpus used for the experiment consisted of 83 full literary works of 36 different Spanish writers from different centuries (17^{th} to 20^{th}) and gender, including a few false examples to prove the consistency of the methods used. As an example of the latter, we used some Catalan and Portuguese texts to verify they were not taken for old Spanish.

2 https://sites.google.com/site/computationalstylistics/

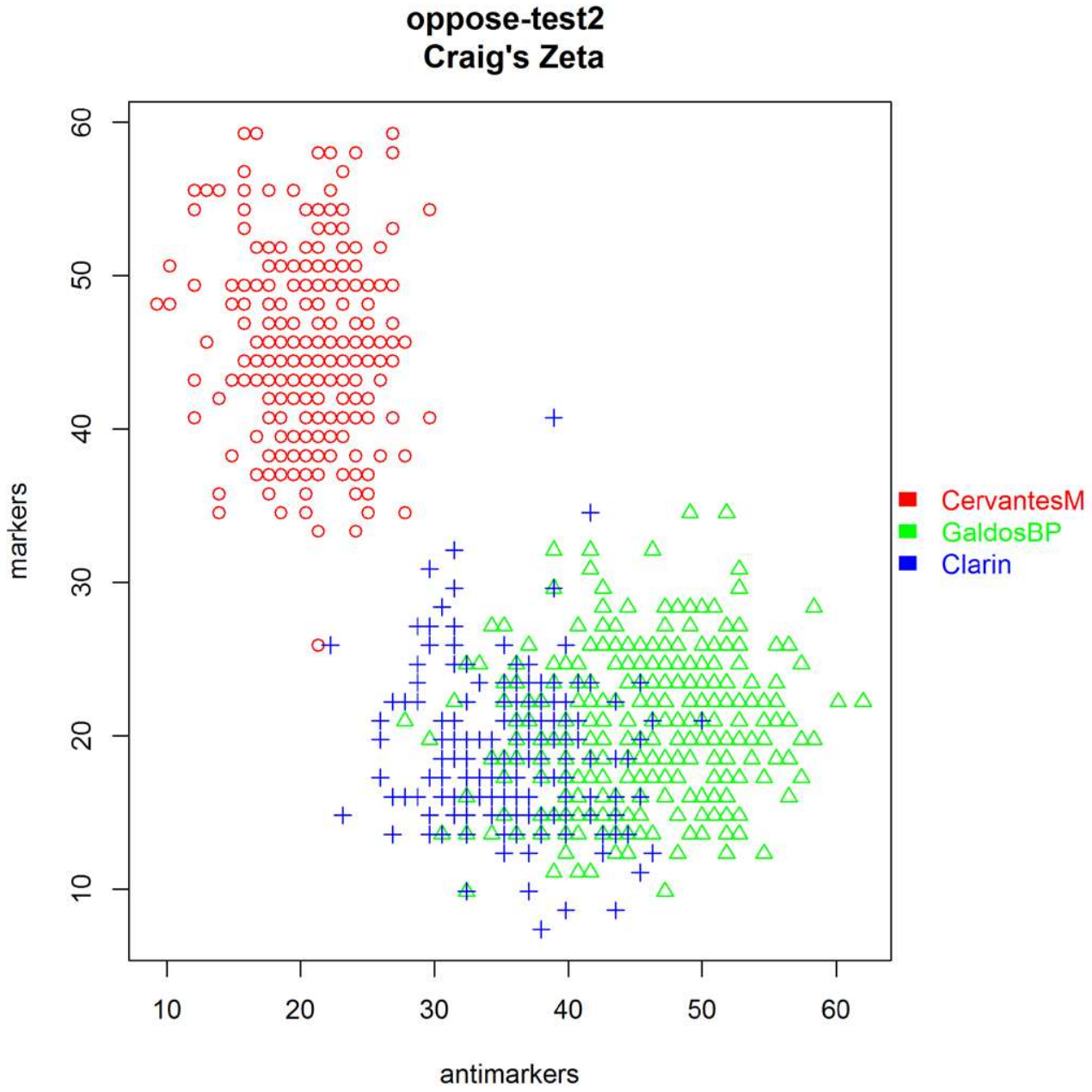


Fig. 1: Oppose-test (Craig): Clustering by time period: Cervantes, Galdos and Clarin.

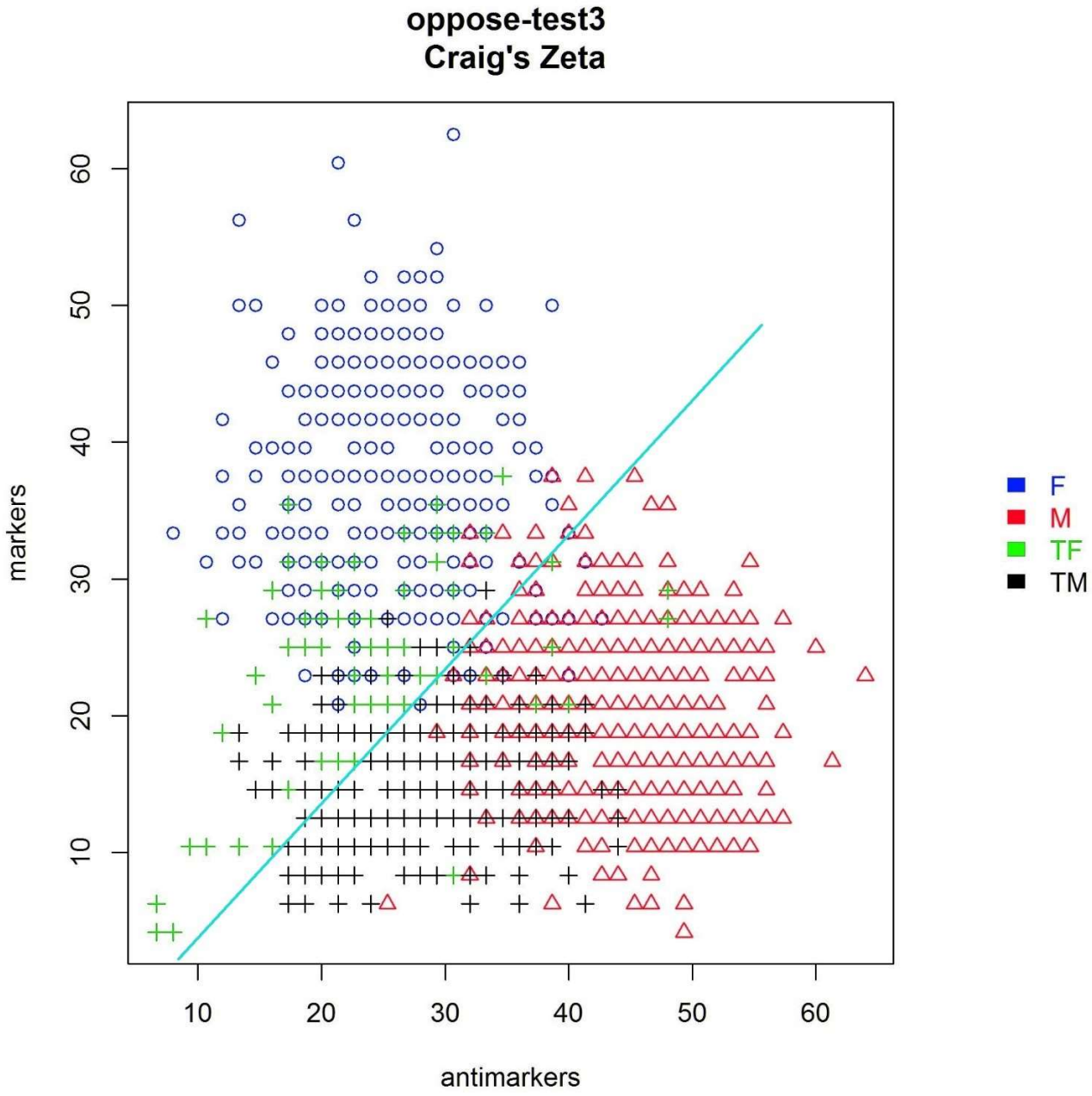


Fig. 2: Oppose-test (Craig's Zeta): Clustering by gender: the test samples are almost perfectly separated by author's gender on each side of the line.

Conclusions

The tests performed with Stylo have shown that by making the necessary adjustments and using the correct parameters, the tool is able to perform time-period classification, gender analysis and authorship recognition in a reliable way on a Spanish corpus of literary works, just as it works when applied to English texts.

From experience, we learned the obvious: that texts must be clean of comments and annotations (when

they are not written by the author), and care must be taken not to use modernized, modified or translated texts, as these were the cases of some faulty classification/clusterization results during our early experiments. In other words, the texts used for stylometry must be pure uncontaminated samples of the writing style of the corresponding authors.

Stylometric tools proved to be very useful, but further human analysis and interpretation of the results is essential to get interesting conclusions. In this, visualization techniques play a very important role. On the other hand, departure hypotheses are desirable in the first place (unless we are looking for unexpected "discoveries"). All this, points to the need of a methodical practice (see figure 3): have a hypothesis or problem to solve, gather and prepare the texts, test/adjust/train the tool, train with a sample collection, test the target research set, visualize and analyse the results, and draw conclusions.

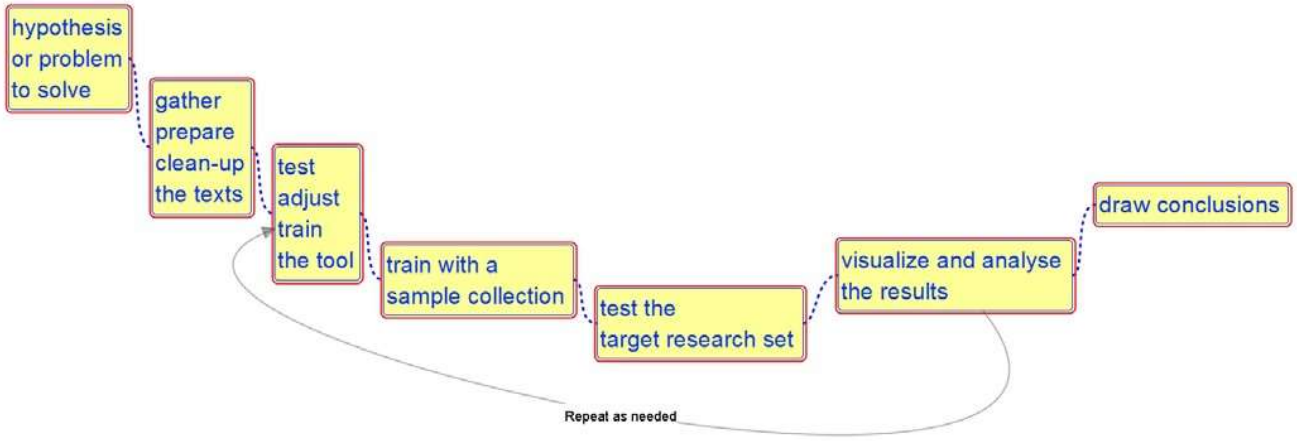


Fig. 3: Proposed Stylometry analysis workflow.

For future work, we intend to explore the use of these tools in computer forensics, to determine authorship and gender of short messages, as shown in the literature (Brocardo et al., 2013) (Calix et al., 2008) (de Vel et al., 2002) (Iqbal et al., 2010).

Although stylometry is not an exact science and depends much on the skills and effort of the researcher choosing the appropriate methods and properly adjusting the test parameters, research on obfuscation corpora has proved that the most robust and accurate methods can be effective, even in cases of deceptive obfuscation (Juola and Vescovi, 2011).

Bibliographic References

Brocardo, M., Traore, I., Saad, S., Woungang, I. 2013. *Authorship Verification for Short Messages Using Stylometry*, Proc. of the IEEE Intl. Conference on Computer, Information and Telecommunication Systems (CITS 2013), Piraeus-Athens, Greece, May 7-8, 2013.

Burrows, J. 2002. *'Delta': A measure of stylistic difference and a guide to likely authorship*. Literary and Linguistic Computing journal, 17(3): 267–287

Burrows, J. 2007. *All the way through: testing for authorship in different frequency strata*. Literary and Linguistic Computing, 22(1): 27–48

Calix, K., Connors, M., Levy, D., Manzar, H., McCabe G., and Westcott S., 2008. *Stylometry for E-*
mail Author Identification and Authentication. In proceedings of CSIS Research Day, Pace University, May 2008.

Chandrasekaran, S., 2013. *Becoming a Data Scientist – Curriculum via Metromap*. 8 Jul 2013. Last seen: 27/11/2016 at http://nirvacana.com/thoughts/becoming-a-data-scientist/

Craig, H. and Kinney A. (eds). 2009. *Shakespeare, Computers, and the Mystery of Authorship.* Cambridge University Press.
Eder, M. 2012. *Computational stylistics and biblical translation: how reliable can a dendrogram be?* In T. Piotrowski and Ł. Grabowski (eds), The Translator and the Computer, 155–170. WSF Press, Wrocław.
de Vel, O., Corney, M., Anderson, A., Mohay, G., 2002, *Language and Gender Author Cohort Analysis of E-mail for Computer Forensics*, In proceedings of The Digital Forensic Research Conference, DFRWS 2002, Syracuse, NY (Aug 6th - 9th), USA
Eder, M., Rybicki, J. and Kestemont, M. 2016. *Stylometry with R: a package for computational text Analysis*. R Journal, 8(1): 107–121. http://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf
Iqbal, F., Khan, L., Fung, B., Debbabi, M., 2010, *E-mail Authorship Verification for Forensic Investigation*, In SAC'10 March 22-26, 2010, Sierre, Switzerland. ACM 978-1-60558-638-0/10/03
Jockers, M. and Witten, D. 2010, *A comparative study of machine learning methods for authorship attribution,* Literary and Linguistic Computing, Vol. 25, No. 2, pp. 215-223.
Juola, P., 2009, *JGAAP: A System for Comparative Evaluation of Authorship Attribution*
JDHCS, Vol. 1, N. 1, Published by: The Division of the Humanities at the University of Chicago, URL: http://jdhcs.uchicago.edu/
Juola, P. and Vescovi, D., 2011, *Analyzing Stylometric Approaches to Author Obfuscation*, In Advances in Digital Forensics VII, G. Peterson and S. Shenoi (Eds.), Chapter 9, IFIP (International Federation for Information Processing) AICT 361, pp. 115–125.
Merriam, T. and Matthews, R., 1994, *Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe,* Literary and Linguistic Computing, Vol. 9, No. 1, Oxfoed University Press.
Nieto V., 2004, *Authorship attribution with help of language engineering,* Homework assignment for Language Engineering, 2D1418, October 16, 2004. URL: https://www.nada.kth.se/kurser/kth/2D1418/uppsatser04/victor.pdf
Ramyaa, He, C., Rasheed, K. , 2004, *Using Machine Learning Techniques for Stylometry.* in Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications (MLMTA'2004), pp. 897-903.
Oakes, M., 2004, *Ant Colony Optimization for Stylometry: The Federalists Papers*, In RASC2004, School of Computing and Technology, University of Sunderland, DGIC, St. Peter's Campus, Sunderland SR6 0DD, United Kingdom.
Rybicki, J. and Heydel, M. 2013. *The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish*. Literary and Linguistic Computing, 28(4):708–717.
Stover, J., Winter, Y., Koppel, M. and Kestemont, M. 2016. *Computational authorship verification method attributes new work to major 2nd century African author*. Journal of the American Society for Information Science and Technology, 67(1):239–242, 2016.
van Dalen-Oskam, K. 2014. *Epistolary voices: The case of Elisabeth Wolff and Agatha Deken*. Literary and Linguistic Computing, 29(3):443–451.

(Big) Data Science e discipline storico-archeologiche: una sfida metodologica, tecnologica e culturale

Claudio Cortese, 4Science S.r.l., claudio.cortese@4science.it

Ci troviamo nella "Big Data Age". Da diversi anni ormai il termine "Big Data" è entrato prepotentemente nel mondo dell'Information Technology, e nell'immaginario collettivo, per le nuove promettenti potenzialità, anche e soprattutto in termini economici, legate all'utilizzo di una nuova generazione di tecnologie e architetture in grado di estrarre valore dall'enorme mole di dati che viene continuamente prodotta negli ambiti più diversi. In questo senso, i più grandi attori del web stanno investendo miliardi di dollari in attività di ricerca e sviluppo di tecnologie per l'analisi di tali masse di dati.

In ambito scientifico i "Big Data" vengono visti come un'opportunità ancora più grande, qualcuno sostiene addirittura epocale: le nuove tecnologie e i moderni strumenti con cui gli studiosi di discipline diverse analizzano i fenomeni di loro interesse permettono, infatti, di produrre e conservare una quantità di dati di diversi ordini di grandezza superiore a quelli disponibili in precedenza (nel 2012 si stimava che 90% dei dati presenti al mondo fosse stato prodotti nel biennio precedente – Silver 2012 -). In questo senso c'è anche chi ha affermato che il cosiddetto "data deluge" (diluvio di dati) renderà obsoleti alcuni dei concetti fondamentali su cui si è finora basato il metodo scientifico (modello, teoria, ipotesi, spiegazione), poiché, in presenza di una tale quantità di dati, i numeri sarebbero ormai in grado di "parlare da soli" (Anderson 2008).

Senza arrivare ad ipotizzare la definizione di un nuovo paradigma scientifico (Wilbanks 2009), sicuramente l'avere a disposizione e l'essere in grado di manipolare e analizzare enormi quantità di dati rappresenta un significativo progresso sia per la scienza sia per altri ambiti di ricerca, non in quanto abolirà la necessità di costruire, raffinare e verificare teorie, ma in quanto permetterà di formulare le ipotesi e di testarle in tempi infinitamente più rapidi e su un campione infinitamente più vasto rispetto a quanto avveniva in passato.

È proprio in quest'ottica che, da qualche anno, il concetto di "Big Data" ha iniziato ad affacciarsi anche tra gli storici e gli archeologi. Se, infatti, è vero che storia e archeologia non sono interessate, al momento, dal fenomeno del "data deluge", esse stanno comunque assistendo ad una crescita dei dati a loro disposizione, grazie alla sempre maggiore diffusione di database, di riviste elettroniche, di digitalizzazioni del patrimonio culturale e di strumenti per l'estrazione e l'analisi dei dati (per una carrellata dei progetti più importanti emersi negli ultimi 10-15 anni si vedano Boonstra, Breure & Doorn 2004; Kristiansen 2014 e Holm 2015). Per la prima volta, dunque, ci si trova di fronte alla possibilità di confrontare migliaia, se non milio ni, di reperti archeologici, libri, opere d'arte, fonti archivistiche, ecc.

In particolare, in questi settori, l'interesse per la problematica dei "Big Data" è legato soprattutto alla possibilità di aggregare, trovare relazioni e analizzare in maniera integrata la molteplicità di dati che sono necessari per rispondere ai quesiti storici. La storia e l'archeologia moderne sono, infatti, spesso caratterizzate da un approccio multidisciplinare e dunque dalla varietà in termini di tipologia, formato, struttura e scala dei dati che, tra l'altro, possono essere patrimonio di istituzioni e gruppi di ricerca differenti. Le ricerche storiche che forse più hanno segnato l'ultimo secolo, del resto, sono proprio quelle che sono riuscite ad utilizzare e integrare tutte le fonti disponibili, secondo i dettami della "lunga durata", della "storia globale" e della "storia comparativa", portati avanti, a partire dai decenni centrali dal secolo scorso, soprattutto da Fernand Braudel (1985, 1993, 2002) e dalla "Scuola delle Annales" (Le Goff 1980; Pomian 1980).

In questo senso, è stato più volte sottolineato il rapporto tra la crescente quantità di dati a disposizio ne e il rinnovato interesse per questo tipo di studi (Van Eijnatten J., Pieters T. & Verheul J. 2013; Kristiansen 2014; Larsson 2014; Guidi & Armitage 2015) in cui l'obiettivo è quello di riuscire a correlare una mole sempre più vasta di fonti diverse, per indagare meglio l'articolazione dei fenomeni storici e dei processi di trasformazione che hanno interessato la storia umana, come, ad esempio, sta già avvenendo con l'integrazione tra i dati paleoclimatici e quelli storici e archeologici (McCormick et al. 2012; Haldon et al. 2014).

Gli strumenti principali per affrontare questa sfida, come peraltro, qualunque progetto che richieda la gestione e l'analisi dei dati (a prescindere dalla loro quantità), sono il Data Management e la Data Science (dalla modellazione dei dati, al text e data mining, ai modelli predittivi, al machine learning, all'analisi del linguaggio naturale, alle simulazioni, all'intelligenza artificiale alla analisi spaziali mediante Geographic Information System, alle tecniche di visualizzazione) che però devono tenere conto della peculiarità e delle caratteristiche dei dati pertinenti agli specifici domini di ricerca (Shennan 1996; Baxter 2003; Boonstra, Breure & Doorn 2004).

In ambito storico-archeologico, infatti, nella maggior parte dei casi, i dati non vengono generati da strumenti ma dagli autori che producono un'opera letteraria o un documento di archivio o dagli studiosi che descrivono un oggetto o un contesto archeologico o, ancora, da chi modella le cosiddette "metafonti" (Genet 1994); quindi, spesso, non sono "neutri" ma possono essere condizionati dalla persona, dal tempo e dal luogo in cui sono stati prodotti. Inoltre le informazioni su cui lavorano la storia e l'archeologia sono frammentarie, parziali e, in molti casi, distorte. Infine il "digita le " rappresenta solo una piccola parte del patrimonio, delle fonti e dei documenti su cui gli studiosi basano le loro interpretazioni e, anche il giorno in cui la maggioranza del patrimonio culturale sarà digitalizzata, essa sarà, comunque, spesso, solo un surrogato che non necessariamente potrà sostituire la visione diretta della fonte informativa originaria (sia essa una pentola in ceramica, un manoscrit to o un dipinto).

Nell'analisi dei dati digitali è fondamentale, quindi, che essi non vengano considerati isolatamente ma congiuntamente a tutte le informazioni contestuali, digitali e non, necessarie a rispondere alle domande della ricerca. È necessario quindi, ad esempio, indagare e analizzare anche il contesto in cui è stato prodotto un documento/monumento (Foucault 1971; Le Goff 1978, 1980), ricostruire e tener conto dei processi formativi che hanno caratterizzato il deposito archeologico da cui i dati provengono (Schiffer 1996) o analizzare le associazioni contestuali che caratterizzano i documenti, i monumenti e i reperti, a diversi livelli e su scale diverse (Le Goff 1978; Hodder 1992), in un processo interpretativo che deve dare il giusto peso e cercare di spiegare anche le assenze, le lacune o i "silenzi " della storia, insomma i dati che non ci sono. È già stato sottolineato, del resto, con particolare riferimento all'ambito epigrafico (Lamé 2015), come proprio un approccio umanistico- informatico favorisca lo studio delle fonti primarie come "dispositivi", nel senso in cui il termine è utilizzato da Michel Foucault, in altre parole come testimonianze di una rete di sistemi eterogenei (sociale, economico, culturale, tecnologico, ecc.) le cui relazioni possono essere indagate attraverso di esse, mediante un'analisi globale e multidimensionale.

Tra l'altro, la gestione del contesto nell'analisi dei Big Data è considerata, anche al di fuori dell'ambito umanistico, una delle sfide fondamentali per arrivare a utilizzarli in maniera efficace, sfruttandone appieno il potenziale (Boyd & Crawford 2012). Importanza decisiva acquisisce dunque il background culturale e la capacità critica dello studioso, passo fondamentale contro il pericolo di una "decontesutalizzazione" dei dati.

In quest'ottica, se un approccio "(Big) Data Driven" può essere auspicabile in ambito storico e archeologico, è necessario essere molto accorti dal punto di vista metodologico per evitare di cadere in modo acritico nel "feticismo del numero". È quindi necessario che i data analyst e i data scientist, oggi, e in un futuro in cui auspicabilmente i Big Data diventeranno sempre più centrali anche nelle humanities, abbiano elevate competenze non solo di tipo informatico e statistico ma anche di dominio. Solo un esperto di dominio, infatti, è in grado di dire, ad esempio, se una correlazione statisticamente

significativa è rilevante anche dal punto di vista storico o archeologico. Già alcuni decenni fa, trattando l'emergere della "storia quantitativa" ed evidenziando come ormai il dato (e non il fatto) costituisse l'unità di informazione fondamentale per lo storico, Jacques Le Goff (1980, p. 37) sottolineava come "la parte essenziale del lavoro storico resta, comunque, ancora da fare quando il calcolatore ha sfornato i suoi risultati" (in questo senso si vedano anche Gattiglia 2015 e Holm 2015).

C'è e ci sarà, quindi, sempre più bisogno di storici e archeologi specializzati nella gestione, nell'integrazione e nell'analisi dei dati, discipline che, a livello universitario, non dovrebbero essere più relegate a laboratori di poche ore o, eventualmente, alla formazione post-laurea magistrale, ma dovrebbero costituire una parte essenziale del bagaglio culturale di uno studente sin dal primo triennio universitario. Del resto i tempi dovrebbero essere ormai maturi perché l'informatica applicata alle discipline umanistiche, smetta di essere vista come una "disciplina ausiliaria", che si può scegliere o meno di utilizzare, ma inizi ad essere considerata per quello che è, ovvero un fondamentale e irrinunciabile strumento di educazione alla modellazione e formalizzazione del ragionamento storico e archeologico, in grado di avere una ricaduta sulla metodologia di tali discipline in senso lato, e, tramite la creazione di "metafonti", di valorizzare, anche mediante l'uso di adeguate tecniche quantitative, la rete di legami contestuali fondamentali per la comprensione dei dati (si vedano in questo senso, ad esempio, Gardin 1995 e 1996; Lamé 2015; Gattiglia 2015; Cortese 2016). Se le humanities potranno in futuro trarre benefici dalle possibilità offerte dai Big Data è una questione legata, dunque, anche a come l'università saprà adattare i suoi programmi a queste esigenze sempre più urgenti.

Un ulteriore passo necessario a favorire l'uso e, soprattutto, la diffusione dei metodi della (Big) Data Science e di una modalità di ricerca scientifica "Data Driven" nell'ambito delle discipline in esame è la disponibilità di infrastrutture e strumenti per l'integrazione, la condivisione, l'analisi e la conservazione dei dataset, e che, nello stesso tempo, rendano possibile interagire con fonti dati esterne.

In questo senso, i VRE (Virtual Research Environment) potrebbero costituire la risposta alle necessità di un "ambiente" scalabile e sostenibile, finalizzato a gestire l'intero ciclo di vita del dato in modo collettivo. Si tratta di piattaforme che, da certi punti di vista, possono essere considerate un'evoluzione dei repository o delle Digital Library, in grado di mettere a disposizione degli studiosi strumenti di condivisione e software di analisi in un ambiente integrato, all'interno del quale i dati possono essere processati con una velocità molto maggiore rispetto a quanto può avvenire con un normale PC. Solo in Europa esistono ormai decine di progetti finalizzati a creare infrastrutture di questo tipo, alcune funzionali a gestire la ricerca relativa ad un singolo dominio, altre invece con scopi più generali (in particolare, per le Digital Humanities, si veda Blanke et al. 2010), che però, sono ancora utilizzate per lo più in ambito sperimentale e da gruppi ristretti di ricercatori, spesso provenienti dalle istituzioni che hanno partecipato allo sviluppo delle stesse. Il prossimo obiettivo dovrebbe essere, dunque, quello di riuscire a inserire gli strumenti tecnologici creati nell'ambito di questi progetti nella pratica quotidiana della ricerca. Ciò sarebbe tanto più utile in contesti come quelli umanistici dove è rara la disponibilità di avanzate infrastrutture tecnologiche e dove, nella maggior parte dei casi, i dati risiedono sui PC personali dei ricercatori. Sulla base di quanto detto in precedenza, tali sistemi dovranno permettere agli studiosi di analizzare i dati, evidenziandone e valorizzandone le relazioni a diversi livelli e di esplicitare le loro interpretazioni rispetto alle dimensioni di variabilità significative e alla rete di legami contestuali che interessano le fonti storico - archeologiche. Dovranno dunque avere nella flessibilità del modello dei dati, oltre che negli strumenti di integrazione e di analisi, una caratteristica fondamentale.

La storia, l'archeologia, e, probabilmente, gli studi umanistici in generale si trovano, quindi, di fronte a una sfida metodologica (adattare al meglio i metodi della Data Science alle peculiarità dei propri dati, facendo tesoro degli strumenti teorici e metodologici messi a punto nel corso della storia degli studi), tecnologica (contribuire alla realizzazione di strumenti che rendano tali metodi più facilmente ed efficacemente utilizzabili da parte della comunità scientifica) e culturale (prendere coscienza del fatto che la Data Science deve entrare a pieno diritto nel percorso formativo dei giovani

che si avvicinano a queste discipline). Solo se sarà in grado di vincere questa sfida, valorizzando dunque i portati del "patrimonio genetico" delle scienze storiche all'interno di un approccio di tipo nuovo in grado di integrare il "tradizionale" lavoro ermeneutico e interpretativo dello storico e dell'archeo logo e le più efficaci tecniche di gestione e analisi dei dati, la comunità degli studiosi nel suo complesso potrà trarre tutti i benefici insiti nella sempre crescente quantità di dati disponibili, che altrimenti rimarranno patrimonio, non condiviso, e forse nemmeno riconosciuto, solo dei gruppi più o meno ristretti che si occupano di "Digital Humanities".

Riferimenti bibliografici

Anderson C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired Magazine*, June 23. Accessed November 7, 2016, https://www.wired.com/2008/06/pb-theory/

Baxter M. 2003. *Statistics in Archaeology*. London: Arnold.

Blanke T., Candela L., Hedges M., Priddy, M. & Simeoni F. 2010. "Deploying general-purpose virtual research environments for humanities research". *Philosophical Transactions of the Royal Society A* 368: 3813-3828.

Boonstra O., Beure L. & Doorn P. 2004. *Past, present and future of historical information science*. Amsterdam: DANS.

Boyd, D. & Crawford, K. 2012. "Critical Questions for Big Data". *Information, Communication and Society* 15:5: 662–679.

Braudel F. 1985. *Una lezione di storia*. Torino: Einaudi.

Braudel F. 1993. *Civiltà materiale, economia e capitalismo. Le strutture del quotidiano (secoli XV- XVIII)*. Torino: Einaudi.

Braudel F. 2002. *Civiltà e imperi del Mediterraneo nell'età di Filippo II*. Torino: Einaudi.

Cortese C. 2016. "Informatica applicata all'archeologia: 'disciplina ausiliaria' o fondamenta le strumento di educazione alla modellazione e formalizzazione del ragionamento archeologico?". In *Archeologia classica e post-classica tra Italia e Mediterraneo. Scritti in ricordo di Maria Pia Rossignani*, a cura di S. Lusuardi Siena, C. Perassi, F. Sacchi & M. Sannazaro, 571-576. Milano: Vita e Pensiero.

Foucault M. 1971. *Archeologia del sapere*. Milano: Rizzoli.

Gardin J. C. 1995. "L'analisi delle costruzioni scientifiche". In *L'architettura dei testi storiografici. Un'ipotesi*, a cura di J. C. Gardin & M. N. Borghetti, 25-110. Bologna: CLUEB.

Gardin J. C. 1996. "Cognitive issues in archaeology". *Archaeologia Polona* 34: 205-232.

Gattiglia G. 2015. "Think big about data: Archaeology and the Big Data challenge". *Archäologische Informationen* 38: 113-124.

Genet J. Ph. 1994. "Source, Métasource, Texte, Histoire". In *Storia & multimedia*, a cura di F. Bocchi & P. Denley, 3-17. Bologna: Grafis.

Guidi J. & Armitage D. 2015. *The History Manifesto*. Cambridge: Cambridge University Press.

Haldon J., Roberts N., Izdebsky A., Fleitman D., McCormick M., Cassis M., Doona O., Eastwood W., Elton H., Ladstatter S., Manning S., Newhard J., Nicoll K., Telelis I. & Xoplaki E. 2014. "The Climate and Environment of Byzantine Anatolia: Integrating Science, History and Archaeology". *Journal of Interdisciplinary History* XLV:2: 113-161.

Hodder I. 1992. *Leggere il passato. Tendenze attuali dell'archeologia*. Torino: Einaudi.

Holm P. 2015. "Climate Change, Big Data and the Medieval and Early Modern". In *Medieval or Early Modern. The Value of a Traditional Historical Division*, edited by R. Hutton, 70-85. Newcastle: Cambridge Scholars Publishing.

Kristiansen K. 2014. "Towards a new paradigm? The Third Science Revolution and its Possible Consequences in Archaeology". *Current Swedish Archaeology* 22: 11-34.

Lamé M. 2015. “Primary Sources of Information, Digitization Processes and Dispositive Analysis”. In *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem*, edited by F. Tomasi, R. Roselli Del Turco, A. M. Tammaro: ACM.
Larsson F. 2014. “The third science revolution and its possible consequences in archaeology. A personal reflection”. *Current Swedish Archaeology* 22: 53-56.
Le Goff J. 1978. “Documento/Monumento”. In *Enciclopedia*, 38-43. Torino: Einaudi.
Le Goff J. 1980. “La nuova storia”. In *La nuova storia*, a cura di J. Le Goff , 7-46. Milano: Arnoldo Mondadori Editore.
McCormick M., Buntgen U., Cane M. A., Cook E. R., Harper K., Huybers P., Litt T., Manning S. W., Mayewsky, More A.M.F., Nicolussi K. & Tegel W. 2012. “Climate Change during and after the Roman Empire: Reconstructing the Past from Scientific and Historical Evidence”. *Journal of Interdisciplinary History* XLIII:2: 169-220.
Pomian K. 1980. “Storia delle strutture”. In *La nuova storia*, a cura di J. Le Goff , 81-110. Milano : Arnoldo Mondadori Editore.
Schiffer M.B. 1996. *Formation Processes of the Archaeological Record*. Salt Lake City: Univers ity of Utah Press.
Shennan S. 1996. *Quantifying Archaeology*. Edinburgh: University
Press. Silver N. 2012. *The Signal and the Noise*. New York. The Penguin
Press.
Van Eijnatten J., Pieters T. & Verheul J. 2013. “Big Data for Global History. The transformat ive Promise of Digital Humanities”. *Low Countries Historical Review*: 128-4: 55-77.
Wilbanks J. 2009. “I Have Seen the Paradigm Shift and It Is Us”. In *The Fourth Paradigm. Data- Intensive Scientific Discovery*, edited by T. Hey, E. Tansley & K. Tolle, 209-214. Redmond, Washington: Microsoft Research.

From digitization to datafication. A new challenge is approaching archaeology

Gabriele Gattiglia, MAPPA Lab – Università di Pisa, gabriele.gattiglia@for.unipi.it

Data, Big Data

Data are what economists call a non-rivalrous good, in other words, they can be processed again and again and their value does not diminish (Samuelson, 1954). On the contrary, their value arises from what they reveal in aggregate. On the one hand, the constant enhancement of digital applications for producing, storing and manipulating data has brought the focus onto data-driven and data-led science (Royal Society, 2012, 7), even in the Humanities, on the other hand, in recent decades, archaeology has embraced digitisation. Moreover, the low cost and improvement in computing power (both software and hardware) gives the opportunity to easily aggregate huge amounts of data coming from different sources at high velocity: in brief we are in a Big Data era. Even if Big Data started in the world of Computer Science and are strongly connected to business, they are rapidly emerging in academic research, with scholars from different disciplines recognising the inherent research potential of analysing composite and heterogeneous datasets that dwarf in size and complexity those traditionally employed in their respective fields (Wesson and Cottier 2014; Gattiglia 2015). In recent years, archaeologists began to ask to themselves if a Big Data approach can be applied to archaeology from both a theoretical and practical point of view (Gattiglia 2015). In the scientific and scholarly world what constitutes Big Data varies significantly between disciplines, but we can certainly affirm that the shift in scale of data volume is evident in most disciplines, and that analysing large amounts of data holds the potential to revolutionise research, even in the Humanities, producing hitherto impossible and unimaginable insights (Wesson and Cottier 2014, 1). For a better understanding of the general concept of Big Data, I adopt the definition proposed by (Boyd and Crawford 2012, 663): "Big Data is less about data that is big than it is about a capacity to search, aggregate, and crossreference large data sets". In other words, Big Data's high volume, high velocity, and high variety do not have to be considered in an absolute manner, but in a relative way. As suggested by (Mayer-Schönberger and Cukier 2013), using Big Data means working with the full (or close to the full) set of data, namely with all the data available from different disciplines that can be useful to solve a question (Big Data as All Data). This kind of approach permits to gain more choices for exploring data from diverse angles or for looking closer at certain features of them, and to comprehend aspects that we cannot understand using smaller amounts of data. Moreover, Big Data is about predictive modelling, i.e. about applying algorithms to huge quantities of data in order to infer probabilities, and it is about recognising the relationships within and among pieces of information. Moreover, a Big Data approach is related to the information content of data. Data are useful because they carry pieces of information. As Clark's DIKW (Data Information Knowledge Wisdom) hierarchy (Clark 2004) and Hey's Knowledge Pyramid pointed out (Hey 2004), data are the building blocks of meaning, they are meaningless except for their relationship to other data. Data become information when they are processed and aggregated with other data, thereby we gain information from data when we make sense out of them (Anichini and Gattiglia 2015). Finally, we can say that data are data because they describe a phenomenon in a quantified format so it can be tabulated and analysed, not because they are digital.

Datafication

Digitisation has changed archaeology deeply. Digitisation usually refers to the migration of pieces of information into digital formats, for transmission, re-use and manipulation. Surely, this process has increased exponentially the amount of data that could be processed, but from a more general point of view the act of digitisation, i.e. turning analogue information into computer readable format, does not by itself involve datafication. Datafication is a new phenomenon brought out by the continuous development of IT technologies. Datafication promises to go significantly beyond digitisation, and to have an even more profound impact on archaeology, challenging the foundations of our established methods of measurement and providing new opportunities. Datafication is the act of transforming something into a quantified format (Mayer-Schönberger and Cukier 2013, 73; O'Neil and Schutt 2013, 406). This is a key issue. As argued by (Cresswell 2014, 57) "two things that are making data suddenly big are the datafication of the individual and the geocoding of everything". To datafy means to transform objects, processes, etc. in a quantified format so they can be tabulated and analysed (Mayer-Schönberger and Cukier 2013). We can argue that datafication puts more emphasis on the I (information) of IT, dis-embedding the knowledge associated with physical objects by decoupling them from the data associated with them (Gattiglia 2015). Datafication is manifest in a variety of forms and can also, but not always, be associated with sensors/actuators and with the Internet of Things (Bahga and Madisetti 2014, 37). Moreover, a key differentiating aspect between digitisation and datafication is the one related to data analytics: digitisation uses data analytics based on traditional sampling mechanisms, while datafication fits a Big Data approach and relies on the new forms of quantification and associated data mining techniques, that permit more sophisticated mathematical analyses to identify non-linear relationships among data, allowing us to use the information, for instance, for massive predictive analyses. In other words, to datafy archaeology would mean to produce a flow of data starting from the data produced by the archaeological practice, for instance, locations, interactions and relations between finds and sites. A flow of data that the archaeological community should have available.

ArchAIDE project

Introduction

The ArchAIDE project goes exactly in this direction.

ArchAIDE is a three-year (2016-2019) RIA project, approved by EC under call H2020-REFLECTIVE-6-2015. The project consortium is coordinated by the University of Pisa with the MAPPA Lab, a research unit of the Department of Civilisations and Form of Knowledge, and includes a solid set of Human Sciences partners (University of Barcelona, University of Cologne and University of York), some key players in ICT design and development (CNR-ISTI and Tel Aviv University), two archaeological companies (BARAKA and ELEMENTS) and one ICT company.

The work of the project includes the design, development and assessment of a new software platform offering applications, tools and services for digital archaeology. This framework, that will be available through both a mobile application and a desktop version, will be able to support archaeologists in recognising and classifying pottery sherds during excavation and post-excavation analysis.

The system will be designed to provide very easy-to-use interfaces (e.g. touch-based definition of the potsherd profile from a photograph acquired with the mobile device) and will support

efficient and powerful algorithms for characterisation, search and retrieval of the possible visual/geometrical correspondences over a complex database built from the data provided by classical 2D printed repositories and images. Our approach is driven by archaeologists needs; since we are aware of the caution of the discipline in front of the replacement of well-established methods, we plan to support this specific Humanities domain by exploiting what is already available in the Archaeology domain in terms of good practices and representation paradigms. We thus plan to deliver efficient computer-supported tools for drafting the profile of each sherd and to automatically match it with the huge archives provided by available classifications (currently encoded only in drawings and written descriptions contained in books and publications). The system will also be able to support the production of archaeological documentation, including data on localisation provided by the mobile device (GPS). The platform will also allow to access tools and services able to enhance the analysis of archaeological resources, such as the open data publication of the pottery classification, or the data analysis and data visualisation of spatial distribution of a certain pottery typology, leading to a deeper interpretations of the past. The integration of cultural heritage information from different sources, together with faster description, cataloguing and improved accessibility can be exploited to generate new knowledge around archaeological heritage. Data visualisation, for instance, would stimulate new research perspectives, and could enable new interpretation and understanding of history, and would bring archaeological storytelling to new audiences in a novel way. By means of a wider dissemination of user-generated content, the framework would permit to develop the culture of sharing cultural resources, research and knowledge.

From digitisation to datafication

The first contribution of ArchAIDE (www.archaide.eu) is an as-automatic-as-possible procedure to transform the paper catalogues in a digital description, to be used as a data pool for an accurate search and retrieval process. This will entail: scanning (2D digitization) of the paper catalogue(s); segmentation and vectorialization of the graphical drawings proposed in those printed catalogues; and linking the graphical representation with the metadata reported in the catalogues. Since we are interested in designing automatic matching and retrieval features, digital description does not mean here only digitisation of the paper catalogues, but includes understanding the meaning of the graphic representation and its conversion in a format that includes shape (in vectorial format, not raster) and semantic. This process, naturally, will also require the definition of a semantically-rich digital vectorial representation for the pottery sherds and of each entire object able to represent not only the shape of the object, but also its subdivision in semantic components (e.g. rim, handle, foot, …). This representation, ideally, should be compliant with the existing representation, description and drawing standards used by archaeologists, to help both the digitisation phase (from “classical” documentation to digital) and the creation of the documentation (from digital back to “classical” documentation). A lightweight set of metadata (the subset considered crucial for the purposes of the project by our users and advisors, e.g. historical period, geographical region...) will be added to the extracted data. On the other hand, the data collected through digitisation will be enriched by data collected by users during the recognition process. This will permit on-time data analysis and data visualisation. In fact, all the information encoded in the pottery identity cards (being them natively digital and including data on location, classification, dating, and so on) will be shared, visualised and integrated with cultural heritage information from different sources (archaeological repositories, Europeana and so on) in order to produce a really significant impact in the advancement of the discipline and in the accessibility for professional and non-professional users. Real time comparisons between different archaeological sites and regions will be made possible, thus highlighting differences and commonalities in the economy of the ancient world. A web-based

visualization tool will improve accessibility to archaeological heritage and generate new understanding about the dynamics of pottery production, trade flows, and social interactions.

Data analysis will be carried on by the MAPPA Lab of the University of Pisa, and will be achieved as an exploratory statistical analysis of data related to pottery. It will be mainly concerned with data about size, density, geo-localisation and chronology. The main objective of the exploratory analysis is to disclose statistical relationships (in statistical sense) between the different variables considered. Moreover, it will provide a comprehensive description of the available data, pointing out important features of the datasets, such as: where the information concentrates and where is missing, or where little data more would imply a relevant gain of information. There are different statistical techniques useful for exploratory data analysis, each one concentrating on particular aspects of the description we would like to give for the data. However, it is important to observe that the statistical techniques are not exploratory as such, rather they are used in order to summarize main characteristics of data, identify outliers, trends, or patterns, i.e. they are used as explorative.

Concerning the analysis of pottery datasets, we will concentrate on the following tools:

- Classification and Clustering techniques, to be used for understanding whether or not some features of the data may possess convenient classifications in a number of categories/groups, subsequently suggesting meaningful interpretation of such categories;
- Dimensionality reduction techniques, to be used in order to extract a small number of specific combination of features describing the greatest part of information and variability contained within the data. These specific combinations provide all at once a way to summarize data, and the identification of the major sources of variability;
- Spatial statistics, point pattern analysis and Kriging methods will be mainly used in order to highlight the possible patterns within the spatial distribution of data;
- Different predictive modelling techniques will be implemented mostly for suggesting where to look for more data in order to get relevant gain of information, or optimal strategies to perform testing.

The results of the data analysis will be made more understandable and easily explicable applying data visualisation techniques. Apart from the quantitative data analysis, data visualization is of extreme importance, in order to: provide an efficient way to understand a vast amount of data; allow non-technical people to do data-driven decision making; communicating the results of the data analysis (Llobera 2011; Gattiglia 2015). An important issue is the communicating the visual information about the relationships among different ceramic classes in the same location, the relationships between the location of the finding and the productive centre, and the relationships with pottery found in different locations. A web-based visualisation tools will be built following the principles of data visualization, pionereed by (Bertin 2010, 83), and developed for instance in (Tufte 1990; Few 2006; Munzner 2014). Following these guidelines, we will classify the different data into types (categorial, ordinal, interval, ratio types), and will determine which visual attributes (shape, orientation, colors, texture, size, position, length, area, volume) represent data types most effectively, so giving rise to the visualization, according to the basic principle of assigning most efficient attributes, such as position, length, slope, to the more quantitative data types, and less efficient attributes, like area, volume, or colors to ordinal or categorical types. The process of building the visualisation will be made interactive, letting the user associating the different variables with the different attributes, at the same time explaining the principles above. Moreover, the different relations within pottery production, trade flows, and social interactions, will be visualised applying the same principles, with graphs.

The possibilities of such system open to research actors, institutions and general public would be a dramatic change in the archaeological discipline as it is nowadays. Its impact on the field would dramatically change the profile of the professionals involved and will generate new markets.

Bibliographic References

Anichini, Francesca and Gattiglia, Gabriele 2015. **"**Verso la rivoluzione. Dall'Open Access all'Open Data: la pubblicazione aperta in archeologia." *Post – Classical Archaeologies* 5: 299-326.

Bahga, Arshdeep and Madisetti, Vija 2014. *Internet of Things (a Hands-On Approach)*. Arshdeep Bahga Vijay Madisetti

Bertin, Jacques 2010. *Semiology of Graphics*. Esripress

Boyd, Danah and Crawford, Kate 2012. "Critical Questions for Big Data. Information." *Communication and Society* 15: 662–679

Clark, Donald 2004. "Understanding and Performance." Accessed November 15. http://www.nwlink.com/~donclark/performance/understanding.html%20.

Cresswell, Tim 2014. "Déjà vu all over again: Spatial Science, quantitative revolutions and the culture of numbers." *Dialogues in Human Geography* 4 (1): 54-58.

Few, Stephen 2006. *Information Dashboard design; The Effective Visual Communication of Data*. Sebastopol, CA: O'Reilly Media.

Gattiglia, Gabriele 2015. "Think big about data: Archaeology and the Big Data challenge." *Archäologische Informationen* 38: 113-124.

Hey, Jonathan 2004. "The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link." Accessed November 15. http://inls151f14.web.unc.edu/files/2014/08/hey2004-DIKWchain.pdf

Llobera, Marcos 2011. "Archaeological Visualization: Towards an Archaeological Information Science (AISc)." *Journal of Archaeological Method and Theory* 18: 193–223.

Mayer-Schönberger, Viktor and Cukier, Kenneth 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Munzner, Tamara 2014. *Visualization Analysis and design*. Boca Raton, FL: CRC Press.

O'Neil, Cathy and Schutt, Rachel 2013. *Doing Data Science*. Sebastopol, CA: O'Reilly Media.

Royal Society 2012. *Science as an Open Enterprise*. London, England: Royal Society.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics* 36(4): 387-389.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, CT: Edward Tufte.

Wesson, Cameron B. and Cottier, John W. 2014. "Big Sites, Big Questions, Big Data, Big Problems: Scales of Investigation and Changing Perceptions of Archaeological Practice in the Southeastern United States". *Bulletin of the History of Archaeology* 24 (16): 1-11.

AIUCD e CLiC-it: citazioni bibliografiche a confronto

Gabriella Pardelli, CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli", Italy
gabriella.pardelli@ilc.cnr.it
Silvia Giannini, CNR, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Italy
silvia.giannini@isti.cnr.it
Federico Boschetti, CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli", Italy
federico.boschetti@ilc.cnr.it
Riccardo Del Gratta, CNR, Istituto di Linguistica Computazionale, "Antonio Zampolli", Italy
riccardo.delgratta@ilc.cnr.it

Introduzione

Il lavoro propone l'analisi e il confronto dei riferimenti bibliografici delle cinque edizioni annuali della Conferenza dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) e del primo biennio (2014-2015) della Conferenza Italiana di Linguistica Computazionale (CLiC-it) per misurare la direzione in cui si muove il trend citazionale. La giovane conferenza italiana di Linguistica Computazionale CLiC-it nasce nel 2014 a Pisa come «un nuovo evento finalizzato alla creazione di un forum di riferimento per la discussione delle ricerche della comunità italiana di Linguistica Computazionale[…] CLiC-it vuole riunire ricercatori da discipline intimamente connesse quali la Linguistica Computazionale, la Linguistica, le Scienze Cognitive, l'Apprendimento Automatico, l'Informatica, la Rappresentazione della Conoscenza, l'Information Retrieval e le Digital Humanities» (http://www.fileli.unipi.it/projects/clic). E a proposito di Digital Humanities, nel 2012 prende vita la conferenza italiana di Informatica Umanistica e la Cultura Digitale AIUCD. Nei cinque anni della sua breve storia l' AIUCD ha avuto come temi principali: l'informatica umanistica e la cultura digitale; la ricerca collaborativa e le piattaforme condivise per l'Informatica umanistica; la metodologia della ricerca umanistica nell'ecosistema digitale; la relazione tra Digital Humanities e Beni Culturali (http://www.umanisticadigitale.it).

Lo studio delle bibliografie permette di applicare metodologie di *distant reading* per osservare somiglianze e differenze relative ai contributi presentati alle conferenze in oggetto. La lettura della bibliografia di un articolo scientifico rivela sin da subito le fonti utilizzate dall'autore nella sua indagine. Idee e opinioni vengono legittimate e consolidate e metodologie di ricerca diverse si confrontano e si integrano su tematiche comuni. Le citazioni diventano così un particolare tipo di *"etichetta"* di riconoscimento del lavoro svolto da un gruppo di ricerca o da un singolo ricercatore attivando una sorta di «… *conversazione fra il passato e il presente…*» (Venuda 2012, 10). La citazione bibliografica nel corso del '900 ha assunto un ruolo strategico nei meccanismi di comunicazione scientifica e di valutazione della ricerca, dovuto alla realizzazione dei servizi di indicizzazione citazionale. Le pubblicazioni «*…sono diventate le banconote della scienza…*» (Di Donato 2010) nonché «*…moneta corrente nel commercio della comunicazione scientifica ufficiale. Moneta di piccolo taglio (costa poco citare), ma dal potere d'acquisto simbolico non indifferente…*» (De Bellis 2005, 9). Gli anni '90 del Novecento, inoltre, hanno fatto da cornice all'affermazione di un concetto altrettanto antico e universale come quello della conoscenza aperta.

La rivoluzione tecnologica e la diffusione del mezzo digitale sono state le due condizioni fondamentali per la realizzazione di questo concetto e la formalizzazione del movimento Open Access (OA). Unitamente all'affermazione dell'OA, il sistema prototipale degli e*print archives*

anni ’90 si è evoluto nella forma degli attuali repositories, divenuti importanti infrastrutture per la raccolta e l’accesso all’informazione.

L’analisi muove dunque dal principio di rilevanza della citazione nella trasmissione della conoscenza in un periodo di grandi cambiamenti socioculturali e di importanti evoluzioni nelle modalità di produzione e diffusione dei risultati della ricerca scientifica attraverso i nuovi canali.

La misura delle risorse citate dagli autori consente di valutare eventuali trasformazioni rispetto alla citazione tradizionale.

Digital Humanities and Computational Linguistics è il titolo di uno special issue di prossima uscita (dicembre 2016) della rivista “Italian Journal of Computational Linguistics IJCol”. In questo contesto la comparazione delle “abitudini citazionali” consente di recuperare analogie e difformità relative alle fonti bibliografiche di due ambiti apparentemente distinti ma storicamente accomunati e interdisciplinari come quelli di AIUCD e CLiC-it.

Materiali e metodo

Le informazioni sono state estratte dai riferimenti bibliografici degli articoli pubblicati negli atti o nei “Book of Abstracts” delle due conferenze. Questa raccolta informativa è andata a costituire il nostro *corpus citazionale* comprendente il materiale necessario per l’elaborazione.

Il lavoro di analisi è suddivisibile in 4 fasi:

1. creazione del corpus
2. classificazione dei dati raccolti in segmenti informativi
3. accorpamento delle tipologie documentarie in macrocategorie
4. analisi statistico/comparativa dei dati

Lo studio condotto sul corpus ha consentito di osservare le caratteristiche e le variazioni nel corso degli anni. E’ stato necessario suddividere il materiale citazionale in macrocategorie documentarie: articoli in atti di convegno; articoli in rivista; documentazione in social network (documenti pubblicati in blog, forum…); documentazione normativa (atti normativi, linee guida, specifiche e standard); documentazione tecnica (guide tecniche, manuali utenti, specifiche tecniche); libri; materiale divulgativo (opuscoli, manifesti, depliant, locandine…); ontologie; preprint (submitted, under review); report (rapporti tecnici, libri bianchi, rapporti di ricerca, working papers); risorse linguistiche (corpora specialistici, vocabolari, thesauri, grammatiche, strumenti linguistici); siti web; software/tools (software scaricabile, demo); tesi (triennale, magistrale o dottorale); materiale didattico (tutorial); voci enciclopediche. Sono stati esaminati 3861 riferimenti bibliografici estratti da 248 articoli.

Analisi comparativa dei dati

Lo scenario

La tabella 1 mostra le macrocategorie documentarie citate nelle bibliografie AIUCD 2012-2016 e CLiC-it 2014-2015. La classificazione effettuata in base ai criteri indicati nel paragrafo precedente soddisfa 15 voci per AIUCD e 12 per CLiC-it.

Come si desume dalla stessa tabella, il dato complessivo 2012-2016 attesta che i libri sono la tipologia più citata in AIUCD (43.2%), seguita dagli articoli in rivista (25%) e dai contributi in

atti di congresso (18.2%). In CLiC-it invece la tipologia più citata è quella dei contributi in atti di convegno (49.1% del totale) e a seguire gli articoli in rivista (23.5%) e i libri (21.5%).

Macrocategorie documentarie	AIUCD 2012-2106		Macrocategorie documentarie	CLiC-it 2014-2015	
	n.	%		n.	%
Libri	640	43.2	Contributi a convegno	1168	49.1
Articoli in rivista	370	25.0	Articoli in rivista	560	23.5
Contributi a convegno	270	18.2	Libri	512	21.5
Report	34	2.3	Tesi	45	1.9
Documentazione normativa	33	2.2	*Report*	31	1.3
Siti web	33	2.2	Documentazione tecnica	12	0.5
Documentazione tecnica	22	1.5	Siti web	11	0.5
Software/Tool	21	1.4	Documentazione normativa	10	0.4
Tesi	19	1.3	*Software/Tool*	10	0.4
Documentazione in *social networks*	17	1.1	Preprint	9	0.4
Ontologie	9	0.6	Risorse linguistiche	8	0.3
Preprint	7	0.5	*Deliverable*	3	0.1
Materiale divulgativo	3	0.2	Documentazione in *social networks*		
Voci enciclopediche	3	0.2	Materiale didattico		
Materiale didattico	1	0.1	Materiale divulgativo		
Deliverable			Ontologie		
Risorse linguistiche			Voci enciclopediche		
Totale riferimenti bibliografici	1482		**Totale riferimenti bibliografici**	2379	

Tabella 1. – Macrocategorie AIUCD e CLiC-it

Il grafico 1 propone il confronto delle tre categorie prevalenti tra le due conferenze. Per quanto riguarda AIUCD è possibile notare una certa discontinuità dei riferimenti ad articoli in rivista nei primi tre anni con una diminuzione significativa, iniziata nel 2015 e confermata nel 2016; più costanti le citazioni a libri con una flessione rilevante nel 2015.

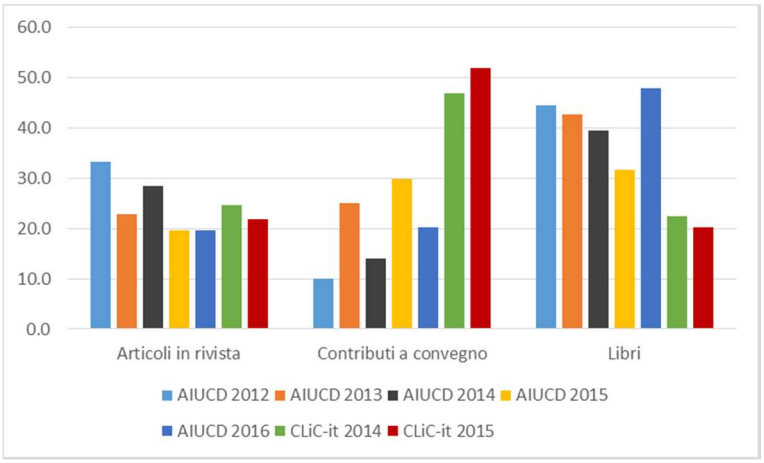


Grafico 1 – Categorie prevalenti in AIUCD e CLiC-it

Anche i riferimenti ai contributi a convegno sono abbastanza discontinui e alternano maggiore e minore presenza all'interno del corpus AIUCD, raggiungendo il valore più alto nel 2015 e quello minimo nel 2012. In CLiC-it, invece, queste tre macrocategorie mostrano variazioni minori nelle due annate di vita della conferenza.

La tabella 2 mostra la distribuzione annuale delle altre categorie documentarie in entrambe le conferenze.

Macrocategorie documentarie	AIUCD										CLiC-it			
	2012		2013		2014		2015		2016		2014		2015	
	n.	%	n.	%	n.	%	n.	%	n.	%	n.	%	n.	%
Deliverable											3	0.2		
Documentazione in social network	2	0.7	2	0.9	11	3.2			2	0.4				
Documentazione normativa	11	3.7	3	1.3	3	0.9	9	7.7	7	1.4	4	0.3	6	0.6
Documentazione tecnica					3	0.9			19	3.9	9	0.7	3	0.3
Materiale didattico									1	0.2				
Materiale divulgativo	3	1.0												
Ontologie					7	2.0	2	1.7						
Preprint	3	1.0			1	0.3			3	0.6	4	0.3	5	0.5
Report	9	3.0	4	1.7	13	3.7	3	2.6	5	1.0	11	0.8	20	2.0
Risorse linguistiche											5	0.4	3	0.3
Siti web	3	1.0	3	1.3	16	4.6	7	6.0	4	0.8	8	0.6	3	0.3
Software/Tool	3	1.0	8	3.4	2	0.6			8	1.6	8	0.6	2	0.2
Tesi	3	1.0	2	0.9	5	1.4	1	0.9	8	1.6	26	1.9	19	1.9
Voci enciclopediche					1	0.3			2	0.4				

Tabella 2. Altre categorie AIUCD e CLiC-it – Distribuzione annuale

Multilinguismo e copertura temporale

Nonostante la prevalenza delle citazioni in lingua inglese nelle bibliografie di entrambe le conferenze, l'osservazione complessiva autorizza a parlare di *multilinguismo citazionale*.

Lingue	AIUCD 2012-2016		Lingue	CLiC-it 2014-2015	
	n.	%		n.	%
inglese	1002	67.61	inglese	2178	91.55
italiano	362	24.43	italiano	158	6.64
francese	64	4.32	francese	30	1.26
tedesco	27	1.82	tedesco	5	0.21
spagnolo	14	0.94	latino	4	0.17
latino	8	0.54	arabo	2	0.08
arabo	3	0.20	rumeno	1	0.04
croato	1	0.07	spagnolo	1	0.04
russo	1	0.07	croato		
rumeno			russo		
Totale	**1482**		Totale	**2379**	

Tabella 3. – Lingue in AIUCD e CLiC-it[3]

La tabella 3 fornisce il prospetto delle lingue dei riferimenti bibliografici. In entrambi i corpora AIUCD e CLiC-it le prime quattro lingue sono inglese, italiano, francese e tedesco. Si rileva tuttavia qualche differenza nelle percentuali di utilizzo.

In CLiC-it i richiami in lingua inglese costituiscono quasi il 92%; l'italiano è presente al 6.7% circa, il francese al 1.26% e in percentuali molto ridotte troviamo il tedesco e il latino così come l'arabo, il rumeno e lo spagnolo. Il multilinguismo sembra essere più accentuato in AIUCD. La varietà delle lingue che si manifesta nei corpora 2012-2016 si differenzia da CLiC-it solo per la presenza di un riferimento in croato e uno in russo, ma le altre lingue risultano più consistenti in termini percentuali. La lingua inglese conta un totale di 1002 riferimenti nell'intero corpus 2012-2016 (68% circa) risultando pertanto meno citata rispetto a CLiC-it mentre l'italiano si trova quasi al 25%, il francese al 4.32% e il tedesco a 1.82%. Il latino e lo spagnolo, sebbene scarsamente

[3] Per visualizzare correttamente alcuni risultati in questa tabella sono state utilizzate due cifre decimali.

presenti raggiungono comunque frequenze superiori a quelle di CLiC-it. Anche in AIUCD sono presenti riferimenti bibliografici in lingua araba.

AIUCD																
	***pre*'50**		**'50**		**'60**		**'70**		**'80**		**'90**		**2000**		**in stampa**	**nd**
	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	n.
2012	8	2.7	1	0.3	4	1.3	12	4.0	10	3.3	28	9.3	230	76.4		8
2013	8	3.4	4	1.7	4	1.7	8	3.4	10	4.3	18	7.8	176	75.9	2	2
2014	9	2.6	1	0.3	8	2.3	9	2.6	20	5.8	37	10.7	236	68.0		27
2015	3	2.6	2	1.7	0	0.0	2	1.7	1	0.9	5	4.3	95	81.2		9
2016	26	5.4	7	1.4	9	1.9	17	3.5	25	5.2	49	10.1	341	70.3	3	8

Tabella 4. Copertura temporale dei riferimenti AIUCD

La copertura temporale dell'intero corpus AIUCD, presentata in tabella 4 è abbastanza estesa e ciò è dovuto alla presenza di fonti testuali distribuite tra il 1666 e il 2016. Escludendo la fonte testuale del 1666 (in AIUCD 2013) e la fonte del 1739 (in AIUCD 2016), il resto delle citazioni precedenti al 1950 si concentra su documenti pubblicati nell'Ottocento (1810-1895) e, con maggiore intensità, nei primi anni del Novecento (1901-1949). L'intervallo di anni dal 1950 al 1990 si assesta su una percentuale del 20%, per cui sono gli anni Duemila ad accogliere la gran parte dei documenti citati (70-80%). All'interno di questi ultimi prevale l'intervallo 2011-2016, ad eccezione del *corpus* 2012 dove sono più numerosi i richiami al periodo 2006-2010. Una piccola percentuale è attribuibile a documenti in corso di stampa o privi di data.

CLiC-it																
	pre'50		**'50**		**'60**		**'70**		**'80**		**'90**		**2000**		**in stampa**	**nd**
	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	**%**	n.	n.
2014	9	0.7	10	0.7	12	0.9	41	3.0	58	4.3	162	11.9	1061	78.0	8	0
2015	2	0.2	2	0.2	17	1.7	17	1.7	24	2.4	65	6.4	884	86.8	7	0

Tabella 5. Copertura temporale dei riferimenti CLiC-it

Il riferimento più remoto di CLiC-it è contenuto nel corpus 2014 ed è attribuibile al 1909, mentre i più recenti sono naturalmente quelli dell'anno della conferenza, il 2015. La copertura temporale, dunque, presenta un numero di citazioni anteriori al 1950 inferiore a AIUCD. Le notizie riferibili agli anni precedenti il 1950 sono circoscrivibili al periodo 1909-1948. In CLiC-it gli anni Duemila sono i più citati e rappresentano il 78% nel corpus 2014 e quasi il 90% nel corpus del 2015. Le citazioni a documenti pubblicati tra gli anni '50 e '90 del Novecento corrispondono al 17% circa. In CLiC-it i riferimenti bibliografici in corso di stampa costituiscono solo lo 0.6% del totale.

In AIUCD una discreta porzione di citazioni è costituita da fonti testuali, ovvero da riferimenti a documenti originali o scholarly editions. La presenza più cospicua di fonti testuali è presente nei corpora 2012, 2013 e 2016. Più bassa è la loro presenza nel 2014 e nel 2015 (1.7% in entrambe le annate). Le fonti riferiscono manoscritti, lettere e testi per cui confluiscono quasi interamente nella macrocategoria libri. Oltre all'inglese, la lingua prevalente, le citazioni sono in italiano, latino, arabo, francese, russo e tedesco. In CLiC-it, la presenza di fonti testuali è trascurabile.

Interdisciplinarità

Entrambe le conferenze conducono a spazi conoscitivi interdisciplinari e trans-disciplinari.

«*The spirit of CLiC-it is inclusive. In the conviction that the complexity of language phenomena needs cross-disciplinary competences, CLiC-it intends to bring together researchers of related disciplines such as Computational Linguistics, Linguistics, Cognitive Science, Machine Learning, Computer Science, Knowledge Representation, Information Retrieval and Digital Humanities*» *(Basili et al, 2014).*

«*L'Associazione per l'Informatica Umanistica e la Cultura Digitale intende promuovere e diffondere la riflessione metodologica e teorica, la collaborazione scientifica e lo sviluppo di pratiche, risorse e strumenti condivisi nel campo dell'informatica umanistica e nell'uso delle applicazioni digitali in tutte le aree delle scienze umane, nonché promuovere inoltre la riflessione sui fondamenti umanistici delle metodologie informatiche e nel campo delle culture di rete*». *(Homepage AIUCD)*

La prima citazione è stata estratta dall'introduzione alla prima Conferenza Italiana di Linguistica Computazionale CLiC-it, tenuta a Pisa nel dicembre 2014, mentre la seconda proviene dal sito web dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, dove compare quale "dichiarazione di intenti" dell'Associazione medesima. L'interdisciplinarità si rileva dai differenti segmenti di conoscenza proposti dai riferimenti bibliografici che intrecciano la tecnologia con le scienze umane e investono differenti tipologie di digital humanities, aprendo così il dialogo tra strumenti e modelli provenienti da diversi ambiti di ricerca: progetti raccontati e delineati sulle pagine di portali dedicati a biblioteche digitali di grandi proporzioni; infrastrutture informatiche che dialogano con i contenuti; ricerche filologiche e di critica testuale coadiuvate da edizioni elettroniche e database testuali. La tecnologia fornisce i mezzi necessari al raggiungimento di obiettivi importanti e complessi. In questo spazio si sviluppano progetti che coniugano sistemi informatici, strumenti semantici e ontologie per l'analisi dei contenuti sempre più spesso discussi e condivisi sui nuovi modelli di comunicazione della scienza come chat, blog e forum.

Tracce di interdisciplinarità emergono anche osservando il lessico. Le citazioni in AIUCD conducono non poche volte all'aggettivo *digital* e l'esame dei sostantivi che accompagnano l'aggettivo conferma i diversi ambiti disciplinari coinvolti: *age*, *archive*, *content*, *demotic*, *discipleship, ecosystem*, e*ditions*, *heritage*, *history*, *images*, *library*, *media*, *natives*, *philology*, *presevation*, *resources*, *world*. Tra questi sostantivi il termine *resources* è molto presente anche nei riferimenti CLiC-it.

Per la classificazione disciplinare dei riferimenti bibliografici delle due conferenze è stato effettuato un raggruppamento concettuale dei contenuti argomentativi estratti dalle titolature citate. In molti casi le Associazioni hanno facilitato il lavoro di etichettatura.

Allo scopo sono stati elaborati 24 raggruppamenti tematici:

1. English Studies / Anglistica - ANG
2. Computational Linguistics - CL
3. Computer Science - CS
4. Corpus Annotation - CA
5. Digital Archives - DA
6. Digital Humanities & Cultural Heritage - DH&CH
7. Digital Libraries - DL
8. Digital Philology - DP
9. History & Philosophy - HP

10. Information Retrieval - IR
11. Italian Studies / Italianistica - ITA
12. Latino - LAT
13. Linguistics - LIN
14. Machine Learning - ML
15. Machine Translation – MT
16. Markup Language-Standard - MLS
17. Named Entities Recognition - NER
18. Natural Language Learning - NLL
19. Ontology - ONT
20. Psycholinguistics - PSY
21. Scholarly Editing - SE
22. Semantic Web - SW
23. Sentiment and Social Media Analysis - SSMA
24. Treebanks Parsing - TP

La tabella 6 presenta la distribuzione degli argomenti citati dalle due conferenze.

	AIUCD										CLiC-it			
Soggetti	**2012**		**2013**		**2014**		**2015**		**2016**		**2014**		**2015**	
	n.	*%*	*n.*	*%*	*n.*	*%*	*n.*	*%*	*n.*	*%*	*n.*	*%*	*n.*	*%*
ANG	21	7.0												
CA											20	1.5	36	3.5
CL	6	2.0	31	13.4	24	6.9	23	19.7	42	8.7	723	53.1	504	49.5
CS	31	10.3	23	9.9	52	15.0	13	11.1	51	10.5	123	9.0	80	7.9
DA	48	15.9	4	1.7	34	9.8	8	6.8	21	4.3	5	0.4		
DH&CH	89	29.6	67	28.9	45	13.0	49	41.9	125	25.8	32	2.4	16	1.6
DL			12	5.2	2	0.6								
DP	15	5.0	21	9.1	18	5.2			11	2.3				
HP	56	18.6	36	15.5	117	33.7	2	1.7	87	17.9	54	4.0		
IR											12	0.9		
ITA			3	1.3	6	1.7	5	4.3	13	2.7				
LAT	12	4.0	3	1.3	13	3.7	4	3.4	14	2.9	34	2.5	4	0.4
LIN	1	0.3	19	8.2	9	2.6	2	1.7	25	5.2	74	5.4	94	9.2
ML											40	2.9		
MLS	6	2.0	9	3.9	3	0.9	3	2.6	38	7.8	3	0.2	10	1.0
MT											46	3.4	38	3.7
NER											37	2.7	16	1.6
NLL													64	6.3
ONT	2	0.7			11	3.2	6	5.1	10	2.1	13	1.0	11	1.1
PSY					1	0.3					68	5.0	11	1.1
SE	8	2.7	1	0.4	4	1.2	2	1.7	28	5.8				
SSMA											13	1.0	74	7.3
SW	4	1.3			1	0.3			8	1.6				
TP											61	4.5	58	5.7
non definibili	2	0.7	3	1.3	7	2.0			12	2.5	3	0.2	2	0.2
Totale	301		232		347		117		485		1361		1018	

Tabella 6. Soggetti in AIUCD e CLiC-it

Alcune tematiche più specifiche sono recuperate soltanto nella conferenza di riferimento, altresì non pochi argomenti sono recuperati in entrambe le conferenze (grafico 2).

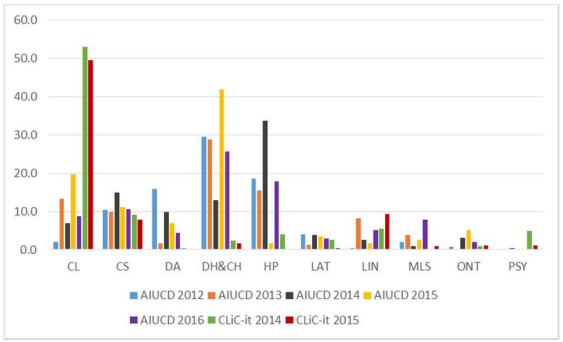


Grafico 2.Soggetti comuni AIUCD e CLiC-it

Text Encoding Initiative (TEI), XML e *standards,* più in generale, sono un tratto in comune; i lessici, gli studi linguistici, i corpora, e più in generale, le risorse linguistiche per le varie lingue sono trasversali ai due ambiti citazionali. Gli studi di anglistica, francesistica, germanistica e italianistica caratterizzano per lo più gli ambienti AIUCD, ma non pochi riferimenti in CLiC-it si avvicinano a questi settori di conoscenza. Le Associazioni di computer science coinvolte nelle citazioni CLiC-it e AIUCD forniscono una misura degli spazi in comune nella creazione di piattaforme e infrastrutture per archivi digitali di varia natura. Testi significativi di storia e filosofia emergono dalle bibliografie, non mancano riferimenti a opere generali come dizionari e enciclopedie.

Le citazioni inerenti tematiche di *digital heritage* sono state annotate con DH&CH, a tal proposito riportiamo una citazione di Fabio Ciotti "… In prima approssimazione diciamo che si tratta di un settore di studi e pratiche interdisciplinare e multi-disciplinare che riguarda l'applicazione dell'informatica, della fisica, della chimica e dell'ingegneria ai beni culturali, e l'uso della tecnologia per la rappresentazione, documentazione, archiviazione, protezione preservazione, comunicazione e valorizzazione del Patrimonio Culturale" (AIUCD, 2015). In questa classificazione abbiamo inserito tutte le citazioni di *digital heritage* e quelle di *digital humanities*.

Il panorama editoriale

Il concetto di multidisciplinarità si accorda anche con la varietà delle associazioni scientifiche incontrate nei *corpora* e dal panorama editoriale che ne emerge.

L'esame dei produttori ha messo in evidenza, soprattutto in CLiC-it, l'importante ruolo di *publisher* svolto dalle Associazioni, sia a livello nazionale che internazionale. Le fonti risultano popolate da acronimi di Associazioni e *SIG-Special Interest Groups*. Questi gruppi si formano all'interno delle Associazioni con lo scopo di costituire un nucleo scientifico specifico. In entrambe le annate di CLiC-it i documenti risultano pressoché pubblicati per il 50% da Associazioni di settore e per il restante 50% da editori commerciali tradizionali. Alcune Associazioni si configurano come dei veri e propri editori commerciali, altre invece svolgono il loro compito in modalità *no-profit*. In ogni caso, è evidente il fondamentale contributo che riescono a fornire alla comunità scientifica.

I servizi svolti accrescono le possibilità di dialogo tra scienziati, incoraggiano lo scambio culturale tra comunità interdisciplinari e valorizzano le attività svolte dai singoli.

Alcune associazioni sono in comune ai due *corpora* citazionali. Gli ambienti informatici per entrambe le conferenze sono condivisi con l'Association for Computer Machinery ACM, l'IEEE Computer Society, l'Association for the Advancement of Artificial Intelligence AAAI; gli ambienti linguistici e linguistico-computazionali sono accomunati dall'Association for Computational Linguistics ACL, dall'European Language Resources Association ELRA, dall'Association pour le Traitement Automatique des Langues ATALA, dall'European Association for Lexicography EURALEX e dalla Linguistic Society of America LSA.

Le associazioni non condivise caratterizzano le due aree come ad esempio l'European Association for Digital Humanities EADH per l'AIUCD e la Global WordNet Association GWA per il CLiC-it.Tra le associazioni, alcune si dimostrano particolarmente sensibili alle tematiche open access e più impegnate nello sviluppo di migliori strategie di divulgazione della letteratura scientifica.

Modalità di accesso e formati

La verifica dei dati ad accesso aperto ha fatto emergere dati molto interessanti. Il grafico 2 mostra le frequenze relative dei documenti ad accesso aperto nei riferimenti bibliografici delle singole annate.

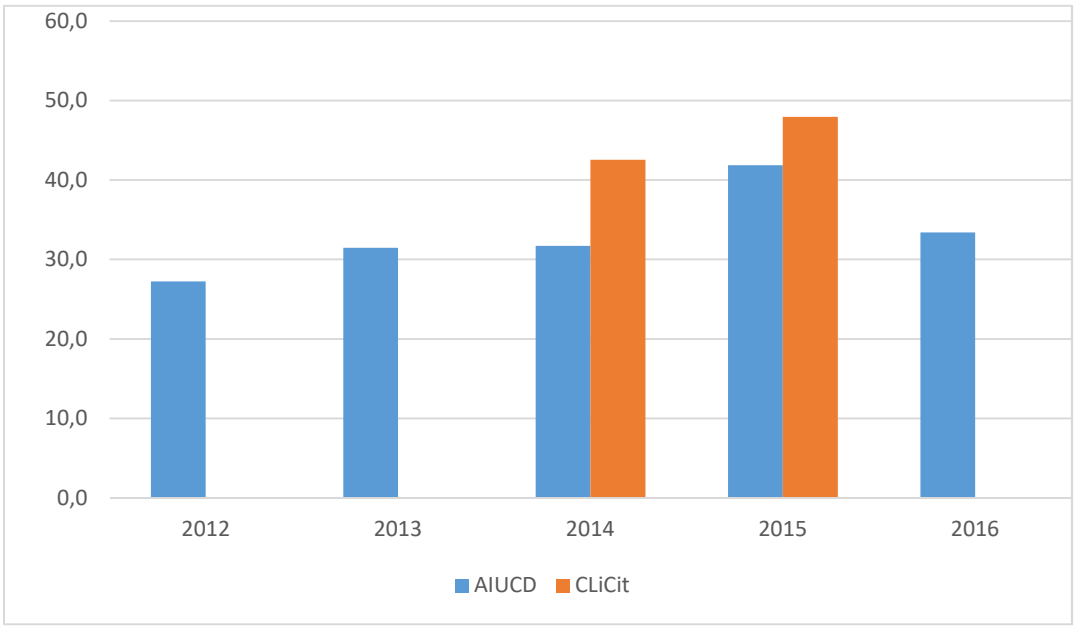


Grafico 3. Open Access in AIUCD e CLiC-it

Osservando i risultati vediamo che in AIUCD il dato open access (OA) è discontinuo. La percentuale più bassa è visibile nel 2012 (27.2%) e la più alta nel 2015(41.9%). Il valore non presenta un aumento costante e parallelo all'affermazione del movimento, nel 2016 anzi la percentuale di documenti ad accesso aperto scende sensibilmente rispetto al 2015. Nelle due annate comparabili si registra una crescita a dimostrazione che gli autori di ambedue le conferenze hanno incrementato le citazioni a ocumenti ad accesso aperto. Il dato complessivo attesta però che sono gli autori di CLiC-it a praticare più intensamente l'OA.

Macrocategorie documentarie	**AIUCD**														
	2012			**2013**			**2014**			**2015**			**2016**		
	n. rif.	*n. rif. OA*	*Frequenza di citazione OA (%)*	*n. rif.*	*n. rif. OA*	*Frequenza di citazione OA (%)*	*n. rif.*	*n. rif. OA*	*Frequenza di citazione OA (%)*	*n. rif.*	*n. rif. OA*	*Frequenza di citazione OA (%)*	*n. rif.*	*n. rif. OA*	*Frequenza di citazione OA (%)*
Articoli in rivista	100	33	33.0	53	18	34.0	99	24	24.2	23	10	43.5	95	39	41.1
Contributi a convegno	30	8	26.7	58	30	51.7	49	14	28.6	35	21	60.0	98	54	55.1
Deliverable															
Documentazione in *social network*	2	2	100.0	2	1	50.0	11	11	100.0				2	2	100.0
Documentazione normativa	11	9	81.8	3	3	100.0	3	3	100.0	9	6	66.7	7	6	85.7
Documentazione tecnica							3	3	100.0				19	17	89.5
Libri	134	11	8.2	99	9	9.1	137	14	10.2	37	0	0.0	233	20	8.6
Materiale didattico													1	0	0.0
Materiale divulgativo	3	3	100.0												
Ontologie							7	7	100.0	2	2	100.0			
Preprint	3	2	66.7				1	0	0.0				3	1	33.3
Report	9	7	77.8	4	0	0.0	13	13	100.0	3	3	100.0	5	5	100.0
Risorse linguistiche															
Siti web	3	3	100.0	3	3	100.0	16	16	100.0	7	7	100.0	4	4	100.0
Software/Tool	3	3	100.0	8	8	100.0	2	2	100.0				8	8	50.0
Tesi	3	1	33.3	2	1	50.0	5	2	40.0	1	0	0.0	8	4	100.0
Voci enciclopediche							1	1	100.0				2	2	**33.4**
Totale citazioni	**301**	**82**	**27.2**	**232**	**73**	**31.5**	**347**	**110**	**31.7**	**117**	**49**	**41.9**	**485**	**162**	**33.4**

Tabella 7. Open Access in AIUCD

Macrocategorie documentarie	CLiC-it					
	2014			2015		
	n. rif.	*n. rif. OA*	*Frequenza di citazione OA (%)*	*n. rif.*	*n. rif. OA*	*Frequenza di citazione OA (%)*
Articoli in rivista	337	48	14.2	223	44	19.7
Contributi a convegno	640	460	71.9	528	390	73.9
Deliverable	3	3	100.0			
Documentazione in *social network*						
Documentazione normativa	4	4	100.0	6	5	83.3
Documentazione tecnica	9	9	100.0	3	3	100.0
Libri	306	2	0.7	206	2	1.0
Materiale didattico						
Materiale divulgativo						
Ontologie						
Preprint	4	2	50.0	5	4	80.0
Report	11	11	100.0	20	19	95.0
Risorse linguistiche	5	5	100.0	3	3	100.0
Siti web	8	8	100.0	3	3	100.0
Software/Tool	8	8	100.0	2	2	100.0
Tesi	26	19	73.1	19	13	68.4
Voci enciclopediche						
Totale citazioni	**1361**	**579**	**42.5**	**1018**	**488**	**47.9**

Tabella 8. Open Access in CLiC-it

L'open access negli articoli in rivista è più esercitato in AIUCD, nel collettivo 2012-2016 e nello specifico 2014 (24.2%) e 2015 (43.5%). La maggiore presenza è rilevabile nel 2015 e nel 2016 (41.1%). In CliC-it l'OA negli articoli in rivista raggiunge soltanto il 14.2% nel 2014 e il 19.7 nel 2015. Anche la citazione a libri o contributi a libro ad accesso aperto è più solida in AIUCD dove nel 2014 supera di poco il 10% rispetto allo 0.7% di CLiC-it 2014 e del 1% di CLiC-it 2015. I riferimenti a contributi a conferenze OA sono quelli che presentano il più alto valore percentuale, raggiungendo quasi il 74% in CLiC-it 2014 e il 60% in AIUCD 2015. In CLiC-it la maggiore incidenza dell'open access, legata in special modo a quest'ultima categoria è attribuibile al grande lavoro di alcune associazioni che hanno creato archivi digitali di grandi dimensioni che ospitano *proceedings* di congressi e altro materiale di ricerca e lo rendono fruibile ad accesso aperto. Questo è il caso, ad esempio, di ACL – Association for Computational Linguistics tramite l'archivio ACL Anthology che accoglie e rende disponibili pubblicazioni ACL e di altre organizzazioni. *Tabella 7. Open Access in AIUCD*

Anche in AIUCD le associazioni che contribuiscono all'open access sono molteplici ed alcune, come ADHO - Alliance of Digital Humanities Organizations pubblicano e forniscono libero ad accesso ad una gamma ancor più vasta di materiali di studio e di ricerca.

Per quanto riguarda l'influenza dell'open access nelle altre categorie documentarie possiamo dire che la maggior parte di esse sono completamente disponibili ad accesso aperto. Il software, che non sempre è *downloadable*, nel nostro contesto è invece totalmente e costantemente fruibile. Soltanto quando la notizia riguarda documentazione normativa riservata, *preprint* sottomessi per la pubblicazione piuttosto che documentazione tecnica scaricabile solo in modalità riservata, le percentuali di accesso aperto si abbassano.

Il formato della documentazione citata è molto spesso sia cartaceo sia elettronico perché ormai, nell'editoria tradizionale commerciale, gli editori forniscono entrambe le versioni. Tuttavia, soprattutto in AIUCD, cominciano ad emergere citazioni a *journal* nativi online e disponibili soltanto in versione elettronica. La maggior parte di questi *journals* è disponibile ad accesso aperto in quanto nasce e si sviluppa a sostegno del movimento OA ed è creata e mantenuta, a sua

volta, con strumenti *open source* in grado di gestire il ciclo editoriale nella sua totalità e complessità.

Questi *journal* sono presenti in AIUCD per il 35% circa mentre in CLiC-it non vanno oltre il 6.5%.

Nelle bibliografie di entrambe le conferenze sono presenti anche citazioni a versioni digitali di alcuni libri o contributi a libri ma sono ancora abbastanza esigue, soprattutto in CLiC-it (0.5%) mentre gli autori AIUCD le indicano nel 6.5% dei casi.

Conclusioni

Il modello tradizionale di citazione - vale a dire l'abitudine di citare la letteratura convenzionale - è ancora molto forte e lascia poco spazio a modelli alternativi. L'indagine restituisce pertanto la tendenza alla citazione di documentazione tradizionale come conferenze, riviste e libri in entrambi i *corpora*. La maggior parte delle citazioni bibliografiche è inserita in circuiti editoriali e appartiene alla "letteratura convenzionale" ovvero pubblicata mediante i classici canali commerciali, mentre un insieme minore è assimilabile alla cosiddetta "letteratura grigia" o "letteratura non convenzionale" contraddistinta dalla appartenenza a circuiti non propriamente editoriali, prodotta dalla comunità scientifica e, più in generale, da istituzioni pubbliche e private. Si tratta molto spesso di tipologie quali *reports* documentazione tecnica, linee guida, tesi… riscontrabili nelle citazioni delle due conferenze. Le categorie di documenti quali blog, forum, dataset, media e social media sonodel tutto assenti in CLiC-it, mentre stanno emergendo in AIUCD, che presenta anche un numero più consistente di riferimenti ad oggetti digitali, nel senso più ampio del termine. Rispetto a CLiC-it, in AIUCD sono riscontrabili percentuali molto più alte di riferimenti dotati di URL, di DOI nonché di collegamenti diretti al documento citato.

Il modello citazionale di AIUCD rispetto a CLiC-it parrebbe quindi più orientato alla fruizione immediata del documento da parte del lettore.

Pur prestando forte attenzione ad oggetti digitali, la peculiare componente filologica di AIUCD richiede che buona parte degli autori debba necessariamente citare libri (manoscritti e testi a stampa), mentre la maggiore componente tecnologica di CLiC-it spinge gran parte dei suoi autori verso articoli pubblicati in atti di congresso. Solo in AIUCD è presente la citazione di materiale divulgativo (nel 2012) e di materiale didattico (nel 2016), mentre solo in CLiC-it si rileva la citazione di *deliverables* e risorse linguistiche.

Sebbene i *repositories* e il modello open access abbiano aperto nuove strade e fornito importanti strumenti per agevolare l'accesso all'informazione, raramente sono state individuate citazioni a documenti archiviati in repository, mentre la consultazione di queste infrastrutture è stata di grande ausilio nel processo di identificazione delle fonti.

Il lavoro ha soprattutto reso evidente l'impegno di molte associazioni e organizzazioni nella condivisione e conservazione delle fonti stesse, talora anche di altro genere quali, ad esempio, demo, poster e materiale informativo. Inoltre, sebbene molti siano i nodi ancora da sciogliere in merito ai prodotti open e al loro impatto citazionale (Guerrini, 2010), (De Bellis, 2005 e 2009), (Eysenbach, 2006), le comunità scientifiche di questi settori disciplinari appaiono particolarmente sensibili alle tematiche OA. È questo un risultato molto incoraggiante: la diffusione del paradigma dell'accesso aperto potrebbe accelerare il riconoscimento e la diffusione dei risultati della ricerca e aprire nuove modalità citazionali, favorendo la rapidità di accesso al testo e la fruibilità dello stesso agli specialisti di questo e di altri settori della conoscenza.

Ringraziamenti

Si ringraziano Roberto Rosselli Del Turco e Francesca Tomasi per aver contribuito all'analisi fornendoci le informazioni relative a AIUCD 2013, 2014 e 2015.

Riferimenti Bibliografici

Barbera, Manuel, and Carla Marello. 2012. «Corpo a corpo con l'inglese della corpus linguistics, anzi della linguistica dei corpora». In *Atti del convegno internazionaleLingua italiana e scienze*, a cura di Annalisa Nesi, e Domenico De Martino, 357-370. Firenze: Accademia della Crusca.

Basili, Roberto, Alessandro Lenci, and Bernardo Magnini, eds. 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, Volume 1. Pisa: University Press.

Bird, Stephen, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark T. Joseph,Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan.2008. «The ACL Anthology Reference Corpus: A reference Dataset for Bibliographic Research». In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC-2008), 1755-1759. Paris: ELRA.

Bosco, Cristina, Sara Tonelli, and Fabio Massimo Zanzotto, eds. 2015. *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*. Torino: Accademia University Press.

Budapest Open Access Initiative. 2002. Ultimo accesso 4 gennaio 2017. <http://www.budapestopenaccessinitiative.org>.

Bethesda Statement on Open Access Publishing. 2003. Ultimo accesso 4 gennaio 2017. http://legacy.earlham.edu/~peters/fos/bethesda.htm.

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. Ultimo accesso 4 gennaio 2017. http://siba.unipv.it/biblioteche/banche_dati/berlin_declaration.pdf.

Cassella, Maria, and Oriana Bozzarelli. 2011. «Nuovi scenari per la valutazione della ricerca tra indicatori bibliometrici citazionali e metriche alternative nel contesto digitale». *Biblioteche Oggi*29:66-78.

Cassella, Maria. 2012. *Open Access e comunicazione scientifica: verso un modello di disseminazione della conoscenza*. Milano: Editrice Bibliografica.

Ciotti Fabio. 2015. *Digital Heritage/Digital Humanities: una linea di faglia*? https://infouma.hypotheses.org/295).

Guerrini, Mauro.2010.*Gli archivi istituzionali. Open access, valutazione della ricerca e diritto d'autore*. Milano: Editrice Bibliografica.

Dale, Robert. 2008. «Last Words: What's the Future for Computational Linguistics?»*Computational Linguistics* 34:621-624.

Dale, Robert, and Adam Kilgarriff. 2010. «Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task». In *Proceedings of 6thInternational Natural Language Generation Conference, (INLG 2010)*, 281-265. ACL.

De Bellis, Nicola. 2005.*La citazione bibliografica nell'epoca della sua riproducibilità tecnica - Bibliometria e analisi delle citazioni dallo Science Citation Index alla Cybermetrica.* Ultimo accesso 4 gennaio 2017. http://docplayer.it/384672-La-citazione-bibliografica-nell-epoca-della-sua-riproducibilita-tecnica.html.

De Bellis, Nicola. 2009. *Bibliometrics and Citation Analysis: from the Science Citation Index to Cybermetrics*. Lanham, MD, US: Scarecrow Press.

De Bellis, Nicola.2014.*Introduzione alla bibliometria: dalla teoria alla pratica.* Roma: AIB.

De Castro, Paola. 2006. «Letteratura grigia: una 'Cenerentola' che si trasforma». *Ricerca e Pratica*22:79-82.

Di Cesare, Rosa,DanielaLuzi, and Silvia Giannini.2013. «Towards an e-Publishing library service in Humanities and Social Sciences: a feasibility study». *Information Services and Use* 33:191-199.
Di Donato, Francesca. 2010. «Le sfide dell'Open Access al sistema di comunicazione della scienza». *Rivista elettronica della Società Italiana di Filosofia Politica*. Ultimo accesso 4 gennaio 2017. http://eprints.sifp.it/245/.
Eysenbach, Gunther. 2006. «Citation Advantage of Open Access Articles».*PLoS Biol*. Ultimo accesso 4 gennaio 2017. doi: 10.1371/journal.pbio.0040157.
Giannini, Silvia. 2015. «Archivi, biblioteche e la comunicazione possibile: il ruolo della tecnologia». *Archivi* X:83-107.
Lawrence, Steven, C. Lee Giles, and Kurt Bollacker. 1999. «Digital Libraries and Autonomous Citation Indexing». *IEEE Computer* 32:67-71.
Luzi, Daniela. 1992. «La letteratura grigia e le basi di dati in linea: primi risultati». In Atti del 1. Convegno Nazionale sulla Letteratura Grigia, a cura di V. Alberani e P. De Castro, 114-124. Roma: Istituto Superiore di Sanità.
Marzi, Claudia, Gabriella Pardelli, and Manuela Sassi. 2010. «Grey Literature and Computational Linguistics: from Paper to Net». *International Journal on Grey Literature* 6:145-148.
Metitieri, Fabio, e Riccardo Ridi. 2002. «La letteratura grigia». In *Biblioteche in rete: istruzioni per l'uso*. Roma: Laterza. Ultimo accesso 4 gennaio 2017.
http://www.laterza.it/bibliotecheinrete/Cap10/Cap10_10.htm.
Pardelli, Gabriella. 2003. «BIBLOS: Historical, Philosophical and Philological Digital Library of the Italian National Research Council». In *Computational Linguistics in Pisa - Linguistica Computazionale a Pisa*, a cura di A. Zampolli, N. Calzolari, L. Cignoni. *Linguistica Computazionale*. Special Issue, XVIII-XIX:519- 546.
Pardelli, Gabriella, Manuela Sassi, and Sara Goggi. 2004. «From Weaver to the ALPAC Report». In *Proceedings of theFourth International Conference on Language Resources and Evaluation* (LREC'04) edited by Maria Teresa Lino et al., 2005 – 2008. Paris, ELRA.
Schreibman, Susan, Ray Siemens, and John Unsworth, eds. 2004. *A Companion to Digital Humanities*.Oxford: Blackwell.
Schöpfel,Joachim. 2010. «Towards a Prague Definition of Grey Literature». In *Proceedings of the 12th International Conference on Grey Literature (GL12)*, 11-26.Amsterdam: Text Release.
Serini, Paola. 2003. «Attualità della letteratura grigia. Il ruolo delle biblioteche nella sua valorizzazione». *Biblioteche Oggi* 21:61-72.
Spärck Jones, Karen. 2007. «Computational Linguistics: What about the Linguistics?» *Computational Linguistics* 33:437-441.
Venuda, Fabio. 2012. *La citazione bibliografica nei percorsi di ricerca. Dalla galassia Gutenberg alla rivoluzione digitale*. Milano: Unicopli.
White, John. 1998. «ACM Opens Portal to Computing Literature». *Communications of the ACM*44:14-28.

Religious and gender issues through the lens of a TV series: watching OITNB

Mariachiara Giorda, Fondazione Bruno Kessler, giorda@fbk.eu
Giovanni Moretti, Fondazione Bruno Kessler, moretti@fbk.eu
Boris Rähme, Fondazione Bruno Kessler, raehme@fbk.eu
Marco Stranisci, Fondazione Benvenuti in Italia, marco.stranisci@benvenutiinitalia.it
Sara Tonelli, Fondazione Bruno Kessler, satonelli@fbk.eu
Federica Turco, CIRSDE, federica.turco@unito.it

Introduzione

In this work, we present an analysis of the widely followed TV series *Orange is the New Black* (OITNB), a US-American comedy-drama produced by Netflix, which tells the stories and the problems of a group of female prison inmates. Its importance is due to two factors: first, OITNB has been, and continues to be, very successful, attracting large audiences and having broad impact. Second, it focuses on topics that are crucial for super-diverse societies (Becci, Burchardt and Giorda, 2015), such as gender, spirituality, and ethnic minorities. In what follows, we provide a multidisciplinary analysis of these cultural and social aspects, bringing together an NLP-approach with perspectives from semiotics and religious studies. In particular, we describe how these issues are represented in OITNB, comparing the outcome of an automatic analysis of subtitles, reviews, and fan-discussions with a semiotic interpretation of the series content and with recent work in the sociology of monastic institutions.

Automated keyword extraction

As a first step towards the automated analysis of OITNB, we created a large corpus containing enough information to explore both the series content and its reception. In particular, we collected a dataset made of 419,500 tokens and divided in three subcorpora: the larger one is made of OITNB subtitles of the first three seasons (235,000 tokens) in English; the second one contains 127 reviews in Italian (95,500 tokens) collected by searching for 'recensione orange is the new black' with Google API and parsed with the *Scrapy* Python library; the third corpus is a collection of posts from the *Italiansubs forum* (http://www.italiansubs.net/forum/), a website where Italian fans discuss about TV shows.

For each corpus, we extracted the top-relevant 50 keywords using KD[1] (Moretti et al., 2015), a freely available tool for key-terms and phrases extraction. KD is a flexible, rule-based tool, which supports English and Italian and can be adapted to any domain by setting a list of parameters to extract a list of ranked keywords. The ranking algorithms is mainly based on frequency, but takes into account also other information such as the keyphrase length (i.e. users can optionally boost longer keywords) and the abstractness level (i.e. boosting terms with specific suffixes). We chose this tool for several reasons: first, its availability in English and Italian, which were required to analyse the corpora considered in this study and perform cross-lingual comparisons. Second it is an off-the-shelf tool that does not require training data (hence, no preliminary annotation work). Third, it is freely available, enabling other researchers to reproduce this work or apply a similar methodology to other sources of information. Finally, since it was developed at FBK, we had a complete control over all parameters and we could easily fine-tune them for our study.

1 http://dh.fbk.eu/technologies/kd

The configuration we chose for this analysis assigned high relevance to specific keywords, thus favouring longer key-phrases (up to 4 tokens). We also made an extensive use of the black-list included in the tool, allowing users to manually define a set of words or expressions to be excluded during keyword extraction. This was done because in subtitles, interjections, colloquial expressions and curse words are extremely frequent, and we had to discard them in order to focus on the actual content of the series, and to make the analysis comparable with reviews and discussion forums. Support of English and Italian provided by KD was necessary in order to compare English subtitles with reviews and discussions leveraged from Italian websites. The partial lists of keywords are reported in Fig. 1.

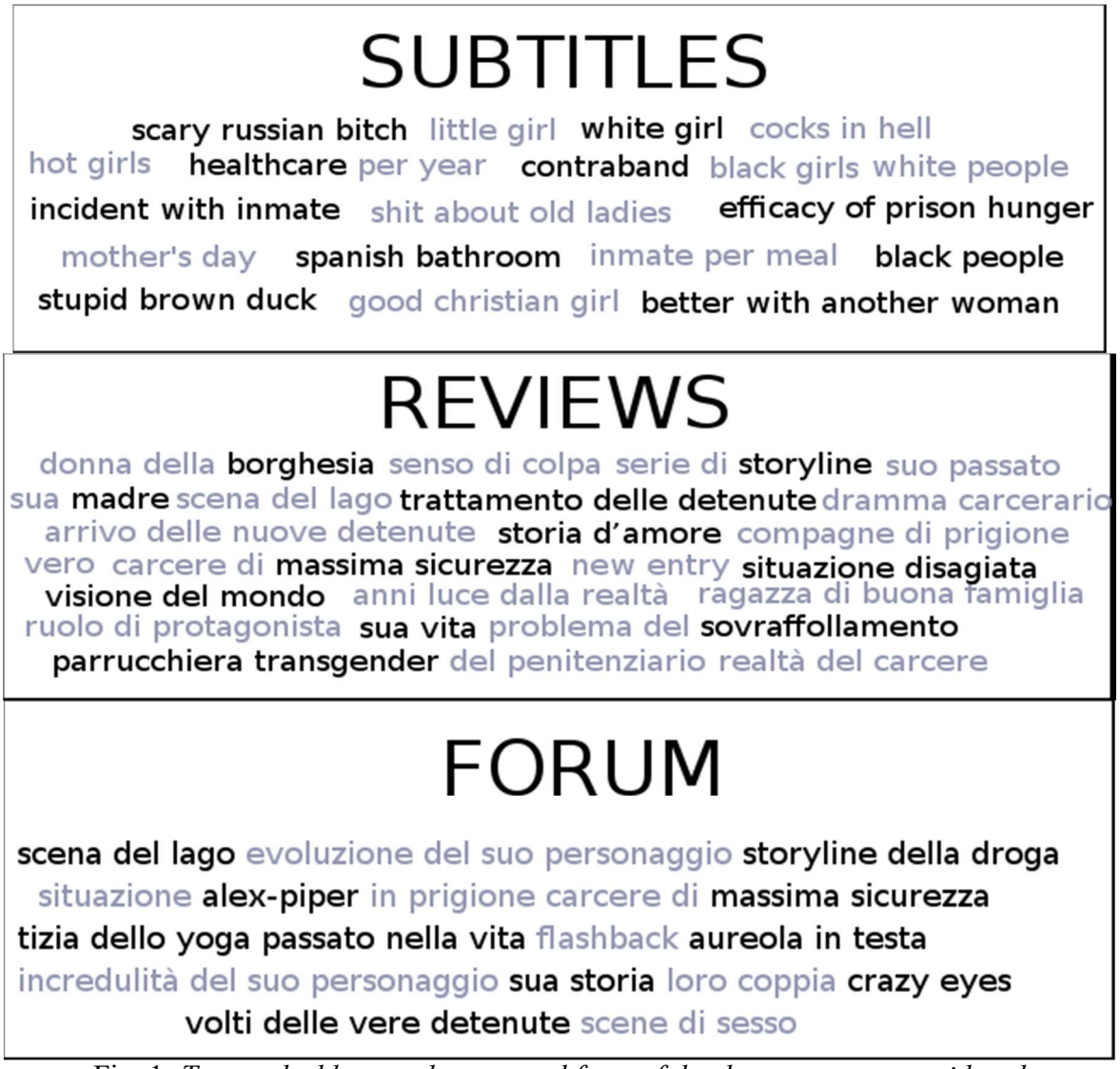


Fig. 1: *Top-ranked keywords extracted from of the three corpora considered*

Analysis and corpus comparison

In this section we provide an analysis of KD results and sketch three possible research directions: first, we investigate how OITNB has been received by critics and fans; second, we discuss how research on religious discourse can be connected with automated analysis. Finally, we compare the NLP-based approach with a traditional semiotic approach, in order to understand if these two theoretical perspectives are complementary.

OITNB keyphrases analysis

The key-phrases extracted from subtitles can be grouped into different categories: most of them refer to gender (e.g.: *little girl, hot girls, good girl, real woman, sexuation of baby girls, mom's saving*); others are hate speech and dirty words, that probably occur with high frequency to mimetically represent the life in prison (e.g.: *fuckin name, fuckin days, same shit*) or as epithets aimed at inmates (e.g.: *scary Russian bitch*). Moreover, we can find some references to the ethnic barriers in the prison, for example a strong polarization between white and black inmates (e.g.: *white girl, other black ladies*). Finally, issues about prison management are highlighted (e.g.: *healthcare per year, contraband after last week, efficacy of prison hunger*). If we compare these results with those obtained analyzing the "community of readers" represented by critics reviews (Jauss, 1967), we interestingly notice that categories do not perfectly fit. The critics focus on the series social criticism *(e.g.: problema del sovraffollamento, trattamento delle detenute, situazione disagiata),* but do not mention ethnic self-segregation, presumably because it is perceived as a marginal problem in Italy. Conversely, reviews show high awareness of gender discrimination, that affects in particular one character of the show (e.g.: *parrucchiera transgender del penitenziario*). Other keywords are related to the specific text-genre of reviews. Indeed, many of them seem to be used to provide the reader with contextual information about the storyline of the show (e.g.: *donna della borghesia, carcere di massima sicurezza, giovane donna, arrivo delle nuove detenute*). Finally, it is important to highlight the occurrence of 'motherhood' both in subtitles and in critics.

The third corpus in our collection represents the voice of OITNB fans and their opinion on the series. It shows important differences from the other corpora: first, fans seem to be more interested in love and sex situations (e.g: *scene di sesso, situazione alex-piper in prigione, loro coppia, ragazzo della biondina*). Moreover, we can notice an evaluative attitude towards characters (e.g.: *aureola in testa, santa donna*). Again, there is an interest in some particular scenes of the show for their aesthetic quality (eg: *scena del lago*); furthermore, the only recurrent topic in the field of social criticism is about drug dealing (e.g.: *storyline della droga, trafficante di droga*). Finally, some key-phrases are stylistic clues about the nature of these text (e.g.: *cosa migliore della serie, cosa bella*), that are considered *written speeches* (Bazzanella, 2010; Pistolesi, 2004).

Comparison between NLP and Semiotic analysis

From a semiotic perspective, each cultural work has a superficial and a deep level, and its elementary structure of signification (Greimas e Courtés, 1979) arises from the second one. In other words, discursive structures exist as a kind of fulfilment of semio-narrative ones and represent the part of the text where deep values are translated into factual elements such as characters, places, colours, figures, etc.

At the deep level, instead, we can find the value structure of the text and these values are the key to comprehend the meaning, the message, the reason why of the text itself.

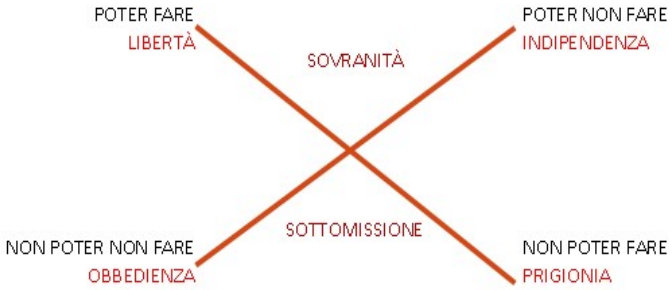


Fig. 2, *The semiotic square of Orange is the New Black*

The semiotic square (ibid.) is a formal representation of this signification, that holds together two opposite concepts (values). Following this theoretical perspective, we provided a semiotic analysis of OITNB, which led us to the creation of this fiction's semiotic square, based on the opposition between 'sovereignty' and 'submission' (Figure 2). As a matter of fact, the series characters are portrayed in the middle of this dichotomy: they live in a present of submission (the prison), that collides with their past of freedom, steadily represented with the technique of flashback in the show. The data extracted using KD would seem to confirm the interpretive hypothesis that the deep level of OITNB is structured by the opposition of sovereignty and submission, i.e., by the semiotic relations depicted in Fig. 2, since many keyphrases can be construed in terms of the opposition between sovereignty and submission.

A first relevant term is *mother*, that appears as a relevant keyword both in subtitles, and in reviews, and represents a strong constraint for inmates. Indeed, most of them are mothers unable to raise their children.

Another example of the correlation between Semiotics and NLP analysis concerns the term *sex*, that appears in all the corpora. Nevertheless, the critics treat it as a social issue, focusing on the storyline of the transgender character (*parrucchiera transgender del penitenziario/transgender hair-dresser of the prison*), while fans are mostly interested in it as a relational issue because it influences love stories of inmates (*scene di sesso/sex scenes*). However, in both cases, sex is linked with the relationship between inmates and their own bodies: on one hand, they use it as a tool of power, on the other they suffer a sort of reification, because they are harassed by prison guards. Another significant keyphrase, which arises from the forum corpus, is *scena del lago/lake scene*. The expression refers to the final scene of the third season: inmates find a hole in the prison fence and take this opportunity to have a bath in the lake near the prison. Thus, a dialogue between inside and outside is staged in OITNB and perceived as relevant by fans. Instead, reviews focus on the doubling of inmates, which takes place just during the lake scene (*arrivo delle nuove detenute/new inmates arrival*). Once again, we have two complementary reactions: fans are interested in emotions and feelings that concern the private sphere. Conversely, critics pay attention to social aspects of the detention problem.

Analysis of religious discourse

Although "religion" has become an important topic for scholars in many fields of the humanities and social sciences (Burchardt, Wohlrab-Sahr, Middell 2016), it is underrepresented in massive cultural works, such as TV series. Issues related to religion are rarely the core topic of the narration, and few characters are explicitly religious. In general, religion has a role only in moments of personal crisis, and religious groups tend to be under-represented, except for Catholics (Engstrom, Valenzano III, 2010) or, more generally, Christians (Clarke, 2005). Religion is, therefore, quite a scenery element – and for this reason the assumption is that it is Christian religion.

In OITNB religion is used by inmates for pragmatic purposes (e.g.: receiving donation from pro-life movements, hiding the phone in the chador, and pretending to be Jewish in order to have a decent meal), but this materialistic attitude is complementary to a deep need of meaning in many inmates. The extracted keywords seem to confirm the underrepresentation of religion in TV series: we found just two keyphrase, *good christian girl* and *visions of hell*, that refer to two minor characters who gain relevance with the continuation of the show. Furthermore, if we consider the Semiotic's superficial level of OITNB, this religious discourse is almost never part of the main plot.

Instead, we can draw a correlation between the deep level of OITNB (Figure 2), and religious studies, that helps interpret and shed light on OITNB's representation of the prison as an

"*institution totale*" (Foucault, 1975; Goffmann, 1968), characterized by a strong inside-outside dialectic and the tension between sovereignty and submission described above. Even though there are obvious limits to the analogy, it is instructive to compare prison settings to monastic settings (Giorda and Hejazi, 2014), since both can be interpreted as a totalizing place where agents live with very few possibilities to communicate with the external world, organizing themselves under strict rules and living the deep and strong presence of borders. Furthermore, the dialectic between exterior and interior involves also the psychology of inmates and, again, we can use the descriptive tools of (female) monasticism in order to interpret it: the prison fence is also to be seen as an inner one, borders are also internal borders and in particular there is a barrier that is physical, metaphorical, and spiritual. This strongly hierarchized space is not an abstract construction. It is created by the inmates' practices and behaviours, which build their new identities as prisoners. The jail is the centre which attracts, towards which the world goes, but also a watershed, a border which separates what and who is an insider from what and who is an outsider (De Certeau M., 1980). As in the case of monastic settings, the relationship between everyday life and environment is pivotal. It shapes not only the definition of the prison as an institution but also the development and continuous re-thinking of the inmates' identities (DeCerteau and Domenach, 1974).

In this perspective, we argue that "religion" appears as an instrument of exploration, starting from the dichotomy of "submission/sovereignty" which emerges from the semiotic square. Through the comparison with monastic life, we can analyse the meanings of identities of the inmates, their relationships and their representation.

Conclusions

In this paper we provided a multidisciplinary analysis of gender and religious issues in *Orange is the New Black* (OITNB), a Netflix series set in an American female prison. In the first part of our work, we described the creation of the corpora and the automated text processing. In the second part, we presented the analysis of the data, which in turn was divided into three sections: a comment on the content expressed by the extracted keyphrases, the semiotic analysis of OITNB in connection with these keywords and, finally, a discussion of religious discourse related to the series. In particular, we analysed religious discourse both at a textual level and at a deep level of meaning, suggesting a parallel between prison life and monastic choice.

Our study suggests that keyphrase extractors are valuable tools for guiding, enriching and validating interpretive hypotheses concerning the semantics of the linguistic level of cultural artifacts. While our focus was on the religious and gender issues figuring in OITNB, the approach can be applied to various other topics and, more generally, to text-based cultural artifacts at large.

In the future, we plan to explore how the domain-specific knowledge provided by religious discourse analysis and semiotics may, in turn, be used to fine-tune KD rules and filters.

Bibliographic References

Bazzanella, Carla. "Contextual constraints in CMC narrative." In *Narrative Revisited*, ed. Christian R. Hoffmann, 19-37. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2010.

Becci, Irene, Marian Burchardt, and Mariachiara Giorda. "Religious super-diversity and spatial strategies in two European cities." *Current Sociology* (2016): 0011392116632030.

Burchardt, Marian, Monica Wohlrab-Sahr, Micheal Middell (eds.). *Multiple Secularities Beyond the West. Religion and Modernity in the Global Age*. Berlin: De Gruyter, 2015.

De Certeau, Michel, Jean-Marie Domenach. *Le christianisme éclaté*, Paris, Éditions du Seuil, 1974

Demaria, Cristina. *Teorie di genere. Femminismo, critica postcoloniale e semiotica,* Milano: Bompiani, 2003.
Demaria, Cristina, and Mascio Lella. "Kill Bill vol.1: Migrazioni interculturali e propagazioni extra-testuali." In *Remix-Remake. Pratiche di replicabilità*, ed. Nicola Dusi and Lucio Spaziante Roma: Meltemi, 2006.
Demaria, Cristina, and Siri Nergaard, eds. *Studi culturali. Temi e prospettive a confronto*, Milano: McGraw-Hill, 2008.
Foucault, Michel. *Surveiller et punir*, Paris: Gallimard, 1975.
Giorda, Maria Chiara, Sara Hejazi. *Monaci senza dio: pratiche, senso, essenza*. Milano: Mimesis, 2014.
Gopalan, Lalitha. *Donne vendicatrici*, in *Lo strabismo di Venere*, DWF, nn. 37-38, 61-72, 1998.
Goffman, Erving. *Asylums: Essays on the social situation of mental patients and other inmates*. AldineTransaction, 1968.
Greimas, Algirdas J., and Jacques Courtés. *Sémiotique. Dictionnaire raisonné de la théorie du langage*. Paris: Hachette, 1979.
de Lauretis, Teresa. *Alice doesn't. Feminism, Semiotics, Cinema*, Indiana University Press, Bloomington, 1984.
de Lauretis, Teresa. *Technologies of gender. Essays on theory, film and fiction*. Bloomington: Indiana University Press, 1987.
Moretti, Giovanni, Rachele Sprugnoli, and Sara Tonelli. "Digging in the Dirt: Extracting Keyphrases from Texts with KD." In *Proceedings of the Second Italian Conference on Computational Linguistics*, Trento, Italy, 2015.
Pistolesi, Elena. *Il parlar spedito: l'italiano di chat, e-mail e SMS*, Esedra, 2004.
Turco, Federica. *Donne assassine nella fiction seriale italiana*, "Lexia Nuova Serie", n. 7-8: 299-316, 2011.

Short Texts in Authorship Attribution. The Case of Robert Musil's War Articles

Berenike Herrmann, Georg-August-Universität Göttingen,
jb.herrmann@phil.uni-goettingen.de[1]
Gerhard Lauer, Georg-August-Universität Göttingen, gerhard.lauer@phil.uni-goettingen.de
Simone Rebora, Georg-August-Universität Göttingen, simone.rebora81@gmail.com
Massimo Salgaro, Università degli Studi di Verona, massimo.salgaro@univr.it

The Case Study: Robert Musil and the *Soldaten-Zeitung*

Robert Musil, one of the most important authors of the twentieth-century German literature, fought in the Austrian army at the Italian front. During the First World War, between 1916 and 1917, Musil was chief editor of the *Tiroler Soldaten-Zeitung* in Bozen and later of the Viennese journal *Heimat*, where he probably authored numerous articles. This activity has always been a philological problem for Musil scholars, who have not been able to attribute a range of texts to the author and so far have not provided a clear description of Musil's style. We suggest that applying methods of formal authorship attribution helps solve both issues.

The first number of the *Tiroler Soldaten-Zeitung* was issued in Vienna in June 1915. In 1916, the publishing was entrusted to the Bozen-based *Heeresgruppekommando Erzherzog Eugen*, to which lieutenant Musil was assigned during the same year. At the beginning of October, Musil became the chief editor of the newspaper. After his arrival, major changes in the magazine's layout were introduced: the issue numbering was restarted and the same month the name was changed to *Soldaten-Zeitung*. Due to the repositioning of the commands and technical problems, the magazine publication ended in April 1917. Subsequently, Musil moved to Vienna, where he collaborated with the war journal *Heimat* from March to October 1918.

All the 43 numbers of *Soldaten-Zeitung* published with the collaboration of Musil are still extant, while only 17 issues from the 34 numbers of *Heimat* survived. Both in the case of the *Soldaten-Zeitung* and in the case of *Heimat*, all the articles were published anonymously. Musil scholars have never been able to define with certainty the number of texts written by the author.

In Musil studies, between 1960 and 2014, at least 40 articles were attributed to the author. However, the surprising aspect of these attributions is the lack of evidence for their assumptions. For example, Marie-Louise Roth lists 19 texts from *Tiroler Soldaten-Zeitung* introducing them with the cryptic phrase: "Anonyme Schriften [...] die bis jetzt noch nicht mit Sicherheit identifiziert wurden" (Roth 1972, 528) (anonymous texts which have not yet been identified with certainty). Subsequent studies, such as the one by Arntzen (1980), refer to Roth without highlighting her gaps in the argumentation. The Italian edition (Fontanari and Libardi 1987) simply includes all the texts previously indicated as Musil's production. The prestigious digital edition of Musil works is no more accurate, since the problem is defined as a "work in progress" (Amann et al. 2009). This paper focuses on Musil's articles in the *Soldaten-Zeitung*, evaluating how the adoption of some of the most recent stylometric methodologies may offer a solution to the problem.

1 The authors appear in alphabetical order.

Methodological Issues: the Problem of Text Length

After extensive testing of multiple stylometric methods, Eder (2015) comes to the conclusion that the minimum dimension for an attributable text chunk is about 5,000 words, independently of the (modern) language of composition—while the ideal 'comfort zone' is around 8,000 words. The average length of the 38 disputed Musil's articles in the *Soldaten-Zeitung* (as listed by Schaunig 2014, 356–357) is however slightly below 1,000 words. For our means, analysis of the individual texts thus poses a challenge.

Combinatory design (1). In an ideal project design, the combination of 8 texts from the *Soldaten-Zeitung* will be sufficient for reaching the 'comfort zone' between 5,000 and 8,000 words, but this approach will generate at least two major issues. First, it will work only on the assumption that at least 8 articles were actually written by Musil (but we lack any sufficient proof to sustain this assumption). Second, it will imply the adoption of a complex combinatory design: in fact, the testing of all the possible combinations of 8 texts out of a total of 38 texts will require a total of 48,903,492 iterations. A strongly demanding computational task even for the most powerful machines: as a rule of thumb, if the machine will perform one iteration per second, the whole process will take 1 year and 8 months to be completed. However, the complexity of this design can be reduced by introducing some careful simplifications.

Combinatory design (2). As shown by Figure 1, some texts fall deeply below the 1,000 word-mean (with a minimum of 47 words). Consequently, these texts may be preliminarily cut off from the experiment—because they are less probably attributable and because they sharply decrease the length of the text combinations. As demonstrated below, we have decided that a reasonable limit can be fixed to 500 words, thus cutting off 9 texts from the experiment. Another text that can be excluded is "Aus der Geschichte eines Regiments," published in the *Tiroler Soldaten-Zeitung* on 26 July 1916, because it is the only text attributed to Musil on solid philological grounds (see Corino 1973). The combinatory design could be repeated on this simplified corpus (composed by a total of 28 texts) reducing the number of combined texts to 6. The text chunks will have an average length of n=6,963 words with a standard deviation of n=909 words and only 0.99% of the texts will fall under the limit of 5,000 words. However, also this design could be highly demanding in computational terms: the iterations will be n=376,740, requiring (if we keep the rule of thumb of one iteration per second) 104 hours of work.

Combinatory design (3). In order to further reduce this complexity, an effective expedient can be the addition of some already attributed articles to the text chunks. "Politik in Österreich" and "Europäertum, Krieg, Deutschtum," published by Musil in *Der Lose Vogel* in December 1912 and in *Die Neue Rundschau* in April 1914, respectively, are the best fitting texts—because of their dates of composition and because of their political contents. These texts will offer n=2,255 words that can be kept as a fixed element in the composition of the text chunks. Iterating only three texts from the (already reduced) *Soldaten-Zeitung* corpus, it will be possible to obtain text chunks with an average length of n=5,736 words and a standard deviation of n=685 words. The percentage of text chunks falling under the limit of 5,000 words will increase to 14%, but the minimum size will be n=4,088 words. The number of iterations will sharply decrease to 3,276.

However, the biggest issue with these project designs will be the interpretability of the results. As demonstrated by preliminary experiments, their quality will primarily depend on the composition of the training set (the group of texts to be compared with the "test set" of the *Soldaten-Zeitung* articles).

In order to effectively attribute the articles to Musil, it will be necessary to compare them with other Musil writings, but also with some effective "impostors" (see Koppel and Winter 2014), as similar as possible to Musil's style. The articles published by Musil in various journals between 1911 and 1921, available in digital format through the Klagenfurter Ausgabe (Amann et al. 2009), offer enough material for the composition of the first part of the training set. Regarding the second

part, two approaches are possible. The first one will be tested in this paper: it will look for the most suitable texts that are available online through the KOLIMO repository (see Herrmann and Lauer 2016). The second one, currently under preparation, will look for the "impostors" in the issues of the *Tiroler Soldaten-Zeitung* published before Musil's collaboration. This approach will profit from the possibility of comparing the test set with texts written by Musil's future collaborators—who could also be the actual authors of some of the disputed articles.[2]

Despite the current limitations of the training set, preliminary experiments offer some promising results. However, methodological research advises against the application of only one approach. As noted by Juola (2015), only the combination of different stylometric methods can provide probabilistically significant results. By consequence, the approach developed here may simply be the first stage of a multilayered authentication chain.[3]

Preliminary Results: on the Feasibility of a Combinatory Approach

A preliminary testing of the methods described in the preceding section has been realized using the Stylo package, developed by Eder, Rybicki and Kestemont (2013) in the statistical programming environment R. The training set has been composed by articles published by Musil between 1911 and 1921. The number of the "impostors" has been fixed to a minimum of three (see Juola 2015, 106), as close as possible in terms of date, genre, and style (following the definition of Musil's style by Thöming, 1970): (1) a selection of literary reviews written by Stefan Zweig between 1902 and 1939; (2) selected passages from the works *Das große Bestiarium der modernen Literatur*, *Formen der Liebe* and *Frauen und Männer der Renaissance* by Franz Blei (published between 1922 and 1930); (3) a selection of letters written by Franz Kafka between 1909 and 1919. The texts were divided into chunks of length between 6,000 and 8,000 words, three for each author, thus composing a training set of 12 text chunks, which generate a stable "consensus tree" (see Figure 2).

In order to test the combinatory design (2), a number of test sets were composed by manually combining different disputed articles written by Musil. Even with this simplified training set, some combinations place themselves out of the 'Musil cluster' (see Figure 3), while the majority regularly group with it—or close to it (see Figures 4 and 5). This first result suggests the presence of texts in the test set which differ from the average configuration of Musil's style. However, their precise individuation—or, more properly, the definition of their probability—will be possible only after comparing the results of 376,740 iterations (which will provide 80,730 different results for each text).

A similar testing for design (3) hardly shows any text chunk placing itself outside of the Musil cluster. Evidently, the presence of 2,255 words actually written by Musil inside the test set, acts as a dominant attractor towards the Musil cluster. However, further investigation reveals how an improvement of the training set actually provides some statistically significant results.

As shown by the consensus trees, the texts closer to Musil are *Das große Bestiarium der modernen Literatur* and *Formen der Liebe* by Franz Blei. Using untested passages from these works and excerpts from another Musil article ("Der Anschluß an Deutschland," published in *Die neue Rundschau* in March 1919), we created an artificial test set, where the percentages of Blei's and Musil's writing were variated over a fixed total of 5,000 words. With artificial test sets

2 The realization of this approach is not possible at the moment, because the digitized version of the *Soldaten-Zeitung* provided by the Österreichische Nationalbibliothek is affected by too many mistakes in the OCR (20%, at a rough estimate) and a reprocessing of the scanned images is advisable.

3 Among the possible alternative approaches that offer promising results with short texts, we list briefly: "rolling stylometry" (Eder 2016), JGAAP (Juola et al. 2008), "bigrams of syntactic labels" (Hirst and Feiguina 2007) and the PAN competitions software (Stamatatos et al. 2014; for some promising results in German language, see Halvani et al. 2016).

composed by 50% Musil and 50% Blei, 60% Musil and 40% Blei, 70% Musil and 30% Blei, respectively, the text chunks always group with Blei's cluster (see Figure 6). A switch in the structure of the consensus tree happens when the "impostor" occupies 20% of the test set (i.e. 1,000 words): testing 4 different textual compositions, the test set groups twice with Musil's cluster and twice with Blei's cluster, generating also an evident instability in the structure of the consensus tree (see Figures 7 and 8). This observation can be explained examining the training set through Principal Component Analysis (PCA): while Zweig's and Kafka's texts group separately, Blei's texts surround Musil's texts (see Figure 9), thus providing an effective distractor for the attribution. With an artificial test set composed by 90% Musil and 10% Blei, finally, the text chunk always groups with the Musil cluster (see Figure 10), confirming how 500 words are not sufficient for modifying the behavior of the test set.

While these results clearly depend on the characteristics of the training set defined, they show that choosing better "impostors" (i.e. building the training set with the issues of the *Tiroler Soldaten-Zeitung* published before Musil's collaboration) allows identification of the texts not actually written by Musil. This even holds when adopting the simplified combinatory design (3). The fact that the threshold length is around 1,000 words is also a promising result, because this corresponds with the average length of the disputed Musil articles. However, once again, these results cannot be generalized and a more systematic research on this type of artificial test set is advisable.

This approach may eventually help tackling the biggest doubt that overshadows the project: if Musil was the chief editor of the *Soldaten-Zeitung*, it is also possible that he copy-edited some of the articles written by his collaborators, thus intermixing his authorial signal with those of others. Therefore, next to expanding the research to a more comprehensive corpus (comprising all issues of *Soldaten-Zeitung* and *Heimat*), our goal is to identify all—or at least the biggest part of—previously mistaken attributions. In relation to stylometry in general, this research has shown how the limit of 5,000 words, while necessary in itself for the construction of an effective project design, is not at all an insurmountable boundary. Especially when the researcher, instead of looking for positive scores or strong attributions, starts looking for negative scores and structural anomalies: where the authorial signals of the "impostors" come into view and a frail attribution may be disproved.

Bibliographic References

Amann, Klaus, Karl Corino, and Walter Fanta. 2009. *Robert Musil, Klagenfurter Ausgabe. Kommentierte digitale Edition sämtlicher Werke, Briefe und nachgelassener Schriften*. Klagenfurt: Robert Musil-Institut der Universität Klagenfurt. DVD edition.

Arntzen, Helmut. 1980. *Musil-Kommentar sämtlicher zu Lebzeiten erschienener Schriften außer dem Roman "Der Mann ohne Eigenschaften."* München: Winkler.

Corino, Karl. 1973. "Robert Musil, Aus der Geschichte eines Regiments." *Studi Germanici* 11: 109–115.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. 2013. "Stylometry with R: a suite of tools." In *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska, 487–489.

Eder, Maciej. 2015. "Does size matter? Authorship attribution, small samples, big problem." *Digital Scholarship in the Humanities* 30(2): 167–182.

Eder, Maciej. 2016. "Rolling Stylometry." *Digital Scholarship in the Humanities* 31(3): 457–469.

Fontanari, Alessandro, and Massimo Libardi, eds. 1987. *La guerra parallela*. Trento: Reverdito.

Halvani, Oren, Christian Winter, and Anika Pflug. 2016. "Authorship verification for different languages, genres and topics." *Digital Investigation* 16: 33–43.

Herrmann, Berenike, and Gerhard Lauer. 2016. “KARREK: Building and Annotating a Kafka/Reference Corpus.” In *Digital Humanities 2016: Book of Abstracts*. Kraków: Jagiellonian University & Pedagogical University, 552–553.
Hirst, Graeme, and Ol’ga Feiguina. 2007. “Bigrams of syntactic labels for authorship discrimination of short texts.” *Literary and Linguistic Computing* 22(4): 405–417.
Juola, Patrick John Noecker, Mike Ryan, and Mengjia Zhao. 2008. “JGAAP3.0 – authorship attribution for the rest of us.” In *Digital Humanities 2008: Book of Abstracts*. University of Oulu, 250–251.
Juola, Patrick. 2015. “The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions.” *Digital Scholarship in the Humanities* 30: 100–113.
Koppel, Moshe, and Yaron Winter. 2014. “Determining if two documents are by the same author.” *JASIST* 65(1): 178-187.
Roth, Marie-Louise. 1972. *Robert Musil. Ethik und Ästhetik*. München: List.
Schaunig, Regina. 2014. *Der Dichter im Dienst des Generals. Robert Musils Propagandaschriften im ersten Weltkrieg*. Klagenfurt: Kitab.
Stamatatos, Efstathios, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. “Overview of the Author Identification Task at PAN 2014.” In *CLEF 2014 (Working Notes)*, 877-897.
Thöming, Jürgen. 1970. “Wie erkennt man einen anonym-veröffentlichten Musil Text?” *Études germaniques* 25: 170–183.

Images

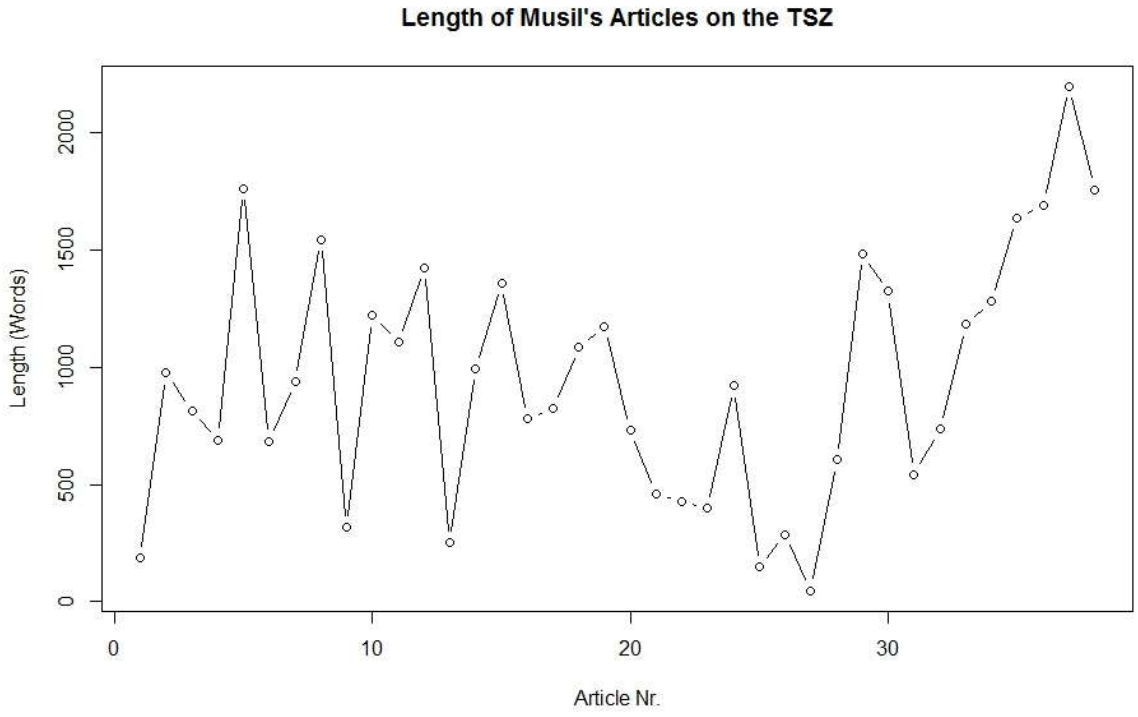


Figure 1. Word-length of the disputed Musil’s articles in the Soldaten-Zeitung

Reverse Binoculars: Combining Bibliographical Data and Topic Modelling in Research into Literary Culture

Maciej Maryl, Institute of Literary Research, Polish Academy of Sciences, maciej.maryl@ibl.waw.pl
Maciej Eder, Institute of Polish Language, Polish Academy of Sciences / Pedagogical University in Kraków, maciejeder@gmail.com

Introduction

This paper engages with the conference's main metaphor (reverse telescope) by proposing a combination of two distant-reading methods in a single study, in order to better account for the evolution of literary scholarship in Poland. Distant reading of literary journals has allowed numerous scholars to draw interesting conclusions about the forms, shapes, currents and dynamics of literary life (cf. Bode 2012, Long and So 2013, Goldstone and Underwood 2012). In the present study we combine two approaches – co-citation network analysis and topic modelling – in order to give a nuanced view of a Polish literary-studies bimonthly *Teksty Drugie* [Second Texts].

The first part is dedicated to the co-citation analysis based on metadata extracted from 1924 texts published between 1990 and 2012 (28,613 bibliographical records and 10,191 unique authors. It enabled the detection of 15 meaningful groups of authors interconnected by references they use. A subsequent part employs topic modelling to analyse 2553 full-text articles (11,3 million words) published between 1990 and 2014, which revealed thematic patterns pertinent to the journal, which are first discussed as an interconnected network and subsequently analysed in a diachronic perspective. Certain events (e.g. political transformation) seem to have influenced the topics discussed by literary scholars. It shows tight links between the humanities scholarship and questions pertinent to the society.

The discussion is dedicated to the "binocular view", i.e. using both perspectives in order to show the development of the literary journal and connections between the topics and authors. Such an approach allows for the multifaceted bird-eye view on the processes of literary scholarship.

Macroanalysis and Literary Scholarship

The availability of digitised full-text resources as well as bibliographical data in standard database format, opened, quite recently, a new chapter in sociology of literature, especially by revaluating empirical approaches and data-driven scholarship. The road to this "empirical turn" in literary scholarship has been paved by such scholars as Franco Moretti (2005, 2013) and Matthew Jockers (2013), who showed how empirical data like bibliographical records, annotations, title words, genre categorization etc. may help in generating new knowledge about literary periods. This approach gathered its momentum as other works exploring applicability of distant reading emerged. Due to the shortage of space we will name just a few that have most influence on this paper, dividing them, arbitrarily, into three research strands. Firstly, the use of bibliographical data for statistical inferences on literary processes, e.g. Bode's (2012) rereading of Australian literary history through the data from *AustLit* (*Australian Literary Bibliography*). Secondly, the study of author co-occurrences and mutual references, e.g. visualising literary

circles on the basis of bibliographic metadata by So and Long (2013). Thirdly, application of topic modelling to uncover pertinent issues in literary scholarship, e.g. Goldstone and Underwood analyses of the evolution of American literary scholarship on the example of *PMLA* (2012) and seven major literary journals (2014). In combining those approaches into a macroanalytical study of *Teksty Drugie* we also adopted the rationale introduced by the internet edition marking the 40th anniversary of *Signs,* literary journal dedicated to feminist criticism[1].

Aim

The aim of this study is to apply macroanalytical methods to trace the transformations of Polish literary studies on the example of *Teksty Drugie*. We hypothesise that the collection of papers published in a leading academic journal on literary scholarship can serve as a reliable approximation to chronological changes and/or breaks in Polish literary theory at the turn of the 20th century, as well as of the most pertinent topics and intellectual influences**.**

Material

Teksty Drugie is a Polish scholarly journal dedicated to literary scholarship. It has been published since 1990 by the Institute of Literary Research of the Polish Academy of Sciences. It focuses on literary theory, criticism and cultural studies, while also publishing articles by authors from the neighbouring disciplines (philosophy, sociology, anthropology). The journal publishes monographic issues, dedicated to particular topics or approaches within literary and cultural studies. All those features make it a good example for exploring the vicissitudes of Polish literary scholarship.

Bibliographic material consists of 28,613 references in 1924 texts, published between 1990 and 2012. Metadata was downloaded from Digital Repository of Research Institutes (RCIN) and transformed to fit the needs of the study. It should be noted that each reference to a particular article or book appears only once in metadata, no matter how many times this work was cited by the author. In order to increase comparability of the data, only names of the authors (or editors in case of collected volumes) were used. Overall there were 10,191 unique authors. More than two thirds of them (67.89%, or 6919) were cited only once and 18.66% (1902) appears more than twice. 385 (3.78%) authors were cited more than 10 times, and only 1% (103 people) received more than 30 references.

The textual corpus consists of the entire collection of papers published in *Teksty Drugie* (excluding letters, surveys, notes, etc.) in the years 1990–2014 (2,553 texts, 11,310,638 words). The material covering the years 1990–1998 was digitised, OCR-ed, and then manually edited, in order to exclude running heads, editorial comments, and so forth. Obviously, some textual noise – e.g. a certain number of misspelled characters – could not be neutralised. The material from 1999 onwards was digitally-born, but even though a small number of textual issues might have occurred. We believe, however, that distant reading techniques are resistant to small amounts of systematic noise (Eder, 2013).

Given the nature of Polish, which is highly inflected, lemmatization was necessary for a reliable processing of texts. The corpus has been lemmatised with LEM 1.0. (Literary Exploration Machine) developed by CLARIN-PL (see: Piasecki, Walkowiak, Maryl, under review)[2].

1 http://signsat40.signsjournal.org/

2 The authors wish to thank Dr. Tomasz Walkowiak (Wrocław Univeristy of Technology / CLARIN-PL) for his extensive help with cleaning and lemmatizing the corpus.

Method

To scrutinise the formulated hypothesis, we applied some standard visualisation procedures (for co-citation data) and topic modelling (involving the LDA algorithm) for textual content. Dynamic visualisation, supporting the interpretive process, was performed in Gephi (Bastian et al., 2009), mostly with the use of community-detection features: Louvain algorithm for computation (Blondel et al. 2008) and OpenOrd-layout for visualisation.

Topic modelling experiments were performed using a tailored script in the R programming language, supplemented by the package 'stylo' (Eder et al., 2016) for text pre-processing, and the package 'mallet' (McCallum, 2002) for the actual LDA analysis. A bimodal network of the relations between topics were produced using, again, Gephi. Other parameters used in the study included: a tailored stop word list containing 327 words (mostly function words, numerals, and very common adverbs), 100 topics extracted in 1,000 iterations, with the obvious caveat that the choice of the hyperparameters was arbitrary.

Results

Study 1. Subsequent trials showed that the best segmentation of the co-citation network is achieved with 27 groups (among which 12 were too small for analysis). Those groups were interpreted (see. table 1) and visualised as a network with 10,543 nodes and 20,952 edges (Figure 1.). The graph is bidirectional (authors referencing others may have also been referenced).

CLUSTER	NUMBER OF AUTHORS	% (N=10543)	TOPIC
A	2510	23.81%	Poststructuralism, philosophy, literature
B	1284	12.18%	Theory of literature and culture (structuralism, hermeneutics, sociology of literature)
C	928	8.80%	History of 20th.-century Polish Literature part 1
D	912	8.65%	History of 20th.-century Polish Literature part 2
E	748	7.09%	Modernism, culture and art.
F	683	6.48%	Romanticism
G	535	5.07%	Holocaust and testimony
H	439	4.16%	Literary criticism, comparative literaturę
I	423	4.01%	Neurosemiotics, darawinism
J	404	3.83%	Feminism
K	387	3.67%	Poetics, versology
L	387	3.67%	Postcolonialism
M	347	3.29%	Old-Polish Literature
N	293	2.78%	Formalism and Prague School
O	220	2.09%	Darwinism
OTHER	43	0.41%	[12 clusters with rare authors]

Table 1. Clusters based on co-citations in *Teksty Drugie*

Topographical distribution of those clusters (Fig. 1) allows for better understanding of relationships between them. Certain authors' position in the network could be explained through the proximity to others (those issues will be raised in more detail during the presentation, with dynamic graphs).

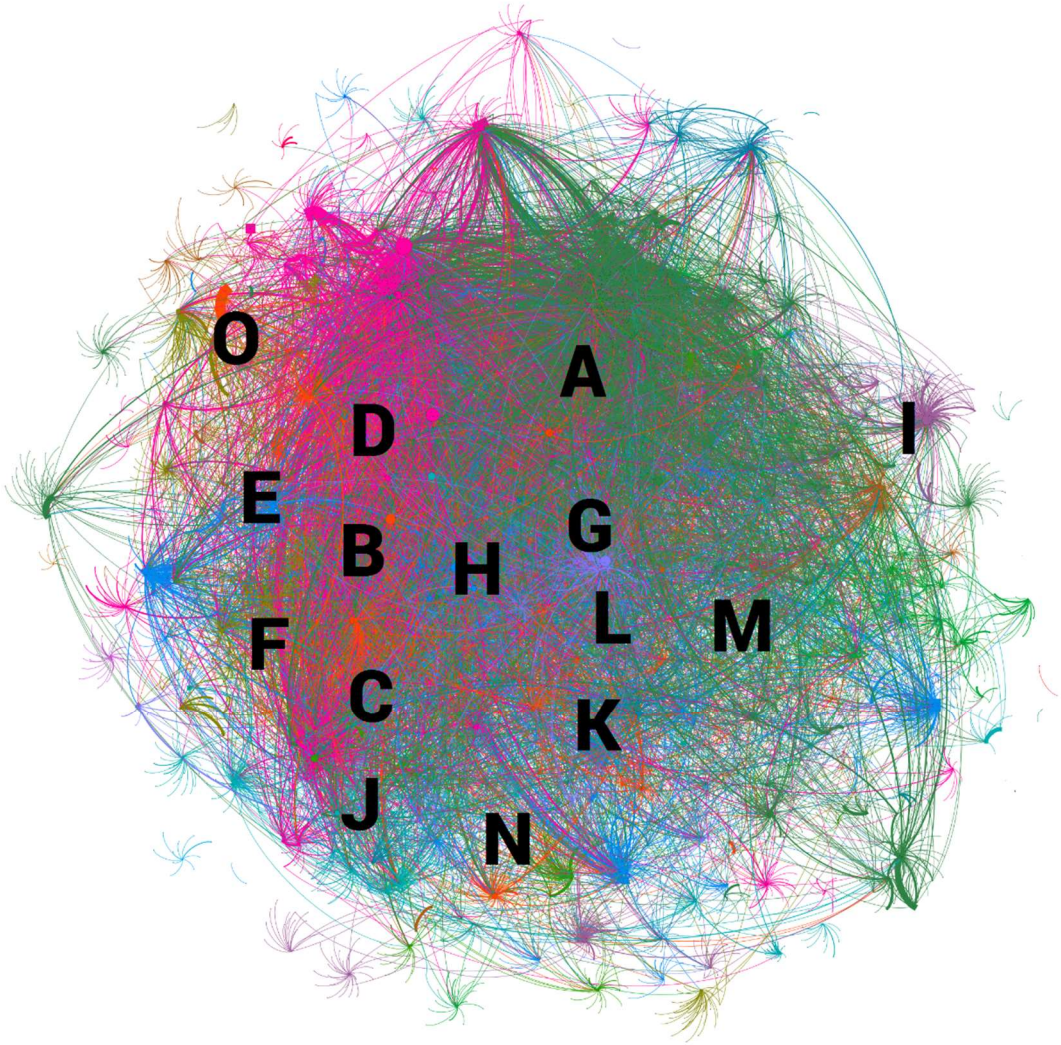


Figure 1, Co-citation network of *Teksty Drugie* with respective clusters.

Study 2. Firstly, we analysed and categorised the topics on the basis of their predominant words. The categories are as follows: literary theory (e.g. literature, fiction, text), poetics (e.g. verse, novel, short story, rhetoric) and methodological approaches (e.g. deconstruction, comparative literature, postcolonial studies, psychoanalysis); history of literature (e.g. romanticism, contemporary poets) and cross-cutting research themes (e.g. death, politics, literacy).

A thorough exploration of such models requires a topographical visualisation capable of showing the connections between various topics, which often share a key word (cf. Goldstone and Underwood, 2012). The network (Fig. 2) is too large to be adequately rendered in this paper, yet even without the knowledge about concrete topics presented, we may see (partly thanks to

ForceAtlas2 layout, which highlighted this feature) that groups of topics in our corpus are concentrically distributed[3].

This onion-like distribution allows us to distinguish between the central topics (i.e. those which appear in many different papers) and those who appear less often or sporadically and hence are not that well connected with other topics. For instance, almost perfectly in the geometrical centre of the network we may find topics and words pertinent to literary scholarship: literature, literary, comparative literature, national literatures, Jewish studies, fiction, together with some names of contemporary authors.

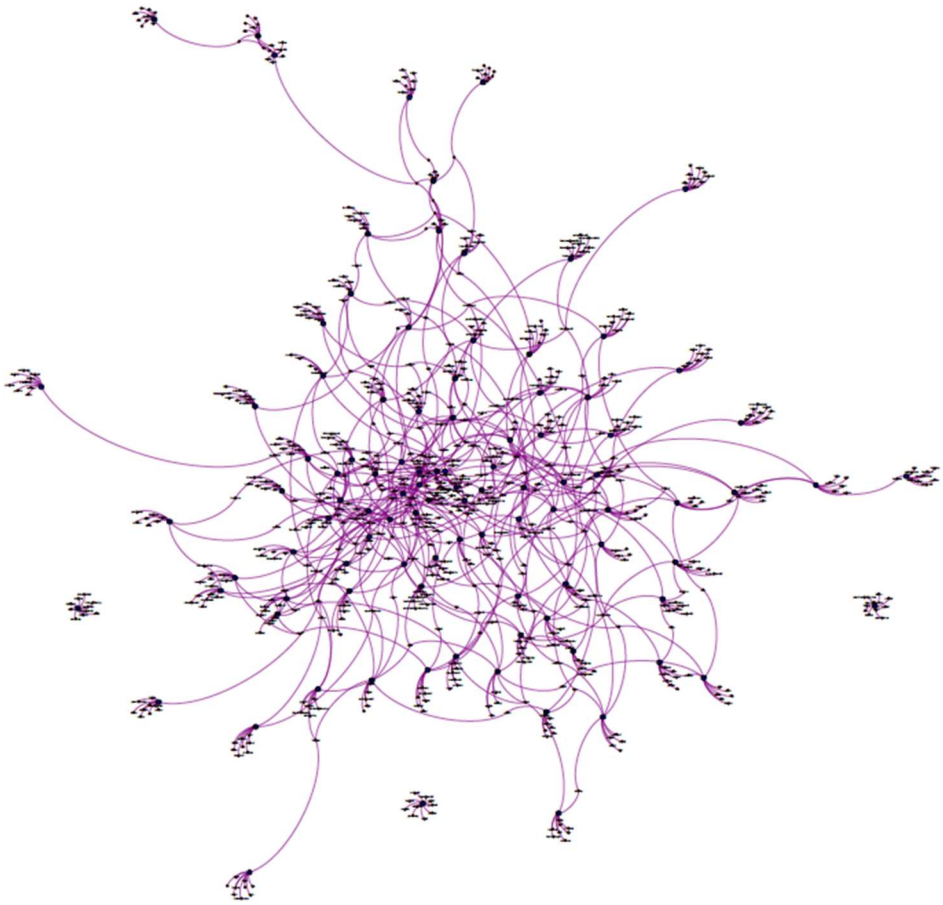


Figure 2. Relationships between topics in *Teksty Drugie*.

Discussion

In the discussion we will elaborate the details of both models and show how both methods may complement each other. Topic models help us understand why certain authors could be connected in the co-citation network. It could also allow us to check whether the dominance of a certain topic stems from the large number of scholars who pursue it, or rather, depends on the fact that a small group of authors has been publishing more often than others.

3 The image in the higher resolution could be downloaded here: https://figshare.com/s/c9e08a1d6b819d5c32a7

Bibliographical References

Bastian, Mathieu, Heymann, Sebastien. and Mathieu Jacomy. 2009. “Gephi: An open source software for exploring and manipulating networks”. *Proceedings of the Third International ICWSM Conference*. San Jose, pp. 361–62.

Blondel, Vincent. D., Guillaume, Jean-Loup, Lambiotte, Renaud and Etienne Lefebvre. 2008. “Fast unfolding of communities in large networks”. *Journal of Statistical Mechanics: Theory and Experiment* (10).

Bode, Katherine. 2012. *Reading by Numbers: Recalibrating the Literary Field*. London & New York: Anthem Press.

Blei, David M. 2012. “Probabilistic Topic Models”. *Communications of the ACM*, 55(4): 77–84.

Eder, Maciej. 2013. “Mind your corpus: systematic errors in authorship attribution”. *Literary and Linguistic Computing*, 28(4): 603–14.

Eder, Maciej, Rybicki, Jan and Mike Kestemont. 2016. “Stylometry with R: a package for computational text analysis”. *R Journal*, 8(1): 107–21.

Goldstone, Andrew and Ted Underwood. 2012. “What can topic models of PMLA teach us about the history of literary scholarship?” *Journal of Digital Humanities*, 2(1).

Goldstone, Andrew and Ted Underwood. 2014. “The quiet transformations of literary studies: What thirteen thousand scholars could tell us”. *New Literary History*, 45(3): 359–84.

Gross, Jan Tomasz. 2000. *Sąsiedzi: Historia zagłady żydowskiego miasteczka*. Sejny: Fundacja Pogranicze.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. University of llinois Press.

Long, Hoyt and Richard So. 2013. “Network science and literary history”. *Leonardo*, 46(3): 274–274.

McCallum, Andrew Kachites. 2002. “MALLET: A machine learning for language toolkit”. http://mallet.cs.umass.edu/.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Moretti, F. (2013). *Distant Reading*. New York: Verso Books.

Piasecki, M., Walkowiak, T., Maryl, M. (under review). “Literary Exploration Machine (LEM 1.0) - New Tool for Distant Readers of Polish Literature Collections”. Paper submitted to DH 2017.

So, Richard and Long, Hoyt 2013. “Network analysis and the sociology of Modernism”. *Boundary 2*, 40(2): 147–82.

Testi e computazioni: fondamenti conoscitivi delle edizioni digitali

Federico Meschini, Università degli Studi della Tuscia, fmeschini@unitus.it

Le edizioni elettroniche sono la trasposizione mediale di contenuti e strutture testuali che per cinquecento anni sono stati indissolubilmente legati al libro stampato. Date le caratteristiche dei calcolatori elettronici, e in particolare l'aspetto computazionale (Ausiello *et al.* 2003), questo nuovo ambiente per la rappresentazione testuale è un qualcosa di totalmente differente e unico rispetto al precedente paradigma, le cui funzionalità sono ancora da esplorare e comprendere appieno (Robinson 2013). Utilizzando un approccio interdisciplinare, che attinge a diversi settori, dalla critica testuale all'informatica teorica, passando per la biblioteconomia, questa relazione vuole concentrarsi sulla natura precipua delle edizioni elettroniche, analizzando da un diverso punto di vista aspetti fino ad ora considerati non rilevanti e accessori.

A un primo sguardo una difficoltà non da poco è relativa al posizionamento cognitivo delle edizioni digitali, in una zona trasversale alla tradizionale separazione tra discipline nomotetiche da un lato e idiografiche dall'altro (Renzi, Andreose 2009). La critica testuale tende molto verso quest'ultimo aspetto, e altrimenti non potrebbe essere, a causa della forte idiosincrasia che caratterizza ogni opera letteraria, sia dal punto di vista del contenuto, sia per ciò che riguarda le modalità di creazione e disseminazione. Il modello cognitivo sottostante nella ricostruzione filologica sembra essere l'abduzione, una forma logica introdotta da C.S. Peirce: partendo dalle conclusioni è possibile supporre, ma non essere certi delle premesse (Peirce 1909). Ciò nonostante, un assoluto relativismo, nelle pratiche editoria, legato solo al giudizio soggettivo del curatore, è stato da tempo superato, così da avere dei criteri e dei procedimenti scientifici in grado di essere riutilizzati in contesti simili (Stussi 2006). Come conseguenza i modelli sviluppati fino ad ora, come ad esempio il metodo Lachmann (Timpanaro 1963), o la teoria del *copy-text* di Bowers-Greg (Greg 1950), hanno la loro natura fondativa basata su criteri formali e nomotetici, basti pensare alla legge di maggioranza o alle differenze strutturali tra *accidentals* e *substantials*. C'è quindi una tensione interna alla critica testuale: da un lato una tendenza verso l'idiografia e dall'altro la necessità di una componente nomotetica. Basti pensare alla relazione tra la filologia romanza e la linguistica e il loro ruolo di mutuale definizione, ben rappresentato dalle "Mani che disegnano" di Escher (Hofstadter 1979). La relazione tra informatica e critica testuale può essere quindi considerata come l'evoluzione di questa dinamicità intrinseca.

Passando a un livello pragmatico, in un saggio di bibliografia testuale, Neil Harris cita una definizione di W. W. Greg, secondo la quale "un libro creato in tipografia è un oggetto fisico, e quindi il primo compito della critica testuale è quello di scoprire i suoi segreti come prodotto" (Harris 2006). Una conseguenza logica è che, quindi, al cambiamento del medium sottostante, questa stessa attenzione verso aspetti "materiali" debba essere mantenuta, e non solo per una curiosità intellettuale, ma come un requisito fondamentale, dato che sono strettamente legati agli aspetti funzionali. Se questi ultimi nel *codex* sono rimasti pressoché invariati, nel supporto elettronico sono in continua evoluzione, in quanto si basano su principi logico-matematici che permettono di modellare infiniti mondi possibili (Torvalds 2001). Il volersi concentrare su un approccio pragmatico potrebbe sembrare apparentemente in contrasto con il livello teorico descritto in precedenza; ne è però in primo luogo la sua incarnazione concreta, e costituisce quelli che sono i *codici tecnologici* di un'edizione: documenti XML, fogli di stile, framework di pubblicazione, basi di dati, template grafici, programmi lato client e lato server, motori di ricerca e indici, ontologie e tutto ciò che va a comporre la complessa architettura di quello peculiare

sistema informatico che è un'edizione. E, di conseguenza la loro importanza può essere paragonata ai *codici bibliografici*, la specifica fisicità dei documenti (McGann, 1983), essendo fondamentali per aspetti come l'interoperabilità e la conservazione.

Di conseguenza se si accetta la teoria di McGann, secondo la quale il libro è una macchina di conoscenza (McGann 2001), possiamo considerare allora l'edizione come un sistema di conoscenza, basata su delle specifiche macchine, e relativi meccanismi, sottostanti, ognuna di esse dotata di caratteristiche particolari, a seconda della tecnologia utilizzata. Fino a ora però la maggior parte delle riflessioni critiche si sono concentrare, per ragioni pragmatiche, sugli aspetti relativi alla trascrizione, che in un'ottica computazionale diventa una codifica testuale basata sui linguaggi di marcatura, come lo standard *de facto* promosso dalla Text Encoding Initiative (TEI Consortium 2014). Molto più scarso, ma non del tutto assente (Meschini 2008; Pierazzo 2011), è lo studio dell'altra estremità del processo editoriale: il meccanismo di pubblicazione, e di conseguenza, la forma dell'edizione stessa. Queste due estremità, codifica da un lato e pubblicazione dall'altro, essendo i passaggi più meccanici, sono quelli più vulnerabili ed esposti ai cambiamenti tecnologici, che ne modificano la natura sia epistemologica sia ontologica. (Mordenti 2001).

Scopo di questa relazione è quindi quello di spostare il focus dei numerosi e precedenti tentativi di analisi in relazione sia alla recente diffusione del *Computational thinking* (Wing 2006), e alle radici culturali dell'informatica, un approccio che inizia ad essere preso in considerazione anche nel settore delle *Digital Humanities* (Berry 2011), e, più in generale, nell'ottica del superamento della divisione tra le due culture (Burnard 2000) o, andando a ritroso, della separazione gutenberghiana dei saperi tra scienze umane e scienze esatte (Chandra 2014). Le edizioni digitali sono da sempre uno degli aspetti più rilevanti dell'informatica umanistica, per la loro storia, diffusione, e combinazione di riflessioni teoriche e applicazioni pratiche; data la loro importanza strategica, sia nel mondo accademico e culturale sia in quello editoriale, una migliore comprensione e divulgazione dei loro aspetti computazionali potrebbe aiutare a ridurre lo spessore del muro che attualmente divide le discipline umanistiche da quelle esatte, aiutandole così a uscire dalla nicchia sempre più stretta in cui vengono al momento collocate, e recuperando quindi il ruolo di primo piano cui hanno diritto. A un livello più generale questo avvicinamento disciplinare sottolineerebbe come sempre di più nel XXI secolo le due abilità fondamentali saranno narrazione da un lato e computazione dall'altro, insieme a tutte le loro possibili interrelazioni (Thaller 2014; Coleman 2015).

Riferimenti Bibliografici

Ausiello, G. D'Amore F. Gambosi G. 2003. *Linguaggi, modelli, complessità, Scienze e tecnologie informatiche*. Milano: Franco Angeli.

Berry, D. M. 2011. "The computational turn: Thinking about the digital humanities." *Culture Machine* 12.0:2.

Burnard, L. 2000. "From two cultures to digital culture: the rise of the digital demotic". <http://users.ox.ac.uk/~lou/wip/twocults.html>.

Chandra, V. 2014. *Geek Sublime: The Beauty of Code, the Code of Beauty*. Graywolf Press, Minneapolis, USA.

Coleman, D. 2015. "Count or Die: Why the Humanities Need Numbers to Survive". *Aspen Ideas Festival 2015*. <http://www.aspenideas.org/session/count-or-die-why-humanities-need-numbers-survive>.

Harris N. 2006. "Filologia dei testi a stampa". In Stussi A. (ed.). *Fondamenti di critica testuale*. 2nd ed. Bologna: Il Mulino.

Hofstadter, D. R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*, New York: Basic Books.

Greg, W. W. 1950. The Rationale of Copy-Text. *Studies in Bibliography* 3 (1):19.

Greetham, D. C. 1994. *Textual Scholarship: An Introduction*. New York: Garland Publishing, Inc.
McGann J. 1983. *A Critique of Modern Textual Criticism*. University of Virginia Press.
McGann J. 2001. *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave Macmillan.
Meschini, F. 2008. “Mercury ain’t what he used to be, but was he ever? Or do electronic scholarly editions have a mercurial attitude?”. *International seminar of Digital Philology*, Edinburgh, 2008.
Mordenti R. 2001. *Informatica e critica dei testi*. Roma: Bulzoni.
Peirce, C. S. 1909. Existential Grapsh. Cambridge, MA: Houghton Library.
Pierazzo, E. 2011. "The Role of Technology in Scholarly Editing". TEI Members Meeting 2011, Würzburg, http://epierazzo.blogspot.it/2011/10/role-of-technology-in-scholarly-editing.html.
Renzi, L., Andreose, A. 2009. *Manuale di linguistica e filologia romanza*, *Manuali*. Bologna: Il Mulino.
Robinson, P. 2013. “Towards a Theory of Digital Editions.” *Variants* 10.
Stussi, Alfredo, ed. 2006. *Fondamenti di critica testuale*. 2 ed. Bologna: Il Mulino.
TEI Consortium (a cura di). 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. 2.6.0*. Thaller, M. 2014. "Are the Humanities an endangered or a dominant species in the digital ecosystem?" *AIUCD 2014*, Bologna, Italy.
Torvalds L. 2001. *Just For Fun*. New York: Harper Business.
Timpanaro S. 1963. *La genesi del metodo del Lachmann*, Firenze: Le Monnier.
Wing, J. M. 2006. "Computational thinking". *Communications of the ACM* 49 (3):33-35.

The *Index Thomisticus* as a Big Data Project

Geoffrey Rockwell, University of Alberta, Canada, grockwel@ualberta.ca
Marco Passarotti, Università Cattolica del Sacro Cuore, Milan, Italy,
marco.passarotti@unicatt.it

The Digital Humanities (DH), as Rob Kitchin reminds us, have always been interested in the building of infrastructure for research (Kitchin 2014, Loc. 222 of 6164). Imagining how emerging technologies could first be tried on humanities problems and then scaled up to infrastructure for others to use has been one of the defining features of the field. Such experimentation began with Father Busa's *Index Thomisticus* (IT) project (Busa 1980; Winter 1999; Busa 1974-1980). Which was probably also the largest digital humanities project of all time, even though the outcome might, by today's standards be considered "small". The project lasted 34 years and at its peak (1962) involved as staff of 71 persons all housed in large ex-textile factory in Gallarate. For their time they were dealing with big data, really big data. If we want to understand what is involved in scaling up to big infrastructure we should look back to the beginnings of the field and the emergence of big projects like the IT. This paper will therefore look at the Busa's project as a way to think through big projects. Specifically we will,

- First, briefly discuss the historiography of the IT project in particular, and DH projects in general. How can we study projects as bearers of ideas? What resources do we need/have?
- Second, look at specific aspects of the project that shed light on DH projects in general. We will look at how the project was conceived, the development of methods/processes, and how communications were managed;
- Finally, reflect on what lessons the IT project has for us at a time when big data has become an end in itself. What can we learn from Busa's attention to data in the face of the temptations of automatically gathered data?

Historiography of Projects

If one believes, as we do, that projects are a form of distributed cognition that create meaning, we should then ask how they work at generating and bearing meaning. Despite a lot of attention being paid to how to manage projects, there is little about how digital projects can be studied as bearers of meaning. The IT project has the advantage of being extremely well documented, and can therefore serve as a case study right at the threshold between traditional concording projects and digital humanities projects. In this paper we will briefly make the case for the study of projects and then make the case for the study of the IT as a paradigmatic project due to its influence and the wealth of materials in the Busa Archives housed at the Università Cattolica del Sacro Cuore in Milan[1]. The Archives cover a time span of around 60 years (from the beginning of the 1950s until 2010) and contain different kinds of materials, which can be summarized as follows:

- Personal materials of Busa, like academic certificates, ordination details, photocopy of his identity card etc.;
- Documentation about conferences, seminars and workshops attended by Busa, including materials used to prepare his contributions, versions of the text of talks given by Busa,

1 Documents like Figure 1 shown above are kindly made available upon request under a Creative Commons CC-BY-NC license by permission of CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy. The image is contained in the *Busa Archives*, held in the library of the same university. For further information, or to request permission for reuse, please contact Marco Passarotti, at marco.passarotti@unicatt.it, or by post: Largo Gemelli 1, 20123 Milan, Italy. For information on the archives contact the archivist Paolo Senna at paolo.senna@unicatt.it .

programs of events, handouts distributed by other speakers (with handwritten notes by Busa) and various materials related to practical matters (flight tickets, Visa bills, hotel reservations etc.);
- Press articles in the Italian and international media on Busa and his research;
- Professional correspondence between Busa and his contemporaries (in academia, cultural heritage and libraries, administration, industry, politics, religious organizations etc.) in Italy and abroad;
- Personal correspondence between Busa and close colleagues and friends in Italy and abroad;
- Materials relating to particular phases of the *Index Thomisticus*, like print outs, punch cards, tapes, budgets, proofs etc.;
- Photographs, each enhanced with the date and the names of the persons pictured;
- *Opera Omnia* of Busa [external to the *Archives*]: one copy of each publication by him.

Following a specific request by Busa, the *Archives* still retain their original organization in sections (and related boxes) arranged by Busa. The *Archives* are freely accessible by sending a request to the Library of the Università Cattolica. The irony is that very few Italian scholars seem to avail themselves of the archive.

The *Index Thomisticus* as Project

The heart of this paper is an examination of some aspects of the IT project that can shed light on DH projects in general and how they have evolved. We will first look at how Busa conceived of the project and how that conception evolved as he developed collaborations with others and the technology changed. Then we will look at the data entry, indexing and concording process to understand the very different role of scholars and data (Busa 1951; Busa 1958). There was a much greater involvement of scholars in defining what was "given" as "data" than what happens today. Scholars also intervened in the processing to identify the "Entry Cards" or headwords of the concordance (see Fig. 1). Unlike modern big data or distant reading practices, scholars and operators were intimately involved in curating the data.

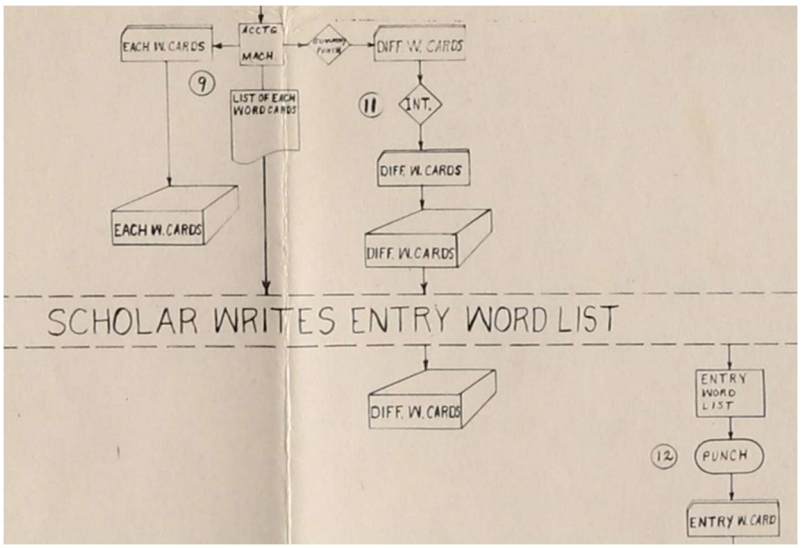


Fig. 1: Detail from the 1952 "Flow Chart" from the Busa Archive

Thirdly, we will talk about how Father Busa and others managed communications (Jones 2016). Today communications would be mostly over email and such email would not be archived.

In Busa's days the correspondence was carefully copied and filed. We can see how important communications were for Busa, not only with experts, but also with potential supporters who could write IBM on the project's behalf.

The Legacy of Father Busa to Reconcile the Two Humanities

In conclusion we will discuss the the legacy of Father Busa as it applies to simplistic ideas about big data analytics in the humanities. The project of the *Index Thomisticus* was started because Busa needed to find a way to deal with a large amount of textual data. This need was motivated by his rigorous approach to empirical evidence. His careful attention to data comes through clearly in the motto "aut omnia aut nihil" ("all or nothing"), which characterized all of Busa's scientific production.

Busa believed the greatest danger lay in considering Computational Linguistics (and DH, too) not as a discipline aimed at doing things better, but rather as a tool to do things faster. He was not satisfied by computer generated quick KWICs (pun intended). Materials in the Archives show an incredible attention to detail from page layout to fonts to language. He feared that the computational linguists (as well as computing humanists) of the third millennium would cease caring for the human data (which should be their bread and butter) and lose the humility to check each analysis, preferring instead to process huge masses of texts quickly and approximately without even reading a line.

Today, we see this fear of Father Busa coming true in opportunistic projects. Although there are projects in the digital humanities and social sciences dealing with enormous amounts of textual data, it turns out that often these aren't carefully checked. Availability of data like social media data replaces the careful gathering, enrichment and curation of appropriate data. The result is projects that don't move from information to knowledge. By contrast, Busa was convinced that striving to formalize language for computing represents an extraordinary method to get to a detailed knowledge of it. He argued that preparing textual data for computer analysis requires the scholar to dedicate more time (and effort) than that required for non-computer-aided research. This is clear if we look at the detailed flowchart that Tasman prepared for the building of word concordances for the *Index Thomisticus* (see Fig. 1 above, dated 1952; see also Tasman 1957).

Today many projects in the digital humanities and other fields are tempted by the big data at hand. This laziness alienates "the two Humanities" namely the "Digital" Humanities and "Traditional" Humanities. Instead, DH should bring the traditional humanist attention to the human record to "data" - that which is "given" to interpretation. What would it mean to return to a rigorous, objective, ethical and, in a word, scientific approach to data? In our presentation we will make the case for best practices from the humanities, which basically consist in the motto "Love Thy Data!" and can be summarized in the following, only apparently obvious, advice:

- Be aware of the properties of the data you are going to process and analyze;
- Check the quality of a (digital) resource (with clearly stated and replicable evaluation criteria) before using the data it provides;
- Check (with clearly stated and replicable evaluation criteria) the results of an automatic analysis before making any inference based on them.

Bibliographical References

Busa, R. (1951). S. Thomae Aquinatis Hymnorum Ritualium Varia Specima Concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate. Milano, Fratelli Bocca.
Busa, R. (1958). The Use of Punched Cards in Linguistic Analysis. Punched Cards: Their Applications to Science and Industry. Eds. R. S. Casey, J. W. Perry, M. M. Berry and A. Kent. New York, Reinhold Publishing: 357 - 373.
Busa, R. (1974-1980). Index Thomisticus. Stuttgart-Bad Cannstatt, Frommann-Holzboog.
Busa, R. (1980). "The Annals of Humanities Computing: The Index Thomisticus." Computers and the Humanities. 14:2: 83-90.
Jones, S. E. (2016). Roberto Busa, S. J., And the Emergence of Humanities Computing: The Priest and the Punched Cards. New York, Routledge.
Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. Los Angeles, SAGE.
Tasman, P. (1957). "Literary Data Processing." IBM Journal of Research and Development. 1:3. 249-256.
Winter, T. N. (1999). "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance." The Classical Bulletin. 75:1. 3-20.

Palantir: Reading the Surveillance Thing
Critical Software Stories as a Way of the Digital Humanities

Geoffrey Rockwell, University of Alberta, grockwel@ualberta.ca
Domenico Fiormonte, Università Degli Studi Roma Tre, fiormont@uniroma3.it

Introduction

In 2011 Anonymous hacked into the server of a security company HBGary and downloaded 70,000 documents including a PowerPoint deck titled "The WikiLeaks Threat" (Coleman 2014, Loc. 2998). This deck showed how HBGary was teaming up with other companies including Palantir to propose disrupting WikiLeaks. A commercial security firm was pitching a bank to get a contract to use surveillance software and cyber attacks to delegitimize WikiLeaks! Tools and techniques that had previously been developed and used by government agencies for counterterrorism were now being deployed by the commercial sector to interfere rhetorically in the public sphere. And that is what this paper is about, the importance to the humanities of analytical tools and services of companies like Palantir. Specifically in this paper we will:

- Discuss what we know about the Palantir analytical tools and the context of services they offer as an example of commercial surveillance tools.
- Talk generally about the critical reading of analytical tools in the Digital Humanities.
- Conclude with by positioning such critical reading as a challenge to naive big data claims.

Palantir and Commercial Surveillance

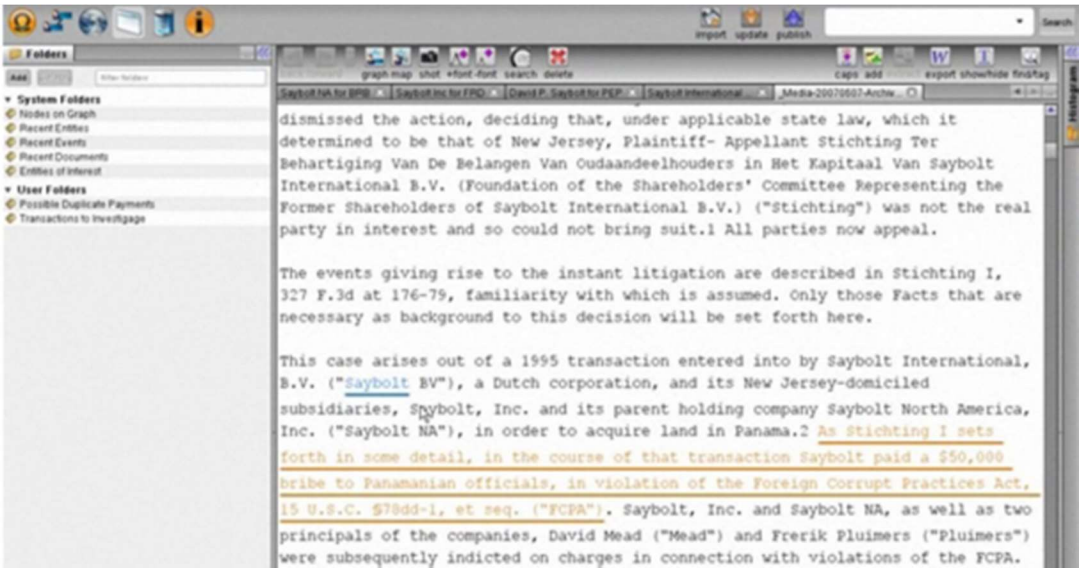


Fig. 1: Slide showing Palantir full text drilling environment

With Edward Snowden's massive leak of intelligence documents we now have a wealth of information about the tools that Five Eyes intelligence services have created (Rockwell & Sinclair 2016b). What is less well understood are the commercial tools that are being sold to intelligence services and other organizations including repressive regimes (Hern 2015). What we do know comes from cases where surveillance companies have been hacked themselves like the Anonymous hack and the later hack of the Italian company Hacking Team (Hern 2015). In this first part of the paper we will use the various sources to reconstruct what Palantir does as a way

of reading a tool meant for the analysis of information about networks of people which, after all is what we do too.

Reading the Surveillance Things

The humanities have traditionally concerned themselves with the interpretation of human expression in its different forms. Software like Palantir can be thought of as a tool or as an instantiation of a hermeneutic. We can read tools like Palantir the way we read other instantiations of ideas about interpretation (Rockwell & Sinclair 2016a). In the second part of this paper we will argue that digital humanists are well positioned to critically read the traces of the tools being used to manage public discourse. We will discuss the approaches one can take, drawing on humanities traditions, to reading the tools as communicative frames.

We will also discuss what we as digital humanists need to learn to study software (and infrastructure) critically. One thing we need to learn to do is to read the forms of information exchange common in business and security circles. This includes the sorts of commercial literature used to promote analytical tools, the forms of documentation shared, and above all PowerPoint decks, which have become a currency in intelligence circles. None of these are really great literature; they are genres generally ignored in the humanities, but they are what we have as traces of surveillance. As modern-day paleographers we need to adapt our hermeneutical methods to these commercial genres. (For a collection of such files see the WikiLeaks Spy Files <https://wikileaks.org/spyfiles/about/>).

In this paper we will argue that these tools build representations of people and their stories. We can consider tools like Palantir as story-telling tools and some of the documentation for Palantir treats them this way. The "literary" side of these surveillance software is their ability to build compelling stories out of data and the mystique of "big data". We will argue that the digital humanities therefore has a responsibility to think through what and how they tell stories that affect us.

Is Data Knowledge?

When thinking about big data we can talk about reading communities. Reading makes meaning, but it does so in networked communities of readers. What do these tools say about their intended communities? Can we create new and critical communities for these tools?

We will conclude by talking about the "original sin" of big data, i.e. the information retrieval paradigm that treats stories as data and data as a resource to be mined. Is information / data extraction "neutral"? And what kind of knowledge constitutes the 'information' extracted from data? Can we recover the role of interpretation or theory (Anderson 2008) in the face of commercial interests? Big data challenges the centrality of human interpretation and the associated idea of the "model" which was at the centre of the debate around humanities computing up to the mid 2000s. (Orlandi 1997). We believe it is doubtful that knowledge can exist without interpretation which is why we need to not only interpret big data, but also the tools of interpretation like Palantir.

Bibliographical References

Anderson, C. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine*. 16:7.

Coleman, G. 2014. *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous*. Kindle Edition. London, Verso.

Gobry, P.-E. Feb. 11, 2011. “Palantir Apologizes For Its Plan To Crush Wikileaks.” *Business Insider*. <http://www.businessinsider.com/palantir-wikileaks-apology-2011-2>
Gorbry, P.-E. Mar. 11, 2011. “REVEALED: Palantir Technologies, The Secretive $735 Million Tech Security Company Helping Hedge Funds And Governments.” *Business Insider Australia*. <http://www.businessinsider.com.au/palantir-technologies-revealed-2011-3#heres-what-palantir-looks-like-when-you-start-it-up-1>
Febelfin, Sept. 24, 2012. “Amazing mind reader reveals his ‘gift’”. *YouTube*. <https://youtu.be/F7pYHN9iC9I>
Hern, A. July 6, 2015. "Hacking Team Hacked: Firm Sold Spying Tools to Repressive Regimes, Documents Claim." *The Guardian*. <https://www.theguardian.com/technology/2015/jul/06/hacking-team-hacked-firm-sold-spying-tools-to-repressive-regimes-documents-claim>
Orlandi, T. 1997. "Informatica, Formalizzazione E Discipline Umanistiche." *Discipline Umanistiche e Informatica. Il Problema Della Formalizzazione*. Roma. 7-17.
Rockwell, G. and S. Sinclair. 2016a. *Hermeneutica: Computer-Assisted Interpretation in the Humanities.* Cambridge, MA: MIT Press. <http://hermeneutic.ca>
Rockwell, G. and S. Sinclair. 2016b. “Watching out for the Olympians! Reading the CSEC Slides.” *Information Ethics and Global Citizenship.* Eds. T. Samek and L. Shultz. McFarland & Company. Forthcoming in 2016.
Ruskin, G. 2013. Spooky Business: Corporate Espionage Against Nonprofit Organizations, Center for Corporate Policy. Report. <http://www.corporatepolicy.org/spookybusiness.pdf>

Digicraft and pensiero sistemico: le prospettive delle DH

Enrica Salvatori, Università di Pisa, enrica.salvatori@unipi.it

In the AIUCD (*Associazione per l'Informatica Umanistica e la Cultura Digitale*) conference held in Bologna in September 2014 Manfred Thaller wondered «Are the Humanities an endangered or dominant species in the digital ecosystem?» (Thaller 2014, see also Thaller 2012). The answer was not simple nor linear and directly involved the Digital Humanities (DH from now on) as a disciplinary and research field that was still poorly and controversially defined. Thaller wondered how, providing certain conditions, DH could bring out the Humanities from the Indian reserve where they are now confined. In particular DH specialists should:

1) conceive of themselves as researchers and not as conversationalists;
2) strive for a vision;
3) change the epistemology of the Humanities;
4) drive technology and not be driven by it.

A few months after the conference Serge Noiret wrote on Digital History (one of my personal fields of research) trying to clarify what actually characterizes this subject within the wider field of DH, and - within the Digital History itself - what is the specific task of the Digital Public History (Noiret 2015; see also Robertson 2014). I place Noiret's article completely under the first point of Thaller's list: Digital History and Digital Public History are clearly seen as areas of research and not merely as new forms of communication of old disciplines. Moreover he answered to items 2 and 3 too, essentially proposing a more accurate taxonomy of DH. In a way, he seems to answer Thaller's question with a more accurate definition of some components of the meta-discipline itself.

I do not want to linger in this paper on the definition of DH as a whole nor of its components: so many authors in recent years have-debated to define what someone thinks to be (or could be) a discipline and others a research or work field (McCarty 2005; Svensson 2010). Every exercise of definition of a "new" area of research is, of course, useful, but at the same time it is potentially frustrating and risky. Frustrating because, as many authors and research centers have declared, despite its now long history, the DH is still an emerging field, and as well as an open, multifaceted, ever-changing one; risky, because each taxonomy of knowledge unavoidably builds walls and fences that encase the knowledge itself in a series of sterile boxes. This could be, in my modest opinion, the risk in Noiret's vision. It's more important to go beyond a possible but also difficult definition of DH and their several sub-disciplines, focusing our attention on items 2 and 3 of the Thaller list instead, namely on the need to have our own vision and on the importance to characterize DH in terms of the emerging changes of method in our daily research.

In particular I will try to connect the concepts expressed by both scholars, looking on the one hand to the recent history of DH in Italy (i.e. degree courses, associations, meetings) and on the other hand to my own research projects at the University of Pisa, especially inside the DH course degree (https://www.unipi.it/index.php/ects/ects?ects_id=IFU-L) and within the Digital Culture Laboratory (http://labcd.humnet.unipi.it/). Starting from some specific cases I wish to reason on one possible "vision,, of the DH.

It is not easy to give an overview of the development of DH in Italy after the "heroic" years of the late twentieth century (Adamo, Gregory 2000; Orlandi 2012), due to the excessive fragmentation of the initiatives that have given birth to centers, workshops, laboratories, regular courses, postgraduate courses, and so on (Ciotti 2014). Some graduate programs born in the begiining of the new century had a more or less long life, partly as a result of the continued reforms of the university rules. Currently the only full course of *Informatica Umanistica* (Humanities Computing*)* with BA and MA degree is in Pisa, but there are BA and MA degrees with different names in other Italian universities, such as Viterbo, Venice or Padua; there are alto several specialization courses after University, a sign that he feels a strong need for training in the sector, although the formalization of this bisdogno is hampered by grid reference standard

Inside the course degree in Pisa I will focus very briefly on some projects.

For what concerns *Epigraphical Studies, Public History and Education*:

Epigrapisa: A re-reading partly driven and partly spontaneous of the epigraphic messages left over time in a city.

Teaching (Digital) Epigraphy: a novel education experience in teaching students to transcribe and interpret Roman inscribed lead tags, using a Digital Autoptic Process (DAP) in a Web environment.

Pisa e l'Islam and Pisan Romanesque meets Contemporary America: two examples of historical web dissemination with a reasoning about both the potential and the limits of the medium to involve the audience..

In the area of *Digital Public History*:

Tramonti. Itinerari tra generazioni lungo i crinali della Val di Vara a complex project aimed at enhancing the cultural heritage of an Italian rural valley through the active participation of residents.

In the field of Digital Editions:

Codice Pelavicino Digitale: the digital edition of a medieval manuscript built to provide all services of the digital world and to invite the readers to actively participate..

By shortly describing these project I will not try to figure out what distinguishes them from each other, but, on the contrary, what characterizes all of them as Digital Culture projects and what they tell us about a possible vision of DH:

- they are digital;
- they are inevitably and necessarily interdisciplinary;
- they are open;
- they were built in a kind of new Renaissance workshop, a digital craft (DIGICRAFT).

They are digital. This may seem trivial but it is not. These are projects "born digital" not because the digital world offers the most useful tools to achieve the same purpose in relation with the "real" world, but because they could not exist outside the incredible interaction between real and digital world that it is now our life. They are digital because they might not otherwise exist.

Interdisciplinarity is compulsory. DH is a field unavoidably and profoundly interdisciplinary and we have to deal with each project as a complex set of activities and skills that crosses, by its true nature, several fields; this change of practice and approach implies by itself a methodological revolution, because it requires an organization of work similar to a Renaissance workshop (a DIGICRAFT), with an articulated division of labor in relation to several levels of skills, where

education and training could be provided by the same learners, coordinated by a strong and mature central idea.

Openness is a result and a choice. Working in a multidisciplinary team built upon research and with different tools, sustainability requires using open source tools, sharing data between individuals and giving everything to the public. Then Openness is a natural result, even it is also an ethical, political and philosophical choice as the Digital Manifesto 2.0 says: “the digital is the realm of the open, open source, open resources".

A DIGICRAFT. In a Renaissance workshop it was possible to produce different objects: statues, paintings, goldsmith or less valuable coroplastic objects. Each handwork was a “project” that included on the one hand a strong artistic and cultural vision (meaning, style, function, purpose, style) and on the other hand a complex set of different techniques made by different workers with different levels of capacity. The owner of the shop (or the head-artist) had not necessarily to know each technique as an expert, but his employees could in many ways be superior and the various members of the workshop could learn from each other. The owner had to keep the team together with a well clear idea of the work itself.

Likewise a DIGICRAFT could work (and actually does in our LabCD in Pisa) in the same way: each project is taken over as an interdisciplinary complex object that requires specific skills and different but profoundly related competences. The basis (first phases) of the work are often composed by students of the Bachelor and Master's degree in DH, who work, in the labCD, as interns or undergraduates. The work is directed by one “manager” but followed at various stages by experts, who assign specific tasks to the students, always ensuring an active connection among everyone in the team through the usual or more useful collaborative tools. While the work goes on, often happens that some student acquire, in a particular technique or phase, a greater capacity and knowledge than the others and then he/she becomes able to propose substantial changes in the work chain. The manager is not required - nor humanly could - to know every aspect in depth, nor to be fully aware of all the problems related to it, or to master each technique: however, he/she must be able:

- to see always clearly the aim and the nature of the work;
- to communicate effectively with everyone in the team.

A “digicraft” is anywhere on a DH project teachers and students exchange knowledge and leverage this interaction to offer innovative and effective solutions, combining the theoretical reasoning with practices and skills. This is possible only if the manager and the team share a common strong vision of what a DH project is, embracing a "systemic" or “organic” or “holistic” thinking of DH itself.

The core of DH is unitary and lies in the conviction that the digital turn has permeated every aspect of our lives as people and scholars modifying them deeply.

In the 70s of XXth century has increasingly gained ground a vision of Humanities Computing that kept almost unchanged the traditional disciplines within their rigid internal divisions and distinguished the humanist from the expert in information technology, hoping and promoting a dialogue between the two main areas (still in Fusi 2011, I, p. 1-2). Today this position is no longer sustainable. The web in first place and the web 2.0 in the second (but also the Big Data emerging field as well as the Data Visualization tools) have slowly but surely changed the research landscape especially demolishing the barrier between tools, methods and ways of sharing. We are obviously still in a transitional phase. Highly specialized sub-areas remain (and also in the future will exist) and obviously several scholars strive to better define the old / new digital disciplines (digital history, digital philology and so on), but there is also a complementary phenomenon pointing to an inclusive and unitary vision of DH.

From the perhaps limited but interesting Italian observatory I believe this change has affected both the terminology used in the establishment of centers, associations and degree programs (AIUCD, Digital Humanities degree, Digital Cultural Heritage, Arts and Humanities School), both the organisation of courses and meetings.

It 's more and more widespread the awareness that we are a new type of scholar (and graduate, and PhD), the digital humanist, someone who has a mixed formation, an open mind, is able to master both languages and the main methodological issues of the two areas without considering one serving the other.

In doing so, we need to maintain the epistemological strictness that each discipline involved in the DH has developed over time: commingling does not mean carelessness or inaccuracy; but in the same time we have to claim the change or the changes in each methodology in order to build a new global epistemology.

In order to do this the digital humanist has to embrace a "systemic" or "organic" or "holistic" thinking of the humanities, leave the enclosure of the academic fields and get away from the temptation to create an old-new rigid taxonomy.

Speaking of "systemic" or "organic" or "holistic" thinking / view, I refer to the epistemological approach that has emerged in some areas of the research over the past thirty years and which tends to oppose the reductionist approach flourished since the seventeenth century onwards and imposed in almost all sectors of the so-called "hard sciences" (Capra-Luisi 2014). As we know reductionism believes that studying in depth a peculiarity of a phenomenon and understanding it completely it will be possible, by progressive addition of discoveries, illuminate the entire system. The reductionist approach has been, as we know, the basis for the scientific revolution of the modern age, but it also led in the eighteenth and nineteenth century to an exasperate fragmentation of the fields of scientific research. This phenomenon has also heavily influenced the Humanities,often creating absurd barriers and hyper-specialized languages, that have closed researches in several walled gardens. I believe that this long wave has exhausted its strength and that precisely the DH can reverse the trend. Now a new methodological approach have arisen alongside the reductionist thinking, considering the "system", the "whole", something more and different than the sum of its components. The "systemic thinking" reasons in terms of relationships, networks, patterns of organizations and processes; it proposes a change of paradigms: from the vision of the world as a machine to the world as a network; it takes account of the fundamental interdependence of all phenomena.

This change of paradigms could and should affect the DH as well for the reasons listed above, promoting a systemic view of this meta-discipline and therefore pushing Digital Humanists to deeply transform the old practice of work.

Bibliographical References

Adamo G., Gregory T. (2000), Informatica Umanistica, Enciclopedia Italiana - VI Appendice (2000), Milano, Treccani, http://www.treccani.it/enciclopedia/informatica-umanistica_(Enciclopedia-Italiana)/

AIUCD Associazione per l'Informatica Umanistica e la Cultura Digitale, http://www.umanisticadigitale.it/ (accessed 27 February 2016)

Boonstra O., Breure L., Doorn P. (2006), *Past*, *Present and Future of Historical Information Science,* Amsterdam

Capra F., Luisi P.L. (2014), *The Systems View of Life: A Unifying Vision,* Cambridge; italian edition *Vita e natura. Una visione sistemica, S*ansepolcro

Ciotti F. (2014), Digital Humanities in Italy and their role in DARIAH research infrastructure, *Leggere, scrivere e far di conto.* Informatica Umanistica e cultura digitale (text of the keynote at

the IV DARIAH VCC meeting in Roma, on 17 September 2014), https://infouma.hypotheses.org/244
The Digital Humanities Manifesto 2.0 http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf
Fusi D. (2011), *Informatica per le scienze umane,* I, Elementi, Roma
McCarty W. (2005), *Humanities Computing,* London-New York
Noiret S. (2015), Digital Public History: bringing the public back*, Public History Weekly*, 3:13 http://public-history-weekly.oldenbourg-verlag.de/3-2015-13/digital-public-history-bringing-the-pu
blic-back-in/
Orlandi T. (2012), Per una storia dell'informatica umanistica, Dall'Informatica umanistica alle culture digitali. Atti del convegno di studi (Roma, 27-28 ottobre 2011), in memoria di Giuseppe Gigliozzi, a cura di Fabio Ciotti e Gianfranco Crupi, Quaderni Digilab, vol. 2, http://digilab-epub.uniroma1.it/index.php/Quaderni_DigiLab/article/view/18/16
Robertson S. (2014), The Differences between Digital History and Digital Humanities, *CHNM Blog Pos*t (23/1/2014), http://drstephenrobertson.com/blog-post/the-differences-between-digital-history-and-digital-humanities/ (accessed 27 February 2016)
Salvatori E. (2015), Il patrimonio genetico della storia digitale (e le nostre paure), *Appunti di viaggio tra il medioevo e la cultura digitale* (05/03/2015), http://esalvatori.hypotheses.org/211 (accessed 27 February 2016)
Salvatori E. (2015), L'identità dell'Informatico Umanista e la visione sistemica*, Appunti di viaggio tra il medioevo e la cultura digitale* (23/01/2015) http://esalvatori.hypotheses.org/204 (accessed 27 February 2016)
Svensson P. (2010), The Landscape of Digital Humanities, *Digital Humanities Quarterly*, 4/1
Thaller M. (2012), *Controversies around the Digital Humanities: An Agenda*, *Historical Social Research / Historische Sozialforschung*, 37/ 3: 7-23 http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html
Thaller M. (2014), *Keynote address* - AICUD 2014, http://aiucd2014.unibo.it (accessed 27 February 2016)

Approcci grafici all'analisi di corpora testuali

Eva Sassolini, Istituto di Linguistica Computazionale "Antonio Zampolli" ILC-CNR, eva.sassolini@ilc.cnr.it
Alessandra Cinini, Istituto di Linguistica Computazionale "Antonio Zampolli" ILC-CNR, alessandra.cinini@ilc.cnr.it

Obiettivi

L'obiettivo prefissato è quello di combinare tecniche di "*distant reading*" e funzionalità classiche di Information Retrieval (IR) su dati testuali. L'esperienza nel trattamento ed interrogazione dei testi, maturata presso l'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR di Pisa (ILC), ci ha permesso sino ad oggi di sviluppare applicazioni web con funzionalità classiche di analisi testuale (Picchi 2003). Attualmente però, in un contesto di maggiore interazione con gli utenti, consideriamo questo approccio non più sufficiente. L'idea è quella di incrementare l'offerta di strumenti di studio e di analisi dei dati con sintesi grafiche e visuali, attraverso un ripensamento delle applicazioni web e, in alcuni casi, studiando una migrazione verso dispositivi mobili e tecnologie App. In questo articolo presentiamo due sperimentazioni fatte tenendo presente questo obiettivo. Nel primo caso abbiamo applicato tecniche di *visual analytics* su un corpus testuale formato da parte della corrispondenza di Galileo Galilei, contenuta nella biblioteca galileiana di Firenze. Il risultato di questo lavoro ha prodotto rappresentazioni visive interattive dei dati che permettono all'utente di acquisire conoscenze sulla struttura interna dei dati e di individuare eventuali relazioni al loro interno. Nel secondo caso abbiamo realizzato un'applicazione per sistema Android finalizzata all'interrogazione di dati testuali relativi ad un censimento di architetture moderne della regione Liguria. L'App produce una visualizzazione su mappe georeferenziate dei dati ed è in grado di far emergere dinamicamente correlazioni ad esempio tra aree geografiche e tipologie di architetture o interventi urbanistici.

Il contesto

ILC rappresenta un punto di riferimento per la comunità scientifica nazionale ed internazionale per lo studio e la realizzazione di procedure per l'acquisizione, la gestione e l'analisi automatica dei testi e di materiale lessicale. La necessità di salvaguardare questo patrimonio testuale ha spinto ILC alla creazione di un progetto di recupero, che oggi prosegue con la collaborazione di molte istituzioni pubbliche e private, impegnate sullo stesso fronte (Sassolini, Cucurullo, Sassi 2014). Nell'ambito delle attività del progetto abbiamo pensato di finalizzare il lavoro di recupero alla realizzazione di nuove modalità di divulgazione dei contenuti. Negli ultimi anni si stanno sperimentando con successo nuovi approcci alla fruizione dei testi, con una forte propensione ad una rappresentazione sintetica e visuale (Rydberg-Cox 2011), (Jänicke 2014). Franco Moretti, che ha coniato il termine "*distant reading*", è un sostenitore dell'utilizzo di tecniche statistiche e metodi quantitativi per "leggere" grandi quantità di testi, utilizzando strumenti e modelli elaborati in altri ambiti di ricerca. Per esempio i grafici della storia quantitativa, le mappe della geografia o gli alberi genealogici utilizzati come forme di rappresentazione astratta dei dati (Moretti 2005). In particolare abbiamo analizzato quali tecnologie e quali supporti tecnologici potessero combinarsi meglio con le funzionalità di analisi

testuale, al fine di produrre una rappresentazione dei testi in forma di "grafici, mappe e alberi".

Un'iniziativa che va in questa direzione è quella intrapresa da "Labex Obvil", laboratorio di eccellenza di Parigi, che si pone come un osservatorio della vita letteraria[1], con il fine di esaminare sotto vari aspetti la letteratura francese classica e contemporanea. Il progetto unisce competenze provenienti da diverse discipline, quali specialisti di letteratura e scienze della cognizione e informatici, con lo scopo di affiancare le modalità classiche di fruizione dei testi a quelle più recenti di rappresentazione grafica e visuale dei contenuti. L'iniziativa francese, infatti, utilizza risorse offerte dalle tecnologie informatiche per il trattamento della lingua e dei testi, ma anche strumenti per la creazione di rappresentazioni grafiche e bidimensionali dei dati.

La sperimentazione con applicazione web

La prima sperimentazione fatta adotta tecniche di *visual analytics* per la gestione dei dati testuali. Abbiamo infatti iniziato a utilizzare le funzioni di rappresentazione visuale che le librerie D3 JS mettono a disposizione. Questa prospettiva porta con se un allargamento della platea dei fruitori, non solo gli "addetti ai lavori", ma anche utenti comuni, che utilizzano oggi dispositivi diversi dal PC per l'accesso ad internet e che hanno esigenze di maggiore sintesi delle informazioni.

I dati utilizzati provengono da un corpus testuale risalente alla prima metà del 1600, il cui contenuto è costituito da lettere, redatte in un linguaggio prevalentemente informale. Il materiale, formato da testi digitali con codifica XML TEI, appartiene ai 20 volumi dell'edizione Favaro[22] della biblioteca di Galileo, che si trova a Firenze presso il Museo Galileo. Degli otto volumi contenenti il "Carteggio" galileiano abbiamo scelto un campione significativo, sia come quantità di lettere presenti, che per la tipologia dei contenuti. In particolare si tratta del volume XV che contiene 462 lettere relative ad un arco temporale che va da gennaio 1633 alla fine di dicembre dello stesso anno. La scelta di questo volume è legata all'esistenza di rilevanti eventi storici, che precedono e si susseguono nell'arco di quel periodo. Ad esempio la pubblicazione del "Dialogo Sopra i Due Massimi Sistemi del Mondo" del 1632, e il successivo processo e condanna dell'autore da parte dell'Inquisizione nel giugno del 1633.

Ogni lettera è organizzata in "campi" che ne consentono una strutturazione funzionale all'estrazione di dati matriciali. In particolare:

- Titolo;
- Mittente;
- Destinatario;
- Destinazione;
- Luogo (da cui è possibile ricavare la data);
- Conservazione.

Solo il titolo è un campo obbligatorio mentre gli altri possono non presentarsi in ogni lettera. La necessità di avere per il progetto dati omogenei, ha portato alla eliminazione di quei documenti che, oltre al titolo della lettera, non presentino almeno i campi principali quali mittente e destinatario.

Le rappresentazioni visuali dei contenuti utilizzano diverse modalità grafiche, ognuna corredata della possibilità di interagire con il motore di analisi testuale. L'accesso al testo è sempre garantito attraverso le funzionalità di analisi testuale più classiche, cambia solo la formulazione della query che viene legata all'interazione con l'oggetto grafico utilizzato:

1 http://obvil.paris-sorbonne.fr/

2 A. Favaro, La libreria di Galileo Galilei, in «Bollettino di bibliografia e di storia delle scienze matematiche e fisiche», XIX, 1886, pp. 219-293 e successive appendici del 1887 e 1896.

- diagrammi a barre per produrre la sintesi degli attori del carteggio:
  - quali e quanti sono i personaggi che scrivono a Galileo e con quale frequenza;
  - a chi lo scienziato scrive con più assiduità;
- diagrammi temporali per mettere in relazione missive ed eventi storici. Con questa modalità è possibile:
  - valutare come lo scambio di messaggi sia strettamente connesso al diffondersi per esempio della notizia dell'avvenuta condanna di Galileo da parte dell'Inquisizione;
  - mettere in evidenza triangolazioni/gruppi di corrispondenze legate da una stretta temporalità, individuando così l'esistenza di possibili "temi di discussione";
- alberi per l'analisi della lingua adottata per le missive. Per esempio emersione di arcaismi evidenti nelle forme, soprattutto verbali:
  - in una prima rappresentazione a cluster (*Container scheme: tree map layout*) si è cercato di mostrare sinteticamente il "formario" come un insieme contenente altri insiemi di parole, organizzate in categorie grammaticali (sostantivi, verbi aggettivi, ecc.) o strutturali (formule di apertura, di cortesia, di saluto), dimensionati secondo le frequenze di attestazione.
  - In una seconda modalità (*Stacked scheme: node link layout*) si è utilizzato lo stesso file di dati matriciali estratti dal corpus (file .json) e utilizzato per la visualizzazione a cluster, per la costruzione di un albero, che consentisse espansioni interattive, sino ad arrivare ad ogni singola forma (foglia). Criteri di razionalizzazione dei dati hanno imposto un taglio delle forme con frequenze basse. In questa seconda modalità è possibile una maggiore interazione con il motore di IR, non solo per vedere contestualizzate le singole occorrenze delle forme e poi accedere al testo completo della lettera desiderata, ma anche l'interazione con funzionalità avanzate di ricerca per lemma o per categoria grammaticale, nel caso di testi con annotazione morfo-sintattica.
- Un grafo orientato che fosse in grado di rendere visibile l'intero carteggio in una sola schermata, mettendo in evidenza le caratteristiche salienti dei dati. Fornendo al tempo stesso indicazioni di volume struttura, relazioni e tipologia dei dati e suggerire ulteriori analisi e studi.

La sperimentazione con App

Un secondo esperimento è stato fatto nell'ambito dei beni culturali, la cui conoscenza può essere diffusa non soltanto per mezzo della divulgazione dei contenuti storico-artistici, ma anche attraverso l'uso di applicazioni software per la georeferenziazione e la visualizzazione su mappe interattive. L'obiettivo della sperimentazione è quello di rendere fruibili i dati raccolti nell'ambito del progetto di ricerca "Censimento e schedatura di complessi di architettura moderna e contemporanea in Liguria"[3]. Nelle intenzioni del progetto da un lato era forte la necessità di divulgare contenuti nuovi o difficilmente reperibili, organizzarli, standardizzarli e metterli a disposizione della comunità, dall'altro offrire una modalità di fruizione intuitiva e diffusa, come avviene con tablet e smartphone. In questo modo si può raggiungere una platea di fruitori più ampia: non solo studiosi del settore ma anche utenti comuni più orientati ad un accesso veloce, sintetico, spesso grafico alle informazioni.

In ambito mobile, la consultazione dei contenuti testuali si intreccia con i dati geografici e spesso avviene in loco: attivando sistemi di notifiche, sono suggerite dinamicamente all'utente

3 Progetto ideato e realizzato dall'allora Direzione regionale per i Beni Culturali e Paesaggistici, oggi Segretariato regionale MIBACT per la Liguria, da Regione Liguria e Dipartimento DSA di Scienze per l'Architettura dell'Università degli Studi di Genova, nell'ambito dell'Accordo di Programma Quadro _Beni e Attività culturali III integrativo_ Intervento BF-10 Progettazioni per lo sviluppo di programmi di valenza strategica in materia di cultura

informazioni correlate al luogo in cui si trova o ad oggetti che sta osservando. Attualmente è stata sviluppata una applicazione per sistema operativo Android, che offre una panoramica di una selezione delle architetture censite, ciascuna corredata di informazioni analitiche e di un "tag" geografico che ne permette la visualizzazione su mappa. I contenuti descrittivi sono disponibili solo per le architetture di maggior rilievo, l'utente può navigare nei testi, con le funzioni base di IR e visualizzare le opere rispondenti ai criteri di ricerca. Sulla mappa si evidenziano ad esempio le aree in cui un determinato progettista ha operato, attribuibili ad uno stile architettonico o legate alla presenza di specifici interventi (ricostruzione, riqualificazione, etc.). Cluster basati sulla prossimità geografica possono suggerire all'utente itinerari di visita.

La sperimentazione è stata fatta con un campione ancora esiguo di documenti ma, nella prospettiva in cui saranno presenti una quantità significativa di materiali testuali, è prevista una maggior interazione tra le due modalità di consultazione: testuale e su mappa; in grado di proporre all'utente una navigazione nei dati più articolata. La scelta di una App nativa è stata infatti dettata dalla modesta quantità di materiali testuali disponibili, che garantiva un accesso alla maggior parte dei contenuti e delle funzioni anche in assenza di connessione. In futuro, prevedendo di incrementare la quantità di informazioni, il database e le funzioni di ricerca testuale saranno demandate all'interazione con un server, mentre all'utente saranno garantite alcune funzionalità o la consultazione di una parte dei dati anche offline. In particolare per l'analisi automatica e l'estrazione di terminologia di dominio si utilizzeranno gli strumenti e le risorse sviluppati presso ILC in grado di gestire grandi quantità di dati testuali.

Riferimenti Bibliografici

Jänicke, Stefan. Franzini, G. Cheema M. e Scheuermann, G. (2015). *On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges*. Proc. of EuroVis - STARs, 83-103.

Sassolini, Eva. Cucurullo, Sebastiana. Sassi, Manuela. 2014, *Methods of textual archive preservation*. In: CLiC-it 2014 – First Italian Conference on Computational Linguistics (Università di Pisa e CNR, Pisa, Italia, 9-10 dicembre 2014). Proceedings, vol. I, ISBN 978-886741-472-7, pp. 334 – 338. Pisa University Press., Pisa.

Sassolini, Eva. Cinini, Alessandra. Sbrulli, Stefano. Picchi, Eugenio. 2013. *Tools and Resources Supporting the Cultural Tourism.* In: GL14 - Fourteenth International Conference on Grey Literature

(National Research Council, Rome, Italy, 29-30 November 2012). Proceedings, vol. (GL conference series, ISSN 1386-2316 ; No. 14) pp. 177 - 180. D.J. Farace, J. Frantzen, GreyNet (eds.). TextRelease, Amsterdam.

Thomas, Jim. Kielman, Joe. (2009). *Challenges for visual analytics*. Information Visualization, 8(4), 309-314.

Rydberg-Cox, Jeff. 2011. *"Social networks and the language of greek tragedy"*, Journal of the Chicago Colloquium on Digital Humanities and Computer Science. Vol. 1. No. 3.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London; New York: Verso.

Picchi, Eugenio. 2003. *PiSystem: sistemi integrati per l'analisi testuale*. In A. Zampolli, N. Calzolari,

L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa. Linguistica Computazionale, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, 597-627.

Modeling Regions from Premodern Geographical Hierarchical Data

Masoumeh Seydi Gheranghiyeh, Leipzig University, m.seydi@uni-leipzig.de
Maxim Romanov, Leipzig University, maxim.romanov@uni-leipzig.de

Introduction

Geographical texts are an important part of the premodern Islamic written legacy. They cover valuable information for historical studies, such as administrative and geographic divisions, routes by which settlements or any other urban/geographic places were connected, distances, descriptions of geographical, social, cultural, and administrative features. This information opens a number of new research perspectives. First, one can study the fluctuations of social and administrative geographies over the time. Second, one can compare the comprehensive visions of the geography of the Islamic world given by geographers from different periods and different parts of the Islamic world. The way the authors describe the history of an area at nearly close or exact time span shows how the area changes over time. At the core of these discussions stands historical data that explains regional divisions. Regions might vary from source to source and from time to time. With such data, we can use computational approaches to model different aspects of regional histories and find answer to questions like:

- how the Islamic world was divided into administrative regions?
- how these regions were changing over time?
- what were the spatial extents of regions?
- how were these regions interconnected with each other?
- how can we use these geographical models to study other historical data from the geographical perspective?

In this paper we examine various mathematical models for geographical hierarchical and route network data that can be collected from premodern sources. Although we focus here on Islamic texts, the same approach can be applied to geographical texts that supply similar type of data in any historical language. We implement corresponding visualizations to provide a better visual insight into relevant data and to demonstrate how they fit into each model. We also evaluate these models by describing their strong and weak sides.

Data

Georgette Cornu's *Atlas du monde arabo-islamique à l'époque classique: IXe-Xe siècles* (Cornu, 1983) is our primary source of geographical data. This data contains only one level of regional divisions (i.e. it divides the Islamic world into major provinces, while premodern geographical sources often also describe the division of provinces into districts). Any other datasets can fit into our model if they offer a structurally similar information on places and their provincial affiliations. The input data should be in one of these data formats: GeoJSON (Butler, 2008), TopoJSON (extension of GeoJSON), or CSV (comma-separated values). We build an hierarchical tree of our data starting from a root representing a big area containing all data which we call "*the working area"*. The working area then is divided into macro regions and in the next levels there are micro/sub regions containing settlements in the leaves. Using the tree structure, Figure 1 illustrates how a sample dataset of regions and settlements look like in five levels of

hierarchies. In Figure 2 we can see part of this data more closely. (NB: This data is collected from *Ahsan al-taqasim fi ma rifat al-aqalim*, a comprehensive geographical text written by al-Muqaddasi[1] (Muqaddasī, 1906 and Muqaddasī, 1994).

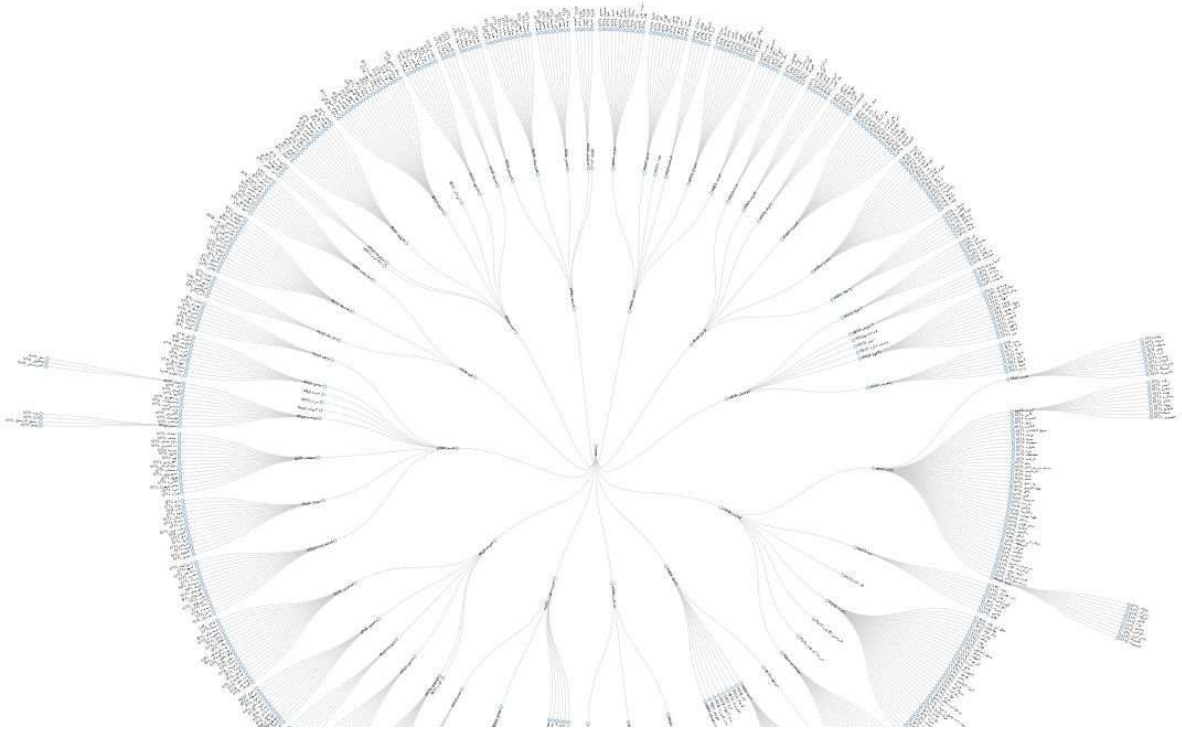


Figure 1: Hierarchical Divisions Data, Visualized by Tree Structure

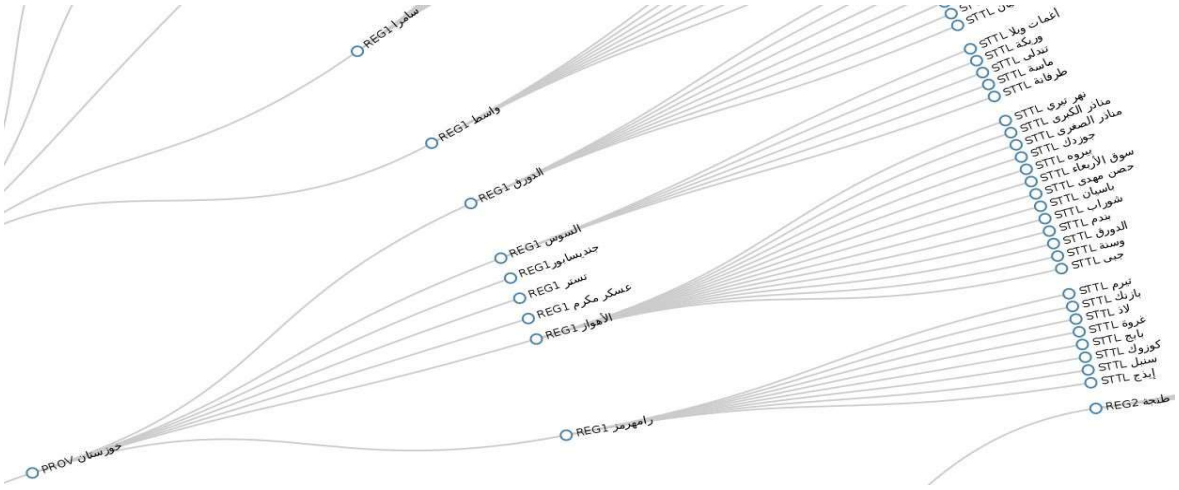


Figure 2: Hierarchical Divisions Data - Zoom in

1 10th century CE

Mathematical Models and Visualizations

The idea is to provide mathematical models to describe and visualize regional divisions. We use the resulting visual perception to improve the current model or think of a new one. We consider the mathematical models quadtree clustering (Samet, 1984), Voronoi diagram (Aurenhammer, 1991), convex hull, and concave hull. Our implementations of these particular models are based on (Meeks, 2016).

Quadtree

We first start with quadtree clustering which divides the space into a multi-level grid cells on the plane such that each cell contains one or more settlements. An illustration is shown in Figure 3. Quadtree divides the space into cells with maximum capacity of points/objects holding inside each cell and when it reaches the capacity, fully splits the cell into four new cells of equal area. We will finally have each individual settlement in the smallest cells in the lowest level of tree structure. Figure 3 shows circles with different radiuses representing distinct levels of zooming. More clearly, bigger circles contain smaller regions or settlements and the smallest circles stand for individual settlements.

Quadtree performs well to depict the density of settlements and thus to show the contrast of settled vs. unsettled areas. Furthermore, this model is useful for datasets without any hierarchical information. However, this is not what literally the regions have been shaped in the real world. It means, this algorithm does not allow to take advantage of hierarchical divisions, as its clustering is based on spatial proximity.

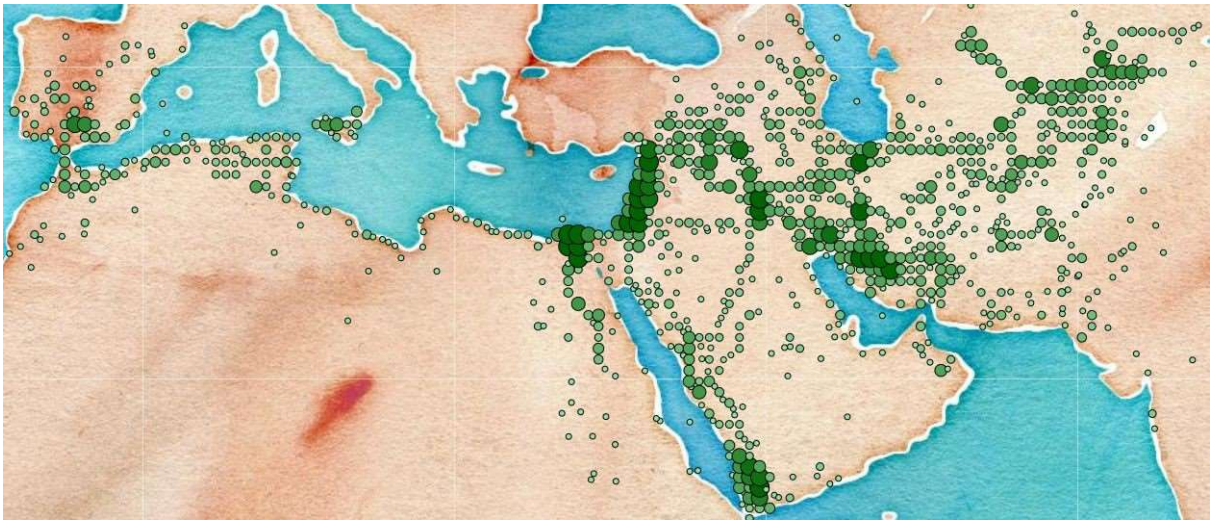


Figure 3: Quadtree Visualization of Settlements

Voronoi Diagram

Voronoi diagram is another model which we examine here. Assuming that each settlement is a generating point on the plane, the Voronoi algorithm divides the plane into cells in shape of polygons such that every single cell contains exactly one generating point. The divisions are calculated so that the distance of every point in a cell to the generating point in that cell is shorter than to any other generating point on the plane. We visualize the calculated Voronoi diagram

for our dataset in Figure 4. All polygons (cells) that belong to the same region—here we can take advantage of hierarchical data— are filled with the same color. Thus, areas formed from polygons of the same color represent distinct regions.. The same idea applies to the higher level of hierarchies by merging the regions.

Figure 4: Voronoi Diagram for Region Visualization

There is an issue with the classic implementation of Voronoi diagram regarding the plane on which the Voronoi diagram is expanded. The points in our dataset belong only to a limited area of the Earth, but Voronoi divides the whole Earth to place the points in polygons. Consequently, we will have a diagram expanded all over the Earth with very large cells where the points are sparse, like those close to the borders of regions. To fix this problem, we limit the plane to the working area (the area in which all coordinates fit into), instead of drawing the diagram on the whole Earth. For this purpose, we clip the Voronoi diagram with the bounding box of the working area. In Figure 4, we also have added a filter to exclude areas bigger than a specific value to visually skip unnecessarily big partitions. We calculate this value by considering the average area of all polygons. Hence, this value may vary for each dataset.

Voronoi diagram provides more insightful shapes of regions compared to quadtree since it actually takes the advantage of the hierarchical data. It forms regions based on the position of points on the plane without any overlap of the regions. However, the extents of regions that Voronoi diagram suggest a more modern-type of a map, where countries border each other with no territory being “no man’s land”. Subsequently, it leaves no unpartitioned space also overextending regions into uninhabitable areas such as seas and deserts. To address these problems we need to make some modifications.

Voronoi Clippings

The first step we took to customize the Voronoi diagram was to clip it by the bounding box of the working area. The result is a rectangular area divided by Voronoi polygon. But, our

working area is not in fact rectangular. Since we are working on pre-modern data, we cannot clip the diagram by modern country borders. Hence, we should find the best estimation of a working area. The convex hull (de Burg, 2000) of the all points is a polygon which encompasses all points. This is a convincing poly which we can utilize to clip the Voronoi diagram. This gives us a combination of two different models to improve the original idea of Voronoi diagram clipping.

As shown in Figure 5, clipped diagram is drawn on tighter area which is no more rectangular and we can see how the polygons are expanded and limited. In this figure, we exclude those partitions which are greater than a specific value to get a better view of the result, as we did before.

Figure 5: Voronoi Diagram Clipping, Using Convex Hull of the Plane

Limiting the Voronoi diagram to a convex hull, we will have still some large empty areas considered as regions (see the non-colored areas in Figure 5). It occurs because of large cells of Voronoi diagram as well as the convexity of generated convex hulls. Since these areas contain insignificant or zero amount of data, we should clip them.

An idea is to clip the diagram by a concave hull (Edelsbrunner, 1983) of all points, instead of convex hull, since concave hull eliminate some unused areas. But it trims again only some parts around the working area and not some parts inside or around the inner regions. Instead, we clip the diagram with a set of convex hulls computed for each region. Accordingly, we dispose the areas belonging to none of the regions and at the same time we fit the whole diagram into the convex hulls of regions, as shown in Figure 6.

Compared to former models, this model depicts more compact and accurate regions. However, there are still empty parts which need to be cut. Moreover, convex hulls may cause some overlaps, as in our data.

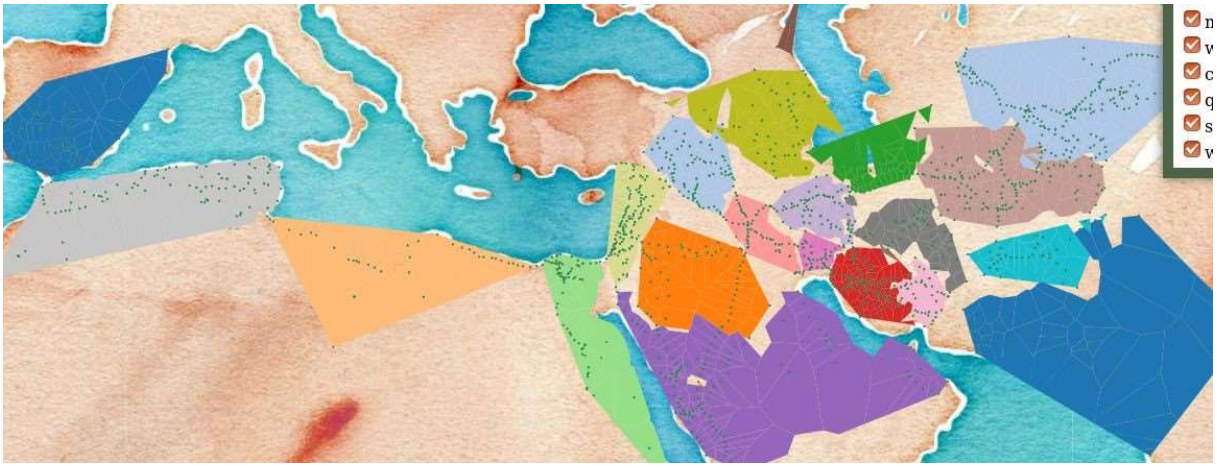


Figure 6: Voronoi Diagram Clipping Using Convex Hull of each Region

Convex Hulls

Following the idea of convex hulls, we introduce another model to make effective divisions of the plane using only the convex hulls of each region without considering Voronoi diagram as illustrated in Figure 7. Convex hulls shape and limit the region to the corresponding points and can be applied to any level of hierarchies. Divisions in each level are convex hulls created by any number of points or regions merging together. Although this model introduces a more effective partitioning which might be more practical in visualization, it leaves some overlappings of regions and some parts which we do not need them in regions and should expunge them.

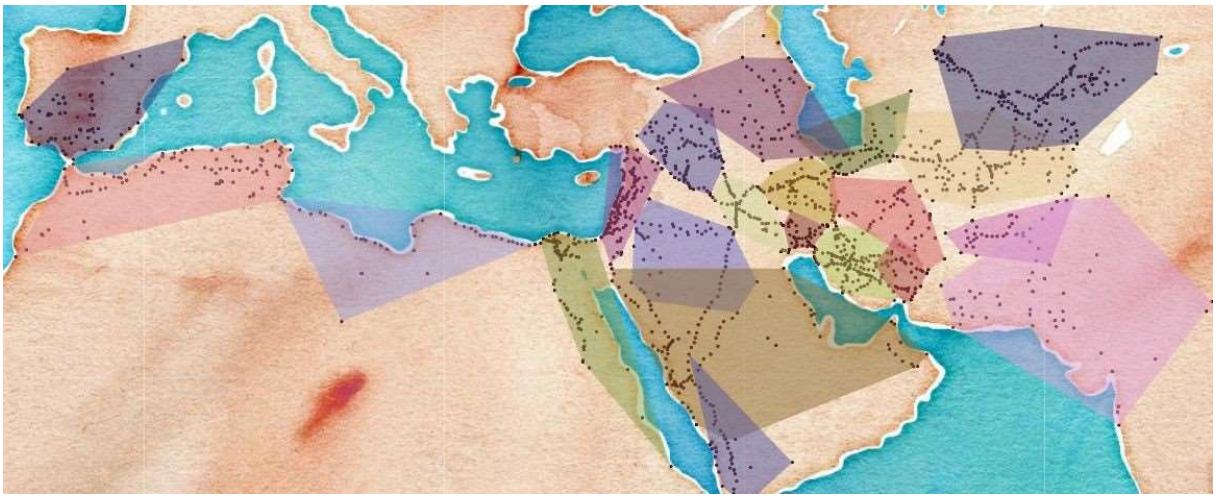


Figure 7: Convex Hulls for Visualizing Regions

Concave Hulls

The idea of convex hulls leads to another model by replacing convex hull by concave hulls of each region. As we see in Figure 8, the regions have more nuanced shapes compared to former models. Here, we exclude some "empty spaces" which do not belong to any region, like deserts. Unlike convex hulls, this model avoids overlapping of regions if there is no inconsistencies in the dataset, as in real world.

Considering all aspects of this model, we still need to clip some parts—like large water bodies that do not belong to any region based on our data. Besides, there are some parts which do not satisfy the explanation of some regions from pre-modern sources. For example, the hull showing Egypt (Misr) is expanded beyond the delta and the valley of the Nile river, which is how it is usually described. The triangular shapes along the Nile that we see in Figure 8, are formed because of the calculations towards shaping a concave hull by a set of points, thus visually grabbing more territory than is desired; the same effect can be observed in other regions that have complex geometries.

Figure 8: Concave Hulls for Visualizing Regions

Route Network

Up to this point, we have been using only the settlements of a region to propose models for visualizing regions from textual data. In addition to settlements, there are also route information in sources which we can use for modeling regions. We combine corresponding route network information with settlements of different regions and build individual shapes to introduce a new model. Each shape is build on a set of points and lines in the same color to depict a region, as illustrated in Figure 9. These shapes show the extent of regions without any need to clip some part and having overlaps or empty spaces. Following the same pattern, we can illustrate subregions and micro regions as well, to represent various levels of hierarchies. This model resolves the mentioned problems in former models and visually seems best matching to the data which does not specifically explain the expansion and shape of the regions.

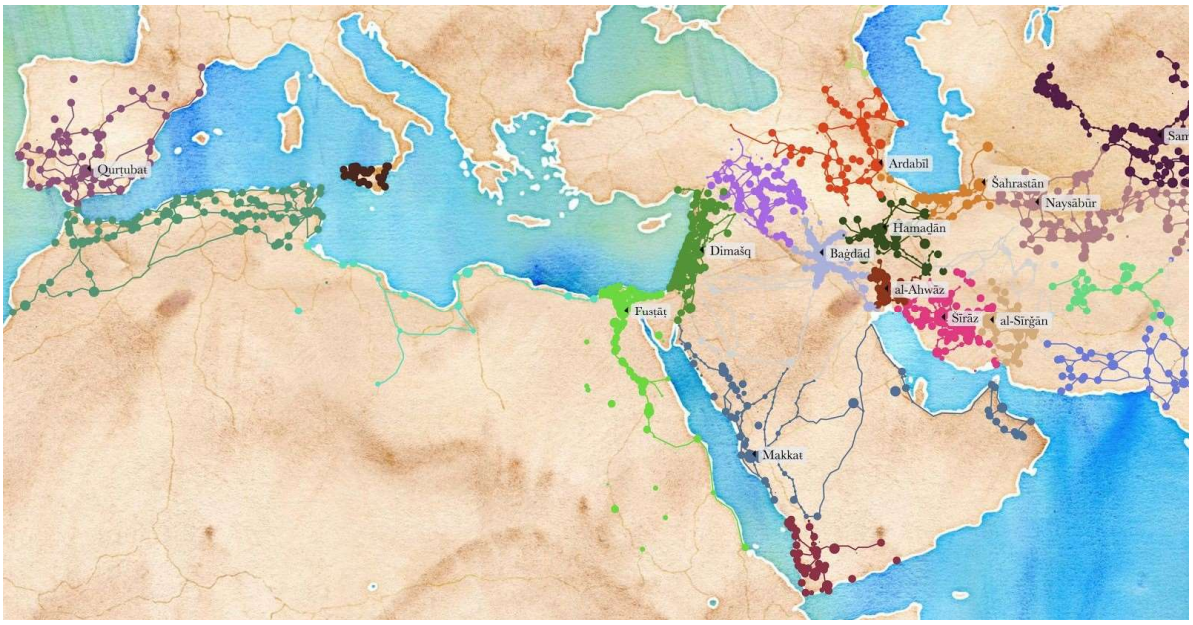


Figure 9: Visualizing Regions by Settlements and Route Network

Conclusion and Future Work

We proposed several models models for visualizing regions based on textual, descriptions of administrative divisions and route networks from premodern sources. We evaluated these models and applied proper alterations to adjust them to our needs. The models that we have discussed are quadtree, Voronoi diagram, convex hull, concave hull, and a model that combines the route network and administrative hierarchies. Each model has its advantages and disadvantages. The route-network model appears to be most useable for general purposes of visualizing historical geographies, while other models may provide valuable analytical insights into particular historico-geographical questions. Thus quadtree is a powerful way to show densely settled areas, while Voronoi diagram can be used to analyze areas of influence. Concave hulls as closing areas for places seem also practically effective for adjusting Voronoi diagram. In addition, our approach can be extended to visualize more hierarchical data with more than level of regional division.

Bibliographical References

Cornu, Georgette, 1985. *Atlas du monde arabo-islamique à l'époque classique: IXe-Xe siècles*. Leiden: Brill.

H. Butler, M. Daly, A. Doyle, S. Gillies, T. Schaub, C. Schmidt, 2008. *The GeoJSON Format Specification,* http://geojson.org/geojson-spec.html.

Hanan Samet, 1984. *The Quadtree and Related Hierarchical Data Structure.* ACM Comput. Surv. 16, 2, 187-260.

Franz Aurenhammer, 1991. *Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure.* ACM Computing Surveys, 23(3):345–405.

M. de Berg, M. van Kreveld, M. Overmars, O. Schwarzkopf, 2000. *Computational Geometry: Algorithms and Applications,* Springer, pp. 2–8.

H. Edelsbrunner, D. Kirkpatrick and R. Seidel, 1983. *On the shape of a set of points in the plane,* IEEE Transactions on Information Theory, vol. 29, no. 4, pp. 551-559.

Elijah Meeks, 2016. *Elijah Meeks's Blocks.* http://bl.ocks.org/emeeks

Muqaddasī, M. ibn Aḥmad., Goeje, M. J. de. ,1906. *Kitāb a san al-taqāsīm fī ma rifat al-āqālīm.* Tab‘a 2. 2 الطبعة. Laydan: Brīl.
Muqaddasī, M. ibn Aḥmad, 1994. *The best divisions for knowledge of the regions: a translation of Ahsan al-taqasim fi ma ifat al-aqalim.* Reading, UK: Centre for Muslim Contribution to Civilisation.

Twitter Data Exploration for Italian History

Rachele Sprugnoli, Fondazione Bruno Kessler and University of Trento, sprugnoli@fbk.eu
Giovanni Moretti, Fondazione Bruno Kessler, moretti@fbk.eu
Sara Tonelli, Fondazione Bruno Kessler, satonelli@fbk.eu

Introduction

Social media texts are a specific type of big data and, as such, constitute a socio-technical phenomenon (Boyd and Crawford 2012) with great potentials but also methodological and conceptual challenges (Tufekci, 2014). Even if issues related to accessibility and authenticity are intrinsic to the nature of big social data (Manovich, 2011), they have become widespread in everyday life, for example to be entertained and stay informed, but they also emerged as a crucial resource in the business domain (Qualman, 2010) and in many research fields. As for this last point, the literature reports many studies that apply the analysis of large-scale social media texts to, among others, crisis management and health communication (Yin, 2012; Moorhead, 2013). As for the Humanities, Lee et al. (2012) and Krutka and Milton (2013) provide examples on how the popular microblogging platform Twitter can be successfully used in schools to implement historical reenactments and encourage historical thinking. Outside classrooms, historians have recognized social media as valuable forums for academic discussion: for example, taking advantage of the so-called Academic Twitter phenomenon (Stewart, 2016) historians live-tweet conferences, disseminate research outcomes and use hashtags, such as #Twitterstorians, to follow and participate in academic debates (Hitchcock, 2014; O'Brien, 2015; Hitchcock and Shoemaker, 2015). Nevertheless, the horizontal and bottom-up communication enabled by social media allows to break the traditional boundaries between academia and laypeople democratizing historical work and facilitating the engagement with the public (Foster, 2014; Myers and Hamilton, 2015). Social media can thus be seen as primary sources useful to study contemporary events, such as the Arab Spring and other recent social protests (Myers and Hamilton, 2014), as well as to understand the way ordinary people engage with the past (Meschini, 2014). This idea is at the basis of the ambitious, and still not operative, project promoted by the Library of Congress of archiving every public tweet for future generations of researchers (Zimmer, 2015). Less all-embracing projects have been carried out on specific historical events: this is the case, for example, of the *Twitterification* of the Finnish Winter War within the #sota39 project and of the D-Day within an initiative of the The British National Archives (The National Archives, 2014). The former project has been analysed taking into consideration the audience participation and the main characteristics of the historical narration as mediated by Twitter, without using any computational tool (Lähteenmäki and Virta, 2016). On the contrary, Clavert (2016) adopts text analysis and social network analysis techniques together with network visualisations to study how the First World War is discussed on Twitter. All these analyses and visualisations are performed using different tools, each requiring separate installation, diverse data format and data manipulation. Our idea is, instead, to provide an interface where all the functionalities are integrated and easy to use.

On the Internet it is possible to find several services for the analysis of tweets but only few of them offer language analyses. Moreover, they are commercial products, thus not free (see for example, NVivo[1] and BlogMeter products[2]). Recently FireAnt, a freeware software for processing social media data, has been released (Anthony and Hardaker, 2016): it allows to collect Twitter data containing specific hashtags or belonging to a specific account and to export these data in different formats and graphics. Anyway, FireAnt does not include text analysis tools, except for a concordancer.

Works cited above focused on various languages such as English, Finnish and French while less attention since now has been given to Italian. Nevertheless, Italian is strongly present in social media: for example, Twitter has 6.4 million active users in Italy[3]. Also institutions are increasingly active online in Italy, with 76% of Italian Universities having a Twitter account (Oppici et al., 2014), 64% of members of Parliament (source: Pokedem, 2016[4]) and 47% of Italian NGOs (source: Sodalitas, 2014[5]). For this reason, we decided to develop MARTIN (*Monitoring and Analysing Real-time Tweets in Italian Natural language*), a free research tool that provides users, interested in the exploration of Italian Twitter data, with flexible, easy-to-use analytics. A first general-purpose version of MARTIN won the second prize at the "IBM Watson Services Challenge"[6] organized in the context of EVALITA 2016, the evaluation campaign of NLP and speech tools for Italian. In this work, we present a modified, domain-dependent version of the tool, designed to meet the research needs of history scholars, and we discuss history-related use cases.

The design of MARTIN builds on previous experiences of developing platforms for supporting Humanities studies (Moretti et al., 2016), where the smooth transition between close and distant reading (Moretti, 2005) is crucial. Scalable reading (Mueller, 2014) is guaranteed because MARTIN allows to zoom in into the tweets by providing the list of retrieved messages and zoom out by looking at the visualisations that display the output of the analyses. Moreover, MARTIN contains analyses that proved to be useful in the context of historical investigations such as key-concept extraction and co-occurrence networks (Sprugnoli et al, 2017).

In this paper, we show how the tool can be employed to investigate the public debate around past historical events, such as the First World War, as discussed in Twitter.

Description of the Tool

MARTIN is a stand-alone application that integrates Twitter APIs[7] to scan real-time information on Twitter and also to recover tweets published in the last 7 days[8]. MARTIN relies on Natural Language Processing (NLP) modules to analyse the language of tweets and on libraries for data visualisation, i.e. ECharts and Highcharts, to display the output of the analyses. Tint[9] (Palmero Aprosio and Moretti, 2016), a modular and open source tool for Italian NLP, is used to perform PoS tagging, morphological analysis, key-concept extraction, and to compute readability metrics.

Readability is calculated by a module that measures the lexical complexity of tweets using the Basic Italian Vocabulary by De Mauro (1999). In particular, it considers the proportion of words in the tweet that appear in such vocabulary, as a proxy for a simple language and high readability of the message. As for key-concepts, MARTIN takes advantage of the KD module included in Tint: KD combines statistical measures with linguistic information given by PoS patterns to extract a list of weighted keyphrases from texts (Moretti et al., 2015). PoS tagging is also used for the creation of co-occurrence

1 http://www.qsrinternational.com/what-is-nvivo
2 https://www.blogmeter.eu/
3 http://www.wired.it/internet/social-network/2016/04/04/social-media-italia-crollo-twitter-esplode-snapchat/
4 http://www.pokedem.it/
5 http://www.sodalitas.it/conoscere/comunicati/social-network-al-nonprofit-piace-facebook-la-ricerca-della-visibil ita-vince-ancora-su-quella-dei-fondi
6 http://www.evalita.it/2016/tasks/ibm-challenge
7 https://dev.twitter.com/rest/public
8 This limitation is due to the Twitter's search index that allows to go back 7 days.
9 http://tint.fbk.eu/

networks, where it is possible to search for a lemma and retrieve all the words (despite their grammatical category or only nouns, only verbs, only adjectives) or expressions (i.e., part-of-speech patterns such as verb+noun) that appear in a context chosen by the user. In addition, MARTIN uses a distributional polarity lexicon made of more than 75,000 Italian lemmas acquired from social media data (Castellucci et al., 2016) to highlight the sentiment (negative, positive, neutral) of words retrieved in the co-occurrence networks.

A screenshot of MARTIN interface is given in Figure 1.

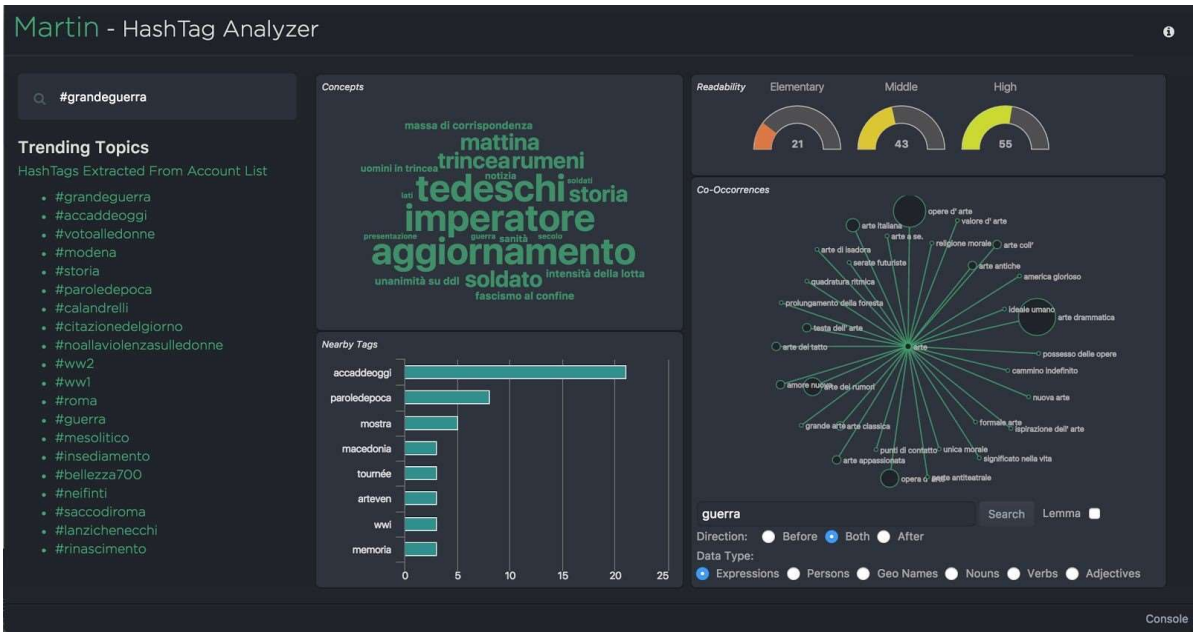


Fig 1. Interface of MARTIN.

Using the MARTIN Tool for History-Related Analyses

While MARTIN was designed as a domain-independent tool to perform Twitter-based analyses, we created a specific configuration to focus on history-related topics. In particular, we manually defined a list of accounts of Italian historical associations (e.g., @Istitutostorico), journals (e.g., @studistorici), and public initiatives related to historical events (e.g., @Centenario14_18). Then, the tool retrieved the hashtags mentioned in the tweets issued by these accounts, and ranked them by frequency. The output of this step is a ranked list of the most popular hashtags in the History domain in the last 7 days. Figure 2 shows the top part of the list as retrieved on November, 26 2016. The most frequent hashtag is the one related to the First World War centenary celebrations in Italy (*#grandeguerra* / "great war"). We find in the list also hashtags used internationally to deal with First and Second World war, such as *#ww1* and *#ww2*. Three very popular hashtags are not connected to a specific historical event but are daily used to chronologically trace what happened in the world (*#accaddeoggi* / "it happened today") and to report quotes by both famous and ordinary people (*#citazionedelgiorno* / "quote of the day" and *#paroledepoca* / "words of the period") of the past. Contemporary issues burst into the list with *#noallaviolenzasulledonne* / "no violence against women" e that is dedicated to the movement against gender-based violence due to the national demonstration that took place on November 26. Another hashtag on women's rights is high in the rank (*#votoalledonne* / "vote for women"), stressing the interest for gender issues in current social media discussions.

After the extraction of the ranked hashtag list, keyword extraction was performed on each group of tweets mentioning a given hashtag. As an example, we present in Figure 3 the key-concepts automatically extracted from tweets with the hashtag *#votoalledonne*. Different font dimensions correspond to different keyword relevance (based mainly on frequency). The keywords display a concise representation of the tweets related to women's vote, covering main issues such as the difficult

journey to obtain this right (*tortuoso percorso* / “tortuous path”) but also the importance of this step (*fautrici delle nostre libertà* / “advocates of our freedoms”). The keyword *convegno* / "conference" shows that historical events are often linked to public celebrations that are promoted and discussed on social media. Another example of this is given by the key-concepts extracted from tweets published in the period of the 50th anniversary of the Florence Flood (first half of November 2016) and containing the hashtag *#alluvione50*. Among those keywords, the majority was about exhibitions, concerts and public meetings.

Fig 2. Trending hashtags in the History domain

Fig 3. Key-concepts extracted from tweets containing the hashtag *#votoalledonne*

Tweets often include more than one hashtag: MARTIN extracts the ones co-occurring with the hashtag searched by the user, thus providing additional insights into the content of tweets. Two examples are reported in Fig. 4. We display on the left hashtags co-occurring with *#grandeguerra*, dealing with the countries involved in different phases of the war (see *#macedonia*, *#serbia* and *#romania*) but also with cultural events related to the centenary. On the right, hashtags associated with *#noallaviolenzasulledonne* are presented, dealing mainly with the national demonstration of November 26 in Rome. This bar chart shows that different hashtags related to the same demonstration are used (e.g., *#nonunadimeno*): this can be a useful indication to expand the search.

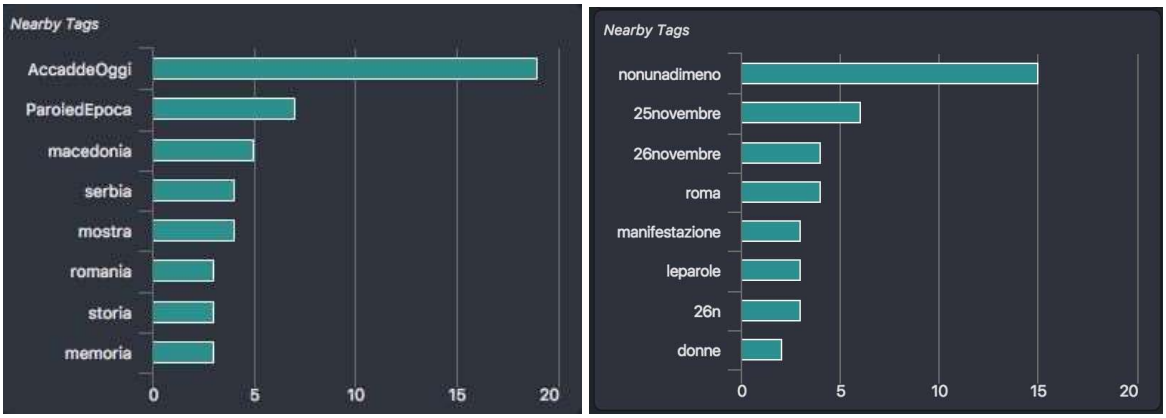


Fig 4. Hashtags co-occurring with *#grandeguerra* (left) and *#noallaviolenzasulledonne* (right)

A more detailed look into the tweets is given by the co-occurrence networks, where the word or an hashtag of interest is displayed as the central node and the other nodes, having a size reflecting their frequency, are the words or expressions surrounding it. The polarity at word level is given by the color of the nodes: red indicates negative words or expressions, green positive ones. All the others are neutral. Figure 5 shows the nouns appearing both in the right and in the left context of the hashtag *#FidelCastro* in the tweets retrieved after his death. The picture shows that Castro is remembered more as a dictator than as the Commander of the revolution. Different critical aspects of his government are mentioned such as repressions, detentions, and the Cuban exiles issue. Clicking on a node is possible to see the corresponding tweets (Figure 6).

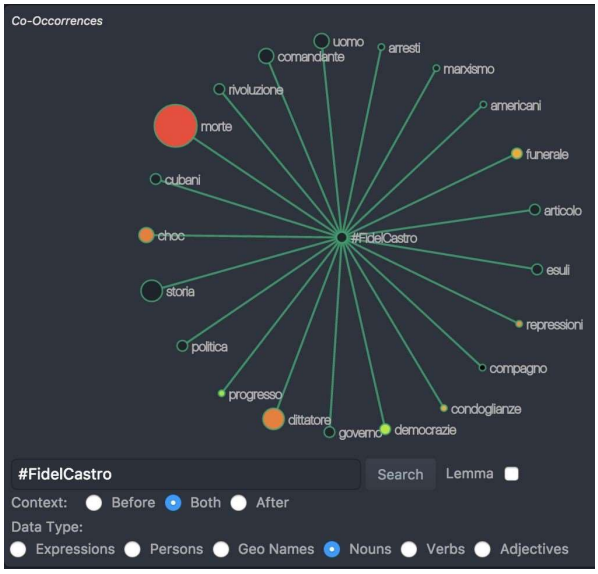


Fig 5. Network of nouns co-occurring with the hashtag #FidelCastro.

#FidelCastro su HugoChavez: «Volete sapere chi era Fidel Castro? Guardate chi lo piange e guardate chi festeggia la sua morte, e lo saprete» https://t.co/CKfnODepAs
"Non c'è peggior fallimento che rallegrarsi della morte di un uomo che non hai potuto vincere da vivo" #FidelCastro
Vi rendete conto che quelli che votano No perché temono dittatura sono gli stessi che piangono per la morte di #FidelCastro
Con la morte di #FidelCastro si chiude con notevole ritardo la stagione delle grandi (e fallimentari) ideologie politiche comuniste
La morte di #FidelCastro chiude definitivamente il #Novecento.
i ns.comunisti ciechi fino alla morte non riconoscono in Castro un dittatore come tutti nei regimi comunisti. COMUNISMO UTOPIA #FidelCastro
Posso dire che i cubani di Miami che ballano e brindano per una morte mi fanno un po'schifo? #FidelCastro.@Corriere @ItalyinCuba @cubadebate
Sebbene non possa tollerare ciò che #FidelCastro è stato, non condivido chi scende in piazza a festeggiarne la morte...ecco cos è l umanità
Cuba alla morte di #FidelCastro Qua si chiedono quale futuro? https://t.co/CKGLuMn9eu
a Cuba 9 gg di lutto nazionale per la morte di #FidelCastro
Attualità: ##Miami ha #accolto così la morte di #FidelCastro, ... (skytg24hd) https://t.co/cRdcPUlvsQ
#FidelCastro "I cubani pontini e la morte di Castro «Niente lacrime, è ora di costruire». Ne parlo su #IlMessaggero

Fig 6.Tweets containing the word *morte* as retrieved from the co-occurring network displayed in Figure 5

Discussion and Future Work

In this work, we presented an adaptation of the MARTIN tool for Twitter analysis to the history domain. In particular, we showed how, starting from a pre-defined list of accounts dealing with history-related topics, it was possible to understand what main events were mentioned, how they were presented and in which contexts they were discussed.

In the future, we plan to extend the tool to support tweets in English, and to include a named entity recognizer so that it will be possible to track which persons are most associated with the considered events. MARTIN will also be released, so that it will be possible to apply this kind of analysis to other

topics or events in the history domain.

Tweet geolocation and visualisation would be easy to implement by relying on Twitter APIs and also useful to find out whether discussions take place in a specific area of Italy or are spread across the whole country. Unfortunately, the percentage of geotagged tweets is very low, between 1.5 and 3% (Paraskevopoulos and Palpanas, 2016), thus geographical information should be integrated through more complex strategies, for example by looking at user short description.

Bibliographical References

Anthony, Laurence, and Claire Hardaker. 2016. FireAnt (Version 1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Boyd, Danah, and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662-679.

Castellucci, Giuseppe, Danilo Croce, Roberto Basili. 2016. A Language Independent Method for Generating Large Scale Polarity Lexicons. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portoroz, Slovenia, 2016.

Clavert, Frédéric. 2016. #ww1. The Great War on Twitter. In *Digital Humanities 2016 Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 461-462.

De Mauro, Tullio. 1999. *Grande dizionario italiano dell'uso*. Garzanti.

Foster, Meg. 2014. Online and Plugged In? Public History and Historians in the Digital Age. *Public History Review*, 21, 1-19.

Hitchcock, Tim. 2014. Doing it in Public: Impact, Blogging, Social Media and the Academy. *Historyonics*. URL: http://historyonics.blogspot.co.uk/2014/07/doing-it-in-public-impact-blogging.html.

Hitchcock, Tim, and Robert Shoemaker. 2015. Making History Online. *Transactions of the Royal Historical Society (Sixth Series)*, 25, 75-93.

Krutka, Daniel and Michael K Milton. 2013. The Enlightenment meets Twitter: Using social media in the social studies classroom. *The Ohio Social Studies Review*, 50.2.

Lähteenmäki, Ilkka, Tatu Virta. 2016. The Finnish Twitter war: the Winter War experienced through the# sota39 project and its implications for historiography. *Rethinking History*, 20.3: 433-453.

Lee, Victor R., Brett E. Shelton, Andrew Walker, Tom Caswell and Marion Jensen. 2012. Retweeting history: Exploring the intersection of microblogging and problem-based learning for historical reenactments. *Designing Problem-Driven Instruction Using Online Social Media*. 2012.

Manovich, Lev. 2011. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2, 460-475.

Meschini, Federico. 2014. Reconstruction of the memory and physicality of the trench: Digital archives and the Great War. *Bollettino di italianistica*, *11*(2), 203-216.

Moorhead, S. Anne, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15.4: e85.

Moretti, Giovanni, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the Dirt: Extracting Keyphrases from Texts with KD. Proceedings of *CLIC-it 2015*, 198.

Moretti, Giovanni, 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.

Moretti, Franco. 2005. *Graphs, maps, trees : abstract models for a literary history*. London: Verso.

Mueller, Martin. 2014. Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment. *Digital Humanities Quarterly*, *8*(3).

Myers, Cayce, and James F. Hamilton. 2014. Social Media as Primary Source. *Media History*, 20:4, 431-444

Myers, Cayce, and James F. Hamilton. 2015. Open genre, new possibilities: democratizing history via social media. *Rethinking History*, 19(2), 222-234.

O'Brien, Laura. 2015. Twitter, Academia and Me. The Society for the Study of French History. URL: http://frenchhistorysociety.co.uk/blog/?p=348.
Oppici, Fiorenza, Juan Carlos De Martin, Federico Morando, Simone Basso, and Giuseppe Futia. 2014. Social University - Le università italiane sui social network (Nexa Working Paper No. 2014-1). Retrieved from Nexa Center for Internet & Society website: http://nexa.polito.it/working-paper/2014-1
Palmero Aprosio, Alessio, and Giovanni Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *arXiv preprint arXiv:1609.06204*.
Paraskevopoulos, Pavlos, and Themis Palpanas. 2016. Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Social Network Analysis and Mining*, *6*(1), 89.
Qualman, Erik. 2010. *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons.
Sprugnoli, Rachele, Sara Tonelli, Giovanni Moretti, Stefano Menini. 2017.Fifty Years of European History through the Lens of Computational Linguistics: the De Gasperi Project. *Italian Journal of Computational Linguistics (IJCoL)* - Forthcoming.
Stewart, Bonnie. 2016. Collapsed publics: Orality, literacy, and vulnerability in Academic Twitter. *Journal of Applied Social Theory*, 1(1).
The National Archives. 2014. Story of D-Day Told through Documents and Twitter 70 Years on – Press Release. URL: http://www.nationalarchives.gov.uk/documents/press-release-d-day-through-twitter.pdf
Tufekci, Zeynep. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014.
Yin, Jie, et al. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27.6: 52-59.
Zimmer, Michael. 2015. The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*, 20.7.

L’uso delle ontologie per la preservazione concettuale dei dati

Marilena Daquino, Università di Bologna, marilena.daquino2@unibo.it
Francesca Tomasi, Università di Bologna, francesca.tomasi@unibo.it

Introduzione

Il fenomeno dei Linked Open Data (LOD) sta ridefinendo le pratiche della rappresentazione informatica dei contenuti culturali. L’approccio documento-centrico del Web 1.0 è stato superato da un Web che vede nel dato l’elemento atomico delle dichiarazioni espresse attraverso triple. Il *Web of data* identifica una “extension of the Web with a global data space based on open standards” (Heath e Bizer 2011, 1).

L’approccio data-centrico impone che le ‘cose’ siano univocamente identificate e adeguatamente correlate attraverso relazioni semanticamente profonde. Ma la progressiva amplificazione della quantità dei dati richiede un controllo sulla qualità degli elementi del Web. Il LOD cloud aumenta esponenzialmente (Schmachtenberg, Bizer e Paulheim 2014), e vanno previsti meccanismi di preservazione del contenuto informativo, al fine di garantire l’accesso alla conoscenza *ad perpetuam memoriam*.

Una delle soluzioni adottate per assicurare la preservazione attraverso meccanismi di metadatazione è rappresentata dalla documentazione relativa alla *provenance*. E fra gli standard di metadati previsti per garantire la preservazione, PREMIS rappresenta senz’altro l’esperimento più compiuto.

È nelle specifiche di PREMIS che gli autori dichiarano (grassetto nostro):

> The **provenance** of the object: Information describing the **custodial history of the object**, potentially stretching back to the **object’s creation**, and moving forward through successive changes in physical custody and/or ownership. Provenance information includes descriptions of the actions that have been taken **to preserve the object over time**. Such information describes aspects of the digital preservation process used to maintain the object; it would also record any consequences of this process that alter the content, or look, feel, and functionality of the object. Related to this would be information that serves to establish and validate the **object’s authenticity**, i.e., that the preserved object is in fact what it purports to be, and has not been altered, intentionally or unintentionally, in an undocumented way. Authenticity would include such elements as fixity and integrity (Lavoie e Gartner 2013, 5).

In senso estensivo la *provenance* andrà concepita come qualunque atto formale di attribuzione di responsabilità di un’azione compiuta su di un oggetto. Sarà allora lecito considerare anche il procedimento di creazione di una tripla RDF come un atto a cui assegnare un *creator*. Attraverso l’attribuzione di paternità agli asserti è possibile riconoscere il responsabile della creazione del dato strutturato, e quindi anche della formalizzazione attraverso l’adozione di un modello di *resource description*. Questo approccio è qualificabile come un meta-livello che lega l’interpretazione del contenuto di una fonte all’autore, ovvero l’agente, dell’interpretazione stessa. Perché ogni dichiarazione finalizzata a costruire una relazione fra URIs, o fra un URI e un letterale, è un procedimento ermeneutico. Ogni dichiarazione, a sua volta, si basa sulla consultazione di fonti che consentono all’interprete di supportare il proprio asserto. Il contesto, inteso archivisticamente come il legame fra la fonte e il soggetto produttore o detentore di quella fonte, diventa allora il meccanismo per preservare il dato stabilendone la provenienza, avvicinando quindi il problema della

conservazione logica a quello dell'autenticità. In senso estensivo il contesto lega l'oggetto a chi, su quell'oggetto, ha espresso un'idea, formulato un'opinione o veicolato un parere utile a stabilire il contenuto dell'oggetto stesso, sulla base dell'assunzione di un individuale punto di vista o alla luce delle proprie conoscenze, ovvero della consultazione delle fonti.

Provenance e ontologie

Date queste premesse il nostro gruppo di ricerca ha elaborato un'ontologia che, partendo da un lavoro preliminare sulla formalizzazione della *provenance*, traducesse l'atto interpretativo corredandolo dell'opportuno contesto necessario ad avvalorare l'autorevolezza dell'asserto.

Il ruolo delle ontologie nel processo di costruzione di LOD assume un valore duplice: da un lato permette di fornire il giusto spessore semantico alle triple; dall'altro fornisce un modello che possa essere scalabile in situazioni diverse. Questo connubio è lo strumento per arricchire i LOD di informazioni utili a garantire la qualità del contenuto informativo dei dati attraverso meccanismi di preservazione concettuale, oltre che logica, e quindi di conservazione della conoscenza.

HiCO, "Historical Context Ontology"[1] (Daquino e Tomasi 2015), è una ontologia nata esattamente per per descrivere il contesto di oggetti culturali in quanto prodotto del *workflow* interpretativo di un agente. Questo modello consente di gestire affermazioni autorevoli su asserzioni, potenzialmente anche contraddittorie, prevedendo meccanismi che possano anche consentire la coesistenza di interpretazioni esito di processi analitici difformi. Ogni asserto è concepito infatti come una lettura soggettiva di un interprete su uno specifico livello della fonte da cui tale lettura viene estratta, come ad esempio un testo, o meglio l'"espressione" del testo della fonte (cfr. fig. 1). Un testo è per sua natura un oggetto multi-livellare: HiCO, acquisendo il modello FRBR[2], cerca di far fronte a questa complessità per determinare i livelli della rappresentazione dell'oggetto informativo (l'opera, la sua espressione, le diverse possibili manifestazioni e l'item).

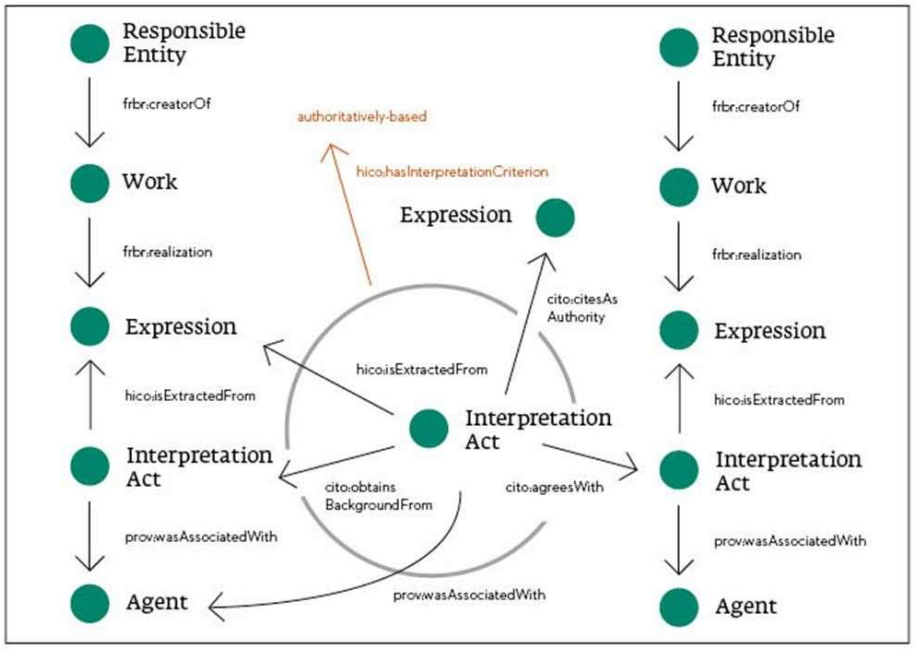


Fig. 1 Il modello HiCO.
A sinistra l'oggetto nativo da descrivere, a destra un altro oggetto culturale, espressione dell'interpretazione dell'oggetto nativo da parte di un agente

1 HiCo, Historical Context Ontology: http://purl.org/emmedi/hico
2 Frbr: http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

Partendo dal problema della *provenance* (Daquino et al 2014), già formalizzato nell'altra ontologia sviluppata dal gruppo di ricerca e cioè PRoles[3] (cfr. fig. 2), HiCO estende l'analisi alle entità coinvolte nel processo interpretativo, proponendo un *workflow* dettagliato per stabilire, secondo criteri condivisibili, quali asserzioni possano essere considerate autorevoli e/o meglio documentate rispetto ad un'asserzione priva di contesto che ne avvalori il contenuto.

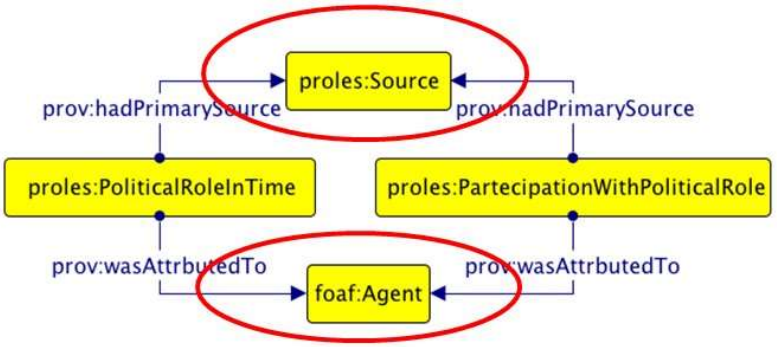


Fig. 2 Il terzo livello di Proles: la dichiarazione di provenance.
Un agente dichiara che - in una fonte (un documento) - è attestato che un altro agente, che ha un ruolo specifico, ha partecipato a un determinato evento

In particolare HiCO – che riusa FRBR[4], Pro[5], N-ary[6] e Prov-o[7] – permette di descrivere l'atto interpretativo di un agente (hico:InterpretationAct class) sulla base di 'tipologie' di Interpretazioni (hico:hasInterpretationType property) e 'criteri' interpretativi (hico:hasInterpretationCriterion property), e, attraverso l'ontologia CiTO[8], consente di creare relazioni fra interpretazioni riferite allo stesso oggetto culturale fatte da autori/editori/commentatori diversi.

I LOD e la preservazione della conoscenza

L'impiego di HiCO come strumento per formalizzare le interpretazioni è stato testato su due *datasets*: un archivio fotografico e un'edizione di lettere. L'ontologia è stata utilizzata quindi per affrontare due opposti metodi di rappresentazione del contenuto di un oggetto culturale: quello data-centrico tipico dei sistemi di descrizione catalografica e quello documento-centrico tipico delle *scholarly editions*. Se quindi nel primo caso lo sforzo di traduzione in LOD si concentra sul processo di identificazione univoca degli elementi atomici e di conversione dei campi descrittivi in classi e proprietà, nel secondo caso il documento deve essere analizzato per estrarre stringhe sotto forma di dati strutturati. Ma in entrambi i casi è necessario legare il processo interpretativo, ovvero la compilazione dei campi della scheda da un lato e il markup dall'altro, all'agente che ha formulato l'interpretazione, e legare l'interpretazione alla fonte che la supporta. Ovviamente l'agente che ha provveduto alle informazioni descrittive dell'oggetto non necessariamente sarà il creatore della tripla. Questo significa che la *provenance* dovrà garantire l'adeguata documentazione di ogni agente coinvolto nel processo ermeneutico: l'interprete della fonte, l'autore della fonte, il creatore/interprete della tripla.

3 Proles: http://www.essepuntato.it/lode/http://www.essepuntato.it/2013/10/politicalroles
4 Frbr: purl.org/spar/frbr
5 Pro: http://purl.org/spar/pro/
6 N-ary: http://www.ontologydesignpatterns.org/cp/owl/naryparticipation.owl
7 Prov-o: https://www.w3.org/TR/2013/REC-prov-o-20130430/
8 CiTO, the Citation Typing Ontology, http://purl.org/spar/cito/

Fig. 3 Esempio di interpretazione dello stesso oggetto, con diversa attribuzione di paternità e di datazione della stessa opera, e autorevolezza dell'interpretazione di Zeri

Il primo caso, si riferisce all'esperienza condotta sull'archivio fotografico Zeri[9], un catalogo ad oggi considerato tra i più imponenti repertori di arte italiana sul Web. Il progetto completo[10] ha previsto, assieme alla modellazione ontologica degli standard di descrizione catalografica Scheda F e Scheda OA[11], la produzione e pubblicazione del dataset in RDF, la creazione di collegamenti al LOD cloud e alle più opportune authorities, e la definizione di un'interfaccia per il browsing (Gonano et al 2014; Daquino et al in corso di stampa).

La partecipazione della Fondazione al progetto Pharos (Reist et al., 2015), un consorzio internazionale di archivi fotografici, permetterà di sperimentare l'uso di HiCO per confrontare le diverse attività, condotte dai catalogatori delle istituzioni coinvolte, nella descrizione delle stesse risorse (per esempio l'attribuzione di paternità ad un'opera d'arte rappresentata in una fotografia) e trarne quindi delle inferenze (cfr. fig. 3). Attraverso regole SWRL applicate ai rapporti tra le fonti, ai criteri e alle informazioni di contesto utilizzati da un agente per spiegare la sua interpretazione, potremmo infatti dedurre formalmente quando una interpretazione può essere considerata autorevole (Daquino et al 2016).

Il secondo caso è l'edizione digitale delle lettere, inviate/ricevute a/dal copista fiorentino Vespasiano da Bisticci, vissuto nell'arco del XV secolo[12] (Tomasi 2013).

In questo caso al testo delle lettere, come trascrizione diretta delle fonti primarie in formato XML/TEI, sono aggiunti una serie di commenti non strutturati che sono stati tradotti in RDF, istituendo una relazione fra il testo della lettera ed il commento effettuato dall'editore. Attraverso l'ontologia CITO è stato possibile stabilire il rapporto fra testo e commento e fra commento e fonti.

L'editore, nell'arricchire il testo con informazioni storiche, prosopografiche, codicologiche e

9 Catalogo della Fondazione Zeri: http://www.fondazionezeri.unibo.it/it

10 Dataset e documentazione: http://data.fondazionezeri.unibo.it/

11 La modellazione ontologica ha previsto l'uso di CIDOC-CRM, delle SPAR Ontologies e di HiCO, ma anche la creazione della FEntry Ontology e della OAEntry Ontology (cfr http://data.fondazionezeri.unibo.it/).

12 Vespasiano da Bisticci, *Lettere*: http://vespasianodabisticciletters.unibo.it/

culturali in senso estensivo, assume posizioni sulla base tanto delle proprie conoscenze quanto alla luce della consultazione di fonti. L'asserto lega dunque l'attribuzione di paternità dell'interpretazione alla fonte che corrobora tale posizione o, volendo anche, alla fonte che la confuta o che dichiara posizioni diverse (cfr. fig. 4).

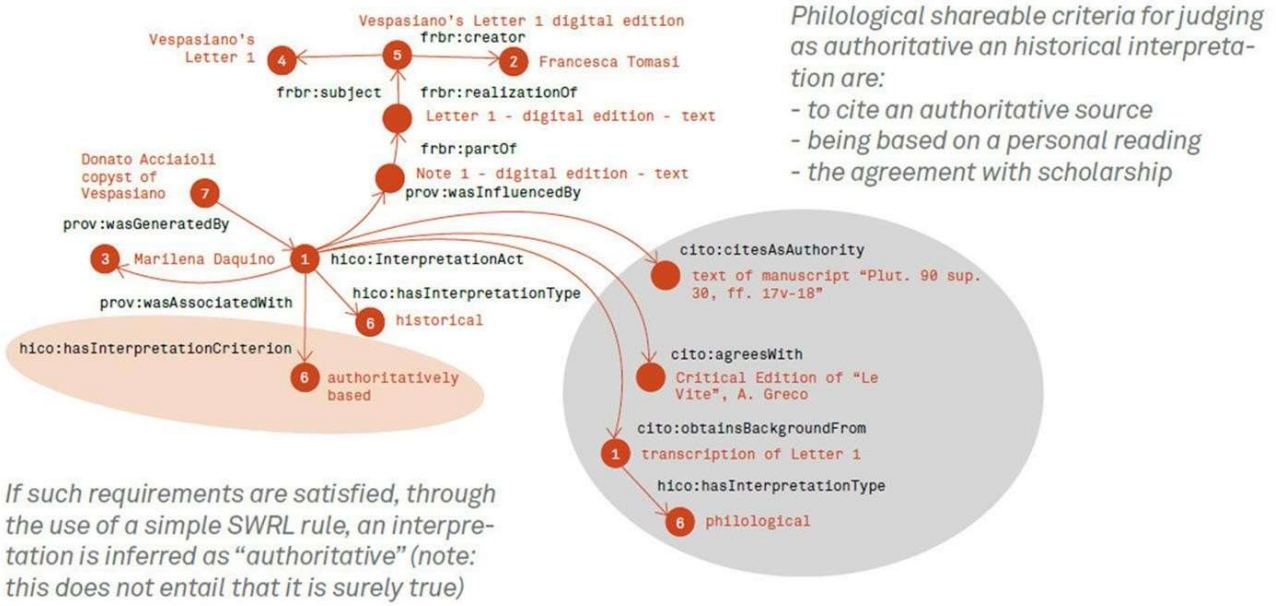


Fig. 4 Esempio del processo di interpretazione di un evento: dichiarazione della fonte e menzione dell'*agreement* di altri commentatori sulla stessa porzione interpretata

Conclusione

I dati della ricerca, rappresentati dalle schede di descrizione catalografica e dal markup del full-text, vengono preservati logicamente attraverso lo schema e concettualmente attraverso l'ontologia. Attraverso la costruzione di opportuni collegamenti fra l'interpretazione e il suo responsabile, ma anche fra l'interpretazione e la fonte utilizzata per dichiarare quella interpretazione, è possibile mettere in relazione interpretazioni diverse, per stabilire una tassonomia di casi, contemplando la contraddittorietà di asserti adeguatamente documentati. Questo significa anche mettere in condizione gli utenti di scegliere quale interpretazione ritenere più autorevole, in base all'attribuzione di paternità delle operazioni. La fiducia di un utente A è veicolata certo dalla conoscenza diretta di utente B, ma anche dalla capacità dello stesso utente B di documentare adeguatamente il contesto di un asserto, per esempio attraverso esplicita menzione delle fonti.

Un avvicinamento importante al *Web of Trust* quindi, per contrassegnare le asserzioni RDF al fine di garantirne l'autenticità e stabilirne la provenienza. Attraverso meccanismi, come firma digitale e crittografia, la rappresentazione ontologica degli asserti avvia il processo della fiducia come proprietà transitiva (se l'utente A si fida dell'utente B, il quale dichiara di fidarsi dell'utente C, si ottiene che A si fida anche di C).

Mentre l'uso di metadati consente, per concludere, la preservazione del dato e, attraverso lo schema logico, del contenuto informativo, con le ontologie la preservazione mira a garantire l'accesso alla conoscenza, attraverso il modello concettuale che attribuisce al metadato lo spessore semantico necessario a correlare il dato al suo contesto.

Potremmo allora chiudere con l'efficace posizione dell'importante progetto DigitalPreservationEurope (DPE):

> Digital preservation is about more than keeping the bits – those streams of 1s and 0s that we use to represent information. It is about maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its 'interrelatedness', and about securing information about the context of its creation and use (Ross 2012, 44).

Rifermenti bibliografici

Daquino, Marilena, Peroni, Silvio, Tomasi, Francesca, and Fabio Vitali. "Political Roles Ontology (PRoles): enhancing archival authority records through Semantic Web technologies." *Procedia Computer Science* 38 (2014): 60-67.

Daquino, Marilena, and Francesca Tomasi. "Historical Context (HiCo): a conceptual model for describing context information of cultural heritage objects." *Communications in Computer and Information Science* 544 (2015): 424-436.

Daquino, Marilena, Peroni, Silvio, Tomasi, Francesca, and Fabio Vitali. "The Project Zeri Photo Archive: Towards a Model for Defining Authoritative Authorship Attributions." In *Digital Humanities 2016: Conference Abstracts*, 472-474. Kraków: Jagiellonian University & Pedagogical University, 2016.

Daquino, Marilena, Peroni, Silvio, Tomasi, Francesca, and Fabio Vitali. In print. "Enhancing semantic expressivity in the cultural heritage domain: exposing the Zeri Photo Archive as Linked Open Data." arXiv.org. arXiv: 1605.01188 (in print for *Journal on Computing and Cultural Heritage* - JOCCH).

Gonano, Ciro Mattia, Mambelli, Francesca, Peroni, Silvio, Tomasi, Francesca, and Fabio Vitali. "Zeri e LODE. Extracting the Zeri photo archive to Linked Open Data: formalizing the conceptual model." In *Digital Libraries (JCDL)*, 289-298. London: IEEE, 2014.

Heath, Tom and Christian Bizer. "Linked data: Evolving the web into a global data space". *Synthesis lectures on the semantic web: theory and technology* 1.1 (2011): 1–136.

Lavoie, Brian and Richard Gartner. *Preservation Metadata (2nd edn.)*. Digital Preservation Coalition Technology Watch Reports, 2013.

Reist, Inge, Farneth, David, Stein, Samuel and Remigius Weda. "An Introduction to PHAROS: Aggregating Free Access to 31 Million Digitized Images and Counting". Paper presented at CIDOC 2015, New Dehli, India, September 5-10, 2015.

Ross, Seamus. "Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries." *New Review of Information Networking* 17.1 (2012): 43–68.

Schmachtenberg, Max, Bizer, Christian and Heiko Paulheim. *State of the LOD Cloud 2014*. Version 0.4, 08/30/2014. Accessed November 17, 2016. http://linkeddatacatalog.dws.informatik.unimannheim.de/state/

Tomasi, Francesca. "L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere". *Ecdotica* 9 (2013): 264-286.

SHORT PAPERS

L'edizione digitale del "Nuovo Liruti": il Dizionario biografico dei friulani

Stefano Allegrezza, Università degli Studi di Udine, stefano.allegrezza@uniud.it
Nicola Raffaele di Matteo, Altaviso-Algoweb, nicoladimatteo@gmail.com

Premessa

La maggior parte dei dizionari biografici oggi esistenti tende sempre più spesso a trasformarsi da edizione cartacea ad edizione on-line. Solo per citare alcuni tra gli esempi più conosciuti, la Treccani ha realizzato una edizione digitale del Dizionario Biografico degli Italiani[1], che risulta più facilmente fruibile e più aggiornata rispetto all'edizione cartacea (la quale - organizzata secondo un ordinamento alfabetico – con le voci relative alla lettera "A" pubblicate nel 1960 e quelle relative alla lettera "M" pubblicate nel 2011 – risulta attualmente del tutto obsoleta); il Dizionario Rosi del Rinascimento Risorgimentale[2] è completamente consultabile on-line, così come il dizionario biografico Donne e Uomini della Resistenza[3] dell'Associazione Nazionale Partigiani d'Italia, il Dizionario Biografico on-line dei Protestanti in Italia[4] della Società di Studi Valdesi e numerosi altri esempi che si ography[5] e il Diccionario Biográfico Español[6], solo per fare alcuni esempi significativi. Tuttavia, spesso questi dizionari on-line presentano un'interfaccia poco intuitiva e spesso non così "attraente" da invogliare l'utente nella navigazione. Inoltre non vengono quasi mai utilizzati gli strumenti del web setico che invece possono trovare una loro peculiare ed interessante applicazione proprio in questo particolare campo.

Il progetto che si intende presentare in questa sede vuole costituire una *best practice* per l'applicazione di metodologie innovative nella realizzazione di edizioni digitali dei dizionari biografici, costituendo, inoltre, un esempio di collaborazione tra umanisti ed informatici, senza la quale il progetto non sarebbe stato possibile.

L'idea del progetto

L'idea del progetto è scaturita dalla felice intuizione dell'Istituto Pio Paschini per la storia della Chiesa in Friuli che nel 2016 si propose di realizzare un'edizione digitale del "Nuovo Liruti. Dizionario biografico dei Friulani" (Scalon 2006-2011) in collaborazione con gli enti e le istituzioni che ne avevano promosso l'edizione a stampa[7]. Le voci già pubblicate sarebbero state riviste ed integrate da circa quattrocento profili bio-bibliografici del cd. "Supplemento Ongaro" (realizzato da Maiko Favaro sulla base dei manoscritti settecenteschi di Domenico Ongaro), e dalle voci dell'*Onomasticon*, previsto in sede di presentazione dell'intero lavoro.

Non si sarebbe trattato, quindi, semplicemente della trasposizione su base digitale dell'edizione a stampa del "Nuovo Liruti" o della messa a disposizione della sua versione elettronica (il file PDF) ma della creazione di un vero e proprio dizionario biografico on-line con un'interfaccia grafica "accattivante" e dotato di numerose funzionalità proficuamente utilizzabili sia dallo studioso che dal

1 Si veda <http://www.treccani.it/biografie>.
2 Si veda <http://www.dizionariorosi.it>.
3 Si veda <http://www.anpi.it/donne_e_uomini/>.
4 Si veda <http://www.studivaldesi.org/dizionario>.
5 Si veda <http://www.anb.org>.
6 Si veda <http://www.rah.es:8888>.
7 Il "Dizionario biografico dei friulani. Nuovo Liruti" è costituito da tre parti: "1. Il medioevo", pubblicata nel 2006; "2. L'età veneta", pubblicata nel 2009; "3. L'età contemporanea", pubblicata nel 2011.

semplice cittadino desideroso di approfondire la storia e la cultura della propria regione. Una volta realizzata l'edizione digitale, sarebbe stato possibile consultare da ogni continente una delle principali opere di riferimento per le ricerche sulla storia e sulla cultura friulana.

Il progetto vede come promotori, oltre all'Istituto Pio Paschini per la storia della Chiesa in Friuli, la Deputazione di storia patria per il Friuli, l'Istituto storico del libro antico (ISLA), la Società Filologica Friulana e il Dipartimento di Studi umanistici e del patrimonio culturale dell'Università degli Studi di Udine. Tra i partner del progetto vi sono l'Accademia San Marco di Pordenone, l'Accademia di Scienze lettere e arti di Udine, l'Archivio diocesano e Biblioteca patriarcale di Udine, l'Archivio di Stato di Gorizia, l'Archivio di Stato di Pordenone, l'Archivio di Stato di Udine, la Biblioteca civica Guarneriana di San Daniele del Friuli, la Biblioteca civica di Pordenone, la Biblioteca civica "Vincenzo Joppi" di Udine, la Biblioteca statale Isontina di Gorizia, l'Istituto di storia sociale e religiosa di Gorizia, l'Università popolare di Udine. La Direzione scientifica è di Cesare Scalon e Claudio Griggio mentre la Direzione tecnico-scientifica è affidata a Stefano Allegrezza. La realizzazione è stata affidata a Nicola Raffaele Di Matteo.

Elementi di forza del progetto

Migliore fruibilità

La fruibilità dei contenuti che verranno resi disponibili on-line migliora rispetto all'edizione a stampa, dal momento che ciò rende possibile non solo una modalità di lettura "sequenziale" (come avviene per l'edizione cartacea) ma anche una modalità di lettura "ipertestuale" (sfruttando i *link* che verranno inseriti nel testo ad evidenziare i collegamenti più interessanti).

Inoltre, i contenuti resi disponibili nell'edizione on-line saranno fruibili secondo modalità di navigazione molteplici: non solo in ordine alfabetico – com'è ovvio - ma anche seguendo i percorsi tematici proposti oppure creando dei propri percorsi personalizzati. Ad esempio, sarà possibile la consultazione per aree tematiche (scienze sociali e storia, arte, lingua e letteratura, scienze naturali, tecnologia e scienze applicate, etc.) oppure utilizzando percorsi trasversali.

Infine, l'edizione on-line renderà disponibili i contenuti in una forma più ricca e fruibile attraverso i moderni dispositivi elettronici: non solo personal computer ma anche tablet, smartphone, phablet, etc., potendo così raggiungere una platea molto più vasta (non solo il ricercatore o lo studioso ma anche lo studente, la persona comune, etc.) e potenzialmente senza confini di tempo o di spazio.

Maggiori possibilità di effettuare studi e ricerche

La versione digitale consentirà di compiere sia ricerche *semplici* (di tipo *full-text*) sull'intero corpus dei contenuti presenti che ricerche *avanzate* (specificando gli opportuni criteri di ricerca in maniera da raggiungere subito il contenuto che si desidera visualizzare). Rispetto alla versione a stampa, che sostanzialmente consente solo ricerche basate sull'ordinamento alfabetico, con la versione on-line l'utente avrà la possibilità di individuare il personaggio o il fatto di suo interesse attraverso i diversi criteri di ricerca che saranno definiti; ad esempio: non solo, com è ovvio, per nome e cognome, ma anche per date o luoghi di nascita e di morte, per periodo storico, per date o luoghi legati al personaggio, per professione (giuristi, letterati, tipografi...), per nome del curatore della voce, etc. Ad esempio, sarà possibile sapere quali sono i friulani illustri legati ad una certa città o territorio; si potrà sapere quali sono gli "anniversari " che cadono in un certo anno e le "ricorrenze" da celebrare (ad esempio, partendo dalle date significative della vita di un personaggio, come la sua data di nascita e di morte); si potrà affinare ulteriormente la ricerca per sapere quali sono gli scrittori, i poeti, i narratori, i filologi, i registi, gli artisti, gli sportivi, etc. legati ad una certa città o territorio; sarà anche possibile effettuare ricerche più mirate, combinando opportunamente le varie chiavi di ricerca. L'individuazione dei metadati da associare a ciascuna voce del dizionario sarà oggetto di approfondito studio per raggiungere la più ampia flessibilità nelle successive fasi di ricerca.

Aggiornamento costante ed in tempo reale

Rispetto all'edizione cartacea, l'edizione on-line presenta indubbiamente il vantaggio di costituire un riferimento sempre aggiornato ed aggiornabile con continuità. Una volta inserita una nuova voce (ovviamente da parte del comitato di redazione), questa sarà immediatamente visibile on-line per cui il dizionario sarà sempre e costantemente aggiornato. Anche la correzione di eventuali errori o refusi è molto facilitata ed immediata: la versione on-line sarà sempre quella riveduta e corretta (questo non è ovviamente possibile con l'edizione a stampa). Rispetto all'edizione a stampa l'edizione on-line costituirà l'occasione per rendere il lavoro più esaustivo: potranno essere aggiunte nuove voci biografiche, sia per completare l'opera che per aggiornarla alle mutate esigenze e sensibilità.

Scalabilità senza limiti

Non vi sono limiti relativamente alla quantità di informazioni che potranno essere ospitate dal Dizionario Biografico dei Friulani; il dizionario potrà essere ampliato nel tempo aggiungendo non solo voci inizialmente non presenti, ma anche contenuti di qualsiasi genere (si pensi, ad esempio, alle registrazioni audio che possono essere associate alla scheda di un musicista o alle registrazioni video che possono essere associate alla scheda di un regista, etc.). Con l'edizione digitale non vi sono sostanzialmente limiti per quanto riguarda le possibilità di espansione del Dizionario.

Interfaccia grafica "accattivante"

Particolare impegno sarà dedicato alla realizzazione grafica del Dizionario Biografico dei Friulani al fine di proporre una interfaccia di navigazione attraente ed accattivante, che sappia presentare i contenuti in maniera tale da "catturare" l'attenzione non solo degli studiosi e degli esperti ma anche di un pubblico variegato, costituito da utenti di ogni età, estrazione sociale e cultura. Per questo, alle "tradizionali" maschere di ricerca che si utilizzano da tempo in progetti di questo genere (generalmente destinate solo ad un utilizzo "specialistico"), saranno affiancate delle interfacce completamente riviste e che punteranno sull'utilizzo di diversi codici mediatici per "coinvolgere" gli utenti ed invogliarli nella navigazione all'interno del Dizionario.

Utilizzo del web semantico e di RDF

Nell'attività di redazione delle voci biografiche si è deciso di inserire una annotazione semantica ad opera di un gruppo di lavoro costituito da umanisti. L'annotazione semantica consente di creare dei linked data esposti in formato RDF, che vengono generati tenendo conto di tutta l'annotazione inserita, sia nel testo che esterna. In aggiunta è possibile utilizzare il linguaggio di interrogazione SPARQL per le interrogazioni che possono essere anche molto complesse (ad esempio: quali sono gli organisti che hanno lavorato nel periodo 1820-1840 nella città di Aquileia? oppure: quali sono gli uomini illustri che la città di San Daniele del Friuli deve festeggiare nell'anno 2017 per una ricorrenza?

Allo stato attuale il progetto di trova in fase di avanzata realizzazione: buona parte delle oltre 2700 voci biografiche sono state riviste per adattarle allo strumento informatico ed annotate semanticamente utilizzando un tool appositamente sviluppato. In particolare è stata operata la migrazione delle 2700 voci biografiche, disponibili nel formato di descrizione di pagina PDF utilizzato per l'edizione a stampa, in un formato ipertestuale con riconoscimento automatico della posizione delle immagini e dei riferimenti bibliografici all'interno della struttura. Il risultato è stato raggiunto utilizzando uno strumento open source (pdf2html) che ha generato dei file XML in cui erano riportati, utilizzando opportuni attributi, le indicazioni sulla formattazione del testo; elaborando i file XML è stato possibile evidenziare dei pattern comuni che hanno consentito di associare ad essi

l'aspetto semantico e ricostruire così gli articoli nella loro struttura (titolo, sottotitolo, corpo, bibliografia) originaria; inoltre è stato possibile estrarre i primi metadati esterni (ad esempio, l'autore della voce biografica).

Affinché il gruppo di lavoro che ha operato sulla revisione delle voci biografiche potesse revisionare i testi e inserire i metadati necessari per l'annotazione semantica (ad esempio: forme alternative del nome, data e luogo di nascita, data e luogo di morte, luoghi di attività, date significative, etc.) in modo semplice, si è scelto di utilizzare un CMS ampiamente diffuso, Wordpress, dotandolo di plugin sviluppati internamente. Il gruppo di lavoro ha potuto così utilizzare l'interfaccia grafica di Wordpress, semplice ed immediata, ed inserire direttamente nel testo i metadati necessari alla notazione semantica in maniera estremamente intuitiva: è sufficiente selezionare l'elemento "oggetto" ed attribuirgli il tag appropriato (che rappresenta la "proprietà") scegliendolo da un elenco a discesa di tipo contestuale. L'annotazione *in text*[8] potrà così essere completata agevolmente e con soddisfazione del gruppo di lavoro.

È stata predisposta anche una sezione per l'inserimento dell'annotazione semantica fuori dal testo, lasciando ai redattori la possibilità di inserire metadati e rimandando a una fase successiva a quella della revisione dei testi la creazione di un vocabolario controllato. È stata scelta questa soluzione per evitare le discussioni per la creazione di una ontologia interna e per evitare il tempo di apprendimento di una esterna. Anche la bibliografia viene annotata ed utilizzata per creare delle strutture RDF che descrivono le risorse esterne. Il testo elaborato viene quindi letto dinamicamente da un parser che crea gli elementi RDF che possono tenere conto delle annotazioni sia *in text* che esterne. È presente una sezione di configurazione dove è possibile scegliere l'ontologia da utilizzare per rappresentare i dati all'esterno e creare le corrispondenze con il vocabolario controllato interno.

In sintesi

Il progetto Dizionario biografico dei Friulani ("Nuovo Liruti on-line") si pone l'ambizioso obiettivo di essere non solamente la "versione" digitale dell'edizione a stampa del "Nuovo Liruti" ma uno dei più ricchi e strutturati depositi di informazione culturale e storica del web italiano − con la possibilità di raggiungere un pubblico molto più ampio e potenzialmente illimitato rispetto all'edizione cartacea − caratterizzandosi così come una delle iniziative più importanti sul piano culturale nell'ambito del più vasto progetto su "L'Identità Culturale del Friuli (ICF)". La conclusione del lavoro è prevista per la prima metà del 2017.

Riferimenti Bibliografici

Scalon, Cesare, Claudio Griggio, Ugo Rozzo e Giuseppe Bergamini (a cura di). 2006-2011. *Nuovo Liruti. Opera completa*. Udine: Editrice Forum

Di Iorio, Angelo, et. al. Describing bibliographic references in RDF, 2014, 11th ESWC 2014 (ESWC2014). http://ceur-ws.org/Vol-1155/paper-05.pdf

Schrott Maximilian, et.al *From Biographies to Data Curation – the Making of www.deutsche-biographie.de*. 2015. http://ceur-ws.org/Vol-1399/paper3.pdf

8 Si veda http://ontotext.com/knowledgehub/fundamentals/semantic-annotation/

L'impatto culturale e sociale dei dati archeologici nella tutela e valorizzazione del paesaggio storico. Un case study: il XV Municipio di Roma

Margherita Bartoli, Sapienza Università di Roma, margherita.bartoli@hotmail.it

Lo studio esplora il complesso rapporto esistente tra la pianificazione territoriale e le evidenze archeologiche nei processi di salvaguardia e valorizzazione del patrimonio culturale. In particolare, si illustra un progetto che ha come obiettivo la raccolta, la gestione di documentazione (fotografica, archivistica e bibliografica) e la creazione di un GIS, finalizzati alla promozione e alla valorizzazione dei beni culturali presenti nel territorio del XV Municipio del Comune di Roma[1].

Questo lavoro si inserisce, pertanto, in quel filone di ricerca archeologica che si occupa dell'elaborazione di protocolli operativi per la gestione di dati archeologici e che è rivolta soprattutto al controllo e al miglioramento qualitativo dei flussi di lavoro nella prassi della catalogazione digitale, uno dei settori di indagine più problematici e complessi nell'applicazione delle nuove tecnologie in ambito archeologico.

Nello studio conoscitivo del bene culturale, la "catalogazione"[2], è intesa, infatti, come procedura che consente di collocare (attraverso sistemi di classificazione, categorie e codici terminologici specifici), ogni prodotto dell'attività umana nello spazio geografico, di definirlo nelle sue caratteristiche materiche e morfologiche, tecnico-esecutive e decorative e, infine, di situarlo cronologicamente e di attribuirlo a una classe di produzione, a un ambito culturale o a un autore. L'obiettivo primario della ricerca è quello di individuare alcune "regole" per strutturare i dati archeologici, affinché sia possibile estrarre conoscenza dalle informazioni raccolte durante la ricerca archeologica.

I dati raccolti provengono da un'attività di censimento delle evidenze archeologiche presenti nel territorio del XV municipio[3] e relative a contesti storico-archeologici assai differenziati. A tal fine, si è deciso di organizzare i dati all'interno di un database sperimentale e di rappresentarli grazie ad un GIS, con l'obiettivo primario di renderli accessibili ad una vasta gamma di utenti, specialisti e non, attraverso una consultazione il più possibile semplice, intuitiva e possibilmente libera. Questo ha naturalmente comportato un'attenta valutazione critica, relativa ai metodi di classificazione e catalogazione del materiale raccolto ritenuti più idonei per la loro più immediata fruizione, ma anche per il loro riuso in contesti scientifico-culturali diversi.

Il database[4] è stato strutturato in due tabelle: la prima chiamata "Scheda Monumento" e una secondaria detta "Scheda di Dettaglio" relazionate tra loro grazie all' "ID Monumento", comune ad entrambe le schede e univoco per ogni oggetto catalogato, anche nel GIS.

La "Scheda Monumento" è la scheda descrittiva dell'evidenza archeologica e ricalca quella ufficiale[5] proposta dall'ICCD, ma semplificata nei suoi tratti essenziali per renderne più semplice e veloce la

1 L'indagine, svolta con il patrocinio della Sovrintendenza capitolina e il XV Municipio di Roma, è circoscritta all'area tra Ponte Milvio, la via Flaminia (fino a Malborghetto) e la Cassia (fino alla sua intersezione con il Raccordo anulare).

2 La catalogazione e la documentazione sono, infatti, inserite tra i principi generali del D. L.vo n. 42 del 22.01.2004.

3 I dati presenti in questo database sono la fusione tra il lavoro di ricerca sul campo, la ricognizione e la campagna fotografica, la ricerca bibliografica ed infine l'integrazione tra dati forniti dal comune e dallo stato (elenchi in *open data* di alcune categorie di monumenti e della loro condizione giuridica e i vincoli a cui sono sottoposti).

4 La scelta del software per il database è ricaduta, in questa fase iniziale, su Microsoft Access e si è deciso di proteggere il sistema con una licenza *Creative Commons*, la CC BY SA.

5 Scheda RA versione 3.00. La scelta di utilizzare il modello ministeriale è stata fatta al fine di avere dati strutturati e interoperabili senza grossi problemi con quelli prodotti dal Ministero.

consultazione, sottolineando la possibilità di un utilizzo trasversale del materiale sia per scopi di ricerca che per la valorizzazione territoriale.
È importante sottolineare la scelta di utilizzare termini indicizzati da vocabolari controllati, perché essa impone al catalogatore l'uso di strumenti di controllo terminologico, contribuendo così a ottimizzare il reperimento dell'informazione e a migliorare sensibilmente la sua comprensione da parte dell'utente[6].
Infine, grazie ad un collegamento ipertestuale è possibile accedere direttamente alla piattaforma GIS costruita con software *open source* QGis, con base cartografica la CTR del Lazio 1:10.000 in formato *raster*, a cui è sovrapposto il poligono del perimetro del XV Municipio[7].
Ad ogni evidenza archeologica è associata una scheda, composta da pochissimi campi informativi, che riprendono in sintesi la scheda del monumento presente sul database.
Con questo strumento, attraverso la realizzazione di buffer, ossia l'individuazione e delimitazione di un'area intorno a un "oggetto cartografico" (misurata, in questo caso, in unità di distanza)[8], è inoltre possibile determinare il "potenziale archeologico", cioè la probabilità più o meno alta, di rinvenire una stratificazione archeologica, di maggiore o minore rilevanza. Si tratta, pertanto, di uno studio complesso, che produce come risultato finale un'informazione geo-spaziale (in GIS), in base alla quale è lecito predire l'esistenza di una stratificazione archeologica conservatasi in un luogo e pertanto di impedire o controllare la fabbricazione di edifici o l'utilizzo dell'area per scopi non idonei alla tutela del bene culturale.
Grazie alla creazione di questo sistema di gestione delle informazioni, per quanto semplice e in fase embrionale, si è resa possibile una lettura del territorio molto più dettagliata e completa di quanto non lo fosse con la semplice ricerca bibliografica o cartografica. L'attività di ricerca si propone altresì di costruire una carta archeologica del Municipio e un sistema di fruizione dei dati finalizzati al loro riuso in ambito scientifico, didattico e turistico.

Riferimenti Bibliografici

Amendolea B. (a cura di) 1999. *Carta archeologica e pianificazione territoriale: un problema politico e metodologico. Primo incontro di studi (Roma 1997).* Roma: F.lli Palombi.

Anichini F., Fabiani F., Gattiglia G., Gualandi M.L. 2012. *MAPPA. Metodologie Applicate alla Predittività del Potenziale Archeologico.* vol.1-2. Pisa.

Antinucci F., "Beni artistici e nuove tecnologie", in Galluzzi P., Valentino P.A. (a cura di). 1997. *I formati della memoria. Beni culturali e nuove tecnologie alle soglie del terzo millennio*, pp. 120-131. Firenze.

Antinucci F. 1998. *La realtà virtuale come strumento di conservazione del sapere*. http://www.mediamente.rai.it/biblioteca/biblio.asp?Id=12&tab=int

Azzari M. Aprile-agosto 2002. "Beni Ambientali e Culturali e GIS", in *Geostorie, Bollettino e Notiziario del Centro Italiano per gli Studi Storico-Geografici* 10, nn. 1-2. Roma.

Azzena G., Tascio M. 1996. "Il Sistema Informativo Territoriale per la Carta Archeologica d'Italia", in Marchi M.L., Sabbatici G. (a cura di), *Venusia (Forma Italiae*, 37), pp. 281-297. Firenze.

Barker G. 1995. *A Mediterranean Valley. Landscape Archaeology and Annales History in the Biferno Valley*. London

Bianchini M. 2008. *Manuale di rilievo e di documentazione digitale in archeologia.* Roma. http://www.aracneeditrice.it/pdf/9788854818262.pdf

6 Per quanto riguarda la "Tipologia" è stato scelto come vocabolario di riferimento il "Soggettario di Firenze" in quanto nelle schede RA dell'ICCD questo campo è libero mentre per quanto riguarda la "Cronologia" e la "Condizione giuridica e vicoli", ho optato per il PICO che rispondeva perfettamente alle mie esigenze di lavoro.

7 Il tutto è georeferenziato con il sistema di riferimento Monte Mario/Italy zone 2 (EPSG 3004).

8 Un buffer è utile per effettuare 'analisi di prossimità ossia una tecnica analitica utilizzata per determinare la relazione tra un punto selezionato ei suoi vicini

Bonamico S., Colini A. M., Fidenzoni P. (a cura di). 1968. "La Carta storico-monumentale dell'agro Romano", in *Capitolium – Rivista di Roma, Quaderni di Urbanistica Romana*, 5,6,7.
Brogiolo G. P. 1996. "Conclusioni"*,* in *La fine delle ville romane: trasformazioni nelle campagne tra tarda antichità e altomedioevo*. Mantova.
Caffo R. 2012. *2002-2012: dieci anni di networking nel digital cultural heritage*, Ministero per i beni e le attività culturali Istituto centrale per il catalogo unico delle biblioteche italiane.
Calaon, D., Pizzinato, C. 2011. "L'analisi archeologica nei processi di valutazione ambientale. Proposta metodologica in ambiente GIS", in *Archeologia e Calcolatori*, 22, pp. 413-439. Roma.
Corti L. 1999. *I Beni Culturali e la loro catalogazione*. Torino, Paravia.
Calci C. 2005. *Roma archeologica. Le scoperte più recenti della città antica e della sua area suburbana*. Roma.
Calci C., Messineo G. 1984. *La Villa di Livia a Prima Porta*. Roma.
Caserta E. 2010. "Mosaici e pavimenti in opus sectile nella Villa di Lucio Vero sulla Via Cassia a Roma. Indagini archeologiche negli anni 2005-2009", in *Atti del XV colloquio dell'associazione italiana per lo studio e la conservazione del mosaico*, pp. 467-478. Tivoli.
Caserta E. 2010-2011. "Roma (via Cassia): La villa di Lucio alla luce delle recenti indagini archeologiche", in *Notizie degli scavi di antichità*, ser. 9, 21-22, pp. 53-159. Roma
Caserta E., "La Villa di Lucio Vero sulla Via Cassia a Roma in località Acquatraversa", in *Journal of Field Archaeology*, 28, 2015, pp. 179-191. Cambridge.
Cifani G. 2013. "Per una definizione storica dei Falisci, tra identità, cultura e territorio", in Cifani G., (a cura di), *Tra Roma e l'Etruria. Cultura, identità e territorio dei Falisci*, *Acts of the British School Seminar, Rome 2011*, pp. 1-53. Roma.
Carbonara A., Messineo G., Pellegrino A. 1996. *La necropoli etrusca di Volusia.* Roma.
Comune Di Roma, Regione Lazio. 1994. "La cartografia dei beni storici, archeologici e paesistici nelle grandi aree urbane dal censimento alla tutela", in *Atti del Convegno (Roma, 26-27-28 aprile 1990)*. Roma.
Comune Di Roma. 2003. *Nuovo Piano Regolatore di Roma*. *Norme tecniche di attuazione*. Roma.
Colonna G. 2007. "Veii", in *CIE* II.2.5, pp. 3-16. Roma.
D'Ambrosio I., Drimmer A., Pascucci P., Rusca F. 2003. "La catalogazione promossa dalla regione Lazio nei musei archeologici: dalle schede di carta alla banca dati condivisa", in *Archeologia e Calcolatori*, 14, pp. 33-71. Roma.
D'Andrea A. 2006. *Documentazione archeologica, standard e trattamento informatico*. Budapest.
D'Andrea A., Niccolucci F. 2001. "L'Informatica dell'archeologo: qualche istruzione per l'uso", in *Archeologia e Calcolatori*, 12, pp. 199-220. Roma.
De Francesco D. 2014. *Dall'età imperiale alla tarda antichità*, pp. 54-62, Roma.
De Santis A. 1997. "Alcune considerazioni sul territorio veiente in età orientalizzante e arcaica", in Bartoloni G. (a cura di), *Le necropoli arcaiche di Veio. Giornata di studio in memoria di Massimo Pallottino*, pp. 101-143. Roma.
Di Nezio P., Maderni M.E., Rossi P. 2005. "Il censimento dell'archeologia industriale a Roma, Atti del Convegno Archeologia industriale. La conservazione della memoria (Roma, 8-9 maggio 2003)", in *Quaderni di Patrimonio Industriale*, 1, pp. 131-144. Roma.
Farinetti E. 2012. *I paesaggi in archeologia: analisi e interpretazione*. Roma.
Ferrari O. 1989. "Esperienza archeologica e catalogazione", in *Bollettino di Archeologia*, 1, pp. 24-25. Roma.

Ferrari O. 1990. “La catalogazione dei beni archeologici e le tecnologie informatiche”, in *Archeologia e Calcolatori*, 2, pp. 113-17. Roma.
Francovich R., Valenti M. 2000. “La piattaforma GIS dello scavo ed il suo utilizzo”, in Brogiolo G.P., (a cura di), *II Congresso Nazionale di Archeologia Medievale*, (Brescia, settembre 2000), pp.11-20. Firenze.
Galluzzi P., Valentino P.A. (a cura di). 1997. *I formati della memoria. Beni culturali e nuove Tecnologie alle soglie del terzo millennio*, Firenze.
Gregori G.L. 2012. “Recinti d’incerta identificazione”, in Rossi D. (a cura di), *Sulla via Flaminia. Il mausoleo di Marco Nonio Macrino*, pp. 162-164. Milano.
Gregori G.L. 2012. “Le sei nuove stele di militari”, in Rossi D. (a cura di), *Sulla via Flaminia. Il mausoleo di Marco Nonio Macrino*, pp. 165-170. Milano.
Gregori G. L. 2012. “Vita e gesta del senatore bresciano Marco Nonio Macrino”, in Rossi D. (a cura di), *Sulla via Flaminia. Il mausoleo di Marco Nonio Macrino*, pp. 286-301. Milano.
Gregori G. L., “Il ′sepolcreto′ di militari lungo la via Flaminia. Nuove stele del V-VI miglio”, in *Archeologia Classica,* Vol. LXIV*,* Roma 2013, pp. 349-366
Grossi G. L., Messineo G., Petracca L., Vigna L.M. 1985. “Contributi alla ricostruzione della rete viaria antica nel settore nord del suburbio di Roma, in Archeologia Laziale” in *Quaderni del Centro di studio per l’archeologia etrusco-italica*, 7, pp. 142-145. Roma.
Guermandi M. P. 1996. “L’informatica come risorsa decisiva nella gestione del patrimonio archeologico: le attività dell’Istituto Beni Culturali della Regione Emilia Romagna”, in *Archeologia e Calcolatori*, 7, pp. 837-848. Roma.
Guermandi M.P. (a cura di). 2001. *Rischio archeologico: se lo conosci lo eviti. Atti del Convegno di studi su cartografia archeologica e tutela del patrimonio (Ferrara 2000).* Firenze: All’Insegna del Giglio.
Kahane A., Murray Threipland L., Ward Perkins J. B. 1968. “The Ager Veientanus, north and east of Veii”, in *Papers of the British School at Rome*, XXXVI. London.
Mancineli M. L. (a cura di), *Nota introduttiva alle normative per la catalogazione dei beni culturali*, ICCD. Roma. http://www.iccd.beniculturali.it/getFile.php?id=280
Mancinelli M. L. 2004. “Sistema informativo generale del catalogo: nuovi strumenti per la gestione integrata delle conoscenze sui beni archeologici”, in *Archeologia e Calcolatori*, 15, pp. 115-128. Roma.
Mari Z. 2004. s.v. “Cassia, via”, in *LTUR Suburbium*, II, pp. 65-75. Roma.
Mastrodonato V. 1999-2000. “Una residenza imperiale nel suburbio di Roma: la villa di Lucio Vero in località Acquatraversa”, in *Archeologia Classica*, pp. 157-235. Roma.
Messineo G.1991. *La via Flaminia da Porta del Popolo a Malborghetto.* Roma.
Messineo G. 2004. s.v. “Flaminia, via”, in *LTUR Suburbium*, II, pp. 252-259. Roma.
Messineo G., Petracca L., Vigna L.M. 1985. “Via Cassia, Km, 11. Località Casale Ghella (circ. XX)”, in *Bullettino della Commissione Archeologica Comunale di Roma,* XC, pp. 177-184. Roma.
Messineo G., Vigna L. M. 1987-1988. “Località Casale Ghella (circ. XX)”, in *Bullettino della Commissione Archeologica Comunale di Roma*, XCII, pp. 504-509. Roma.
Moscati P. 2002. “L’informatica in archeologia”, in *Il mondo dell’archeologia*, I, pp. 318-323. Roma: Istituto della Enciclopedia Italiana Treccani.
Mucci A. 1997-1998. “Antonio Maria Colini e la Carta dell’agro Romano”, in *Rendiconti della Pontificia Accademia Romana di Archeologia*, LXX. pp. 267-279. Roma.
Niccolucci F., Pardi G., Zoppi T. 1996. “Un archivio georeferenziato di insediamenti archeologici”, in *Archeologia e Calcolatori*, 7, pp. 161-177. Roma.

Orlandi T. 1996. “Formalizzazione dei dati, semiotica e comunicazione”, in *Archeologia e Calcolatori*, 7, pp. 1247-1258. Roma.
Piranomonte M. 2013. “Costantino e i luoghi della Battaglia”, in *Costantino 313 d.C.*, pp. 27-32. Milano.
Piranomonte M. 2012-2013. “Nuovi ritrovamenti sulla via Flaminia”, in *Atti della Pontificia Accademia Romana di Archeologia, Rendiconti*, 85, pp. 129-170. Roma.
Piranomonte M. (a cura di). 2014. *Via Flaminia*. Roma.
Ricci A. (a cura di). 2002. *Archeologia e urbanistica, XII Ciclo di Lezioni sulla ricerca applicata in archeologia (Certosa di Pontignano 2001)*. Firenze: All’Insegna del Giglio.
Rossi D. (a cura di). 2012. *Sulla via Flaminia. Il mausoleo di Marco Nonio Macrino.* Milano.
Rossi P., Cimino M. G., Le Pera S. 2004. “Salvaguardia e valorizzazione dei Beni Culturali nel territorio del Comune di Roma”, in *Beni Culturali e catalogazione integrata (IV corso di formazione e di aggiornamento per il personale dei Musei Civici, Roma, 13-19 novembre 2003)*. Roma.
Sgambati G. 2009. “Programmi di digitalizzazione, accesso in rete e conservazione del patrimonio culturale”, in *Progetto Centri e-learning*, 2. Napoli.
Terrenato N., Becker J.A. 2009. “Il sito di Monte delle Grotte sulla via Flaminia e lo sviluppo della villa nel suburbio di Roma”, in *Suburbium II*, pp. 393-401. Roma.
Valenti M. 2000. “La piattaforma GIS dello scavo. Filosofia di lavoro e provocazioni, modello dei dati e soluzione GIS”, in *Archeologia e Calcolatori*, 11, pp. 93-109. Roma.
Valentino P.A. 1997. “Criteri e metodi per la scelta delle tecnologie informatiche applicabili ai beni culturali”, in Galluzzi P., Valentino P.A. (a cura di), *I formati della memoria. Beni culturali e nuove tecnologie alle soglie del terzo millennio*, pp. 167-201. Firenze.
Volpe R. 2000. “Il suburbio”, in Giardina A. (a cura di), *Roma Antica*, pp. 183-210. Roma-Bari.
Ward Perkins J.B. 1955. “Notes on Southern Etruria and the Ager Veientanus”, in *Papers of the British School at Rome*, XXIII, pp. 44-58. Roma.
SITAR
http://www.archeositarproject.it/
http://webais.archeositarproject.it/webgis/login.php

Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni

Andrea Bellandi, Istituto di Linguistica Computazionale "A. Zampolli", CNR – Pisa, andrea.bellandi@ilc.cnr.it
Federico Boschetti, Istituto di Linguistica Computazionale "A. Zampolli", CNR – Pisa, federico.boschetti@ilc.cnr.it
Anas Fahad Khan, Istituto di Linguistica Computazionale "A. Zampolli", CNR – Pisa, fahad.khan@ilc.cnr.it
Angelo Mario Del Grosso, Istituto di Linguistica Computazionale "A. Zampolli", CNR – Pisa, angelo.delgrosso@ilc.cnr.it
Monica Monachini Istituto di Linguistica Computazionale "A. Zampolli", CNR – Pisa, monica.monachini@ilc.cnr.it

«Il dividere le sentenzie dalle parole è un dividere l'anima dal corpo»
B. Castiglione *(Il Cortegiano*, 1-33)

Descrizione del Contributo

Il dizionario storico è il luogo d'incontro privilegiato di linguistica e lessicografia e filologia e critica letteraria. Nella prima parte prendiamo in considerazione un caso di studio piuttosto noto, relativo all'espressione "provando e riprovando", per mostrare come perfino i luoghi citati nei dizionari, che sono introdotti con lo scopo di disambiguare i termini in contesto, non siano privi di controversie interpretative. Nella seconda parte, molto più dettagliata e più tecnica, tentiamo di aggiungere ai modelli lessicali e citazionali già esistenti ed aperti soluzioni minime che ci permettano di collegare voci, citazioni e interpretazioni all'interno dell'universo dei Linked Open Data.

Caso di studio: "provando e riprovando"

Ci sembra degno di nota il fatto che proprio i social media, che per l'autorevole (e un po' narcisista)[1] Umberto Eco danno "diritto di parola a legioni di imbecilli"[2], consentano ad un navigatore nascosto dietro all'assai poco autorevole pseudonimo "Bubu7" di smentire tramite un post, con argomenti ineccepibili, un'osservazione dello stesso Umberto Eco pubblicata a stampa.

In un articolo dal titolo *Provare e riprovare* nella rubrica *la bustina di Minerva* del settimanale *l'Espresso* del 27 luglio 2004, Umberto Eco elogia il metodo scientifico e scrive:

1 "Il narcisismo di Eco, che esiste, è un narcisismo rovesciato, come il binocolo: anziché avvicinare e mostrare, allontana il soggetto lo sposta verso l'orizzonte. E lui si è sempre compiaciuto di questo." (Roberto Cotroneo, Fenomenologia di Umberto Eco, 2016, https://robertocotroneo.me/2016/03/21/umberto).

2 In realtà, nel discorso di Umberto Eco disponibile su YouTube (http://bit.ly/2glGIWa), il soggetto è un più restrittivo "il fenomeno Twitter"; la generalizzazione "i social media" e, addirittura, "Internet" è frutto più propriamente della (sovra)interpretazione che ne ha dato immediatamente la stampa.

> la scienza procede correggendo continuamente se stessa, falsificando le sue ipotesi, per 'trial and error' (tentativo ed errore) [...] Che è poi quello che sosteneva secoli fa l'Accademia del Cimento, il cui motto era 'provando e riprovando' - e 'riprovar è non significava provare di nuovo, che sarebbe il meno, ma respingere (nel senso della riprovazione) quello che non poteva essere sostenuto alla luce della ragionevolezza e dell'esperienza.

Come anticipato sopra, il motivo per cui l'interpretazione *difficilior* sostenuta da Eco non regge, è ben argomentato in un post presente sul sito dell'Accademia della Crusca (http://bit.ly/2fw7iKV), che contiene una citazione di Magalotti e una citazione di Dante:

> [...] Ma ecco come il motto [...] è commentato da Lorenzo Magalotti nel Proemio ai lettori dell'unica pubblicazione dell'Accademia, i Saggi di Naturali Esperienze (1667):
> Non vi ha cui meglio rivolgersi che alla fede dell'esperienza, [...] se non di primo lancio come la geometria tanto che provando e riprovando le riesce talora di dar nel segno.
> [….]
> Il significato di riprovare si evince chiaramente dal mio grassetto ed è proprio quello di provare di nuovo, accademicamente negato dal professor Eco.
> Il quale si dev'essere confuso col significato che ha il termine nella seguente terzina della Commedia dantesca (Par. III, 1-3):
> Quel sol che pria d'amor mi scaldò 'l petto,
> di bella verità m'avea scoverto,
> provando e riprovando, il dolce aspetto;

Dove il sole è Beatrice che spiega a Dante come stanno le cose e confuta le sue credenze errate.

Il post è perfettamente in linea con il dizionario Treccani (http://www.treccani.it/vocabolario), che divide la voce Riprovare$_1$ (come "provare di nuovo", nei suoi vari sensi) dalla voce Riprovare$_2$ (come "disapprovare"). Sotto la prima voce, al senso 2, si trova scritto: "per il motto provando e riprovando, v. provare, n. 2", che dice:

> fare un tentativo, tentare [...] Provando e riprovando, motto assunto verso il 1666 dall'Accademia fiorentina del Cimento, tratto dalla Divina Commedia (Par. III, 3, dove però significa «approvando e disapprovando»), che vuole alludere al carattere fondamentale che per gli scienziati di quest'accademia doveva avere la ricerca sperimentale.

Quindi, per gli Accademici del Cimento vale il senso corrente di tentare e ritentare, diverso dal significato dantesco di approvare e confutare. E infatti il passo del Paradiso è nuovamente citato come attestazione del senso Riprovare$_{2\ 1.b}$, glossato con "ant. Dimostrare falso, confutare".

Ciò che apparentemente sorprende è invece la disposizione sia del motto dell'Accademia, sia della citazione dantesca, sotto la sola voce Riprovare$_1$ del Grande Dizionario della Lingua Italiana fondato da Salvatore Battaglia con la glossa:

> Provare nuovamente o ulteriormente; ripetere una prova, un tentativo, un'esperienza, ecc. - Anche assol., con partic. riferimento all'impresa dell'Accademia del Cimento Provando e riprovando.

Ma basta leggere l'articolo di Aldo Tuccione, "Provando e riprovando": *Beatrice e l'accademia del Cimento* (Tuccione 1998), per seguire la storia dell'origine e della propagazione di questo anacronismo esegetico, che fa Dante precursore del metodo sperimentale, nato da padre Antonelli, poi riportato ma non condiviso da Tommaseo e infine attribuito più e più volte nel corso del Novecento addirittura a Tommaseo stesso.

Nei dizionari, le fonti letterarie e documentali hanno lo scopo di indicare in modo evidente l'impiego di un senso (e possibilmente di fornire un *terminus post quem* per la datazione dell'uso). Ma visto che la citazione in sé non garantisce l'univocità dell'interpretazione, l'associazione di un senso ad una voce, oltre al supporto della citazione, dovrebbe avere il sostegno della letteratura secondaria che ne giustifichi l'interpretazione.

Tecnologie, modelli e standards del Semantic web per le risorse linguistiche

Il caso di studio precedente ha mostrato come lessicografia e filologia debbano restare in continuo dialogo. Il nostro scopo è ora di proporre degli accorgimenti tecnici per connettere in modo sempre più efficace il mondo delle risorse digitali lessicografiche, quello delle risorse digitali testuali e quello delle risorse digitali relative alla letteratura secondaria.

Il paradigma dei Linked Open Data (LOD) fornisce un metodo di pubblicazione di dati strutturati, tramite il protocollo standard HTTP, che permette sia di interconnettere tra loro diversi datasets, sia di riusare vocabolari condivisi per descrivere dati e metadati. L'idea di fondo è infatti quella del Semantic Web, in cui si prevede di associare alle risorse informative sul Web una descrizione formale del loro significato, attraverso la sovrapposizione di uno o più livelli di metadati semantici (Ciotti 2012). Tali metadati sono espressi in formalismi che fanno parte della famiglia dei sistemi di rappresentazione della conoscenza sviluppati nell'ambito dell'intelligenza artificiale. Il paradigma dei LOD fonda le proprie basi sul framework RDF (Resource Description Framework), attraverso il quale è possibile costruire i propri datasets mediante la dichiarazione di una serie di statements nella forma di tripla soggetto-predicato-oggetto (Ciotti and Tomasi 2106).

Negli ultimi anni c'è stato un notevole aumento del numero di risorse linguistiche codificate attraverso il paradigma dei LOD. La codifica e la condivisione di tali risorse è quasi sempre basata sul modello lessicale LExicon Model for ONtologies (*lemon*) che ha quindi conseguentemente assunto la connotazione di standard de facto per la rappresentazione di lessici computazionali[3]. É stato infatti utilizzato anche nella modellazione, tra le altre risorse, della Princeton WordNet, di Framenet e di VerbNet. Da un certo punto di vista *lemon* può essere considerato come un modello lessicale per arricchire descrizioni ontologiche con dati linguistici. Esso si basa sul principio che ogni senso di ogni entrata lessicale deve essere collegato ad un concetto ontologico. In (Khan et al 2014) e (Khan et al 2016) è stata proposta un'estensione del modello *lemon*, chiamato *lemonDia* per la rappresentazione computazionale del fenomeno della diacronia nei lessici. Tale esigenza nasce evidentemente dal trattamento di lingue "classiche", come il greco antico, il latino o il sanscrito in cui si ha accesso a un corpus di testi che coprono un lungo periodo e che quindi presentano una evoluzione della semantica del linguaggio. È anche il caso di lingue moderne come il francese, l'italiano o l'inglese dove è possibile contare su un patrimonio di testi che attestano, in periodi differenti, lo sviluppo di tali lingue. *lemonDia* permette di modellare tali aspetti, sia tenendo traccia delle variazioni nel tempo del significato delle parole nel lessico di una lingua, sia categorizzando tali variazioni (per esempio metonimia). Tale rappresentazione permette di utilizzare le tecnologie e gli standard del Semantic web, come ad esempio SPARQL e Ontology Web Language (OWL) per interrogare e fare inferenze su tali dataset lessicali.

Quando si tratta di codificare lessici preesistenti, specialmente nel caso di dizionari cartecei, tramite il paradigma dei LOD, è importante essere in grado di rappresentare tutti o la maggior parte dei diversi tipi di informazione lessicografica contenuta all'interno di ogni voce, utilizzando risorse di vocabolari[4] esistenti già codificati nel paradigma LOD (Daquino and Tomasi 2015). Sia *lemon* che *lemonDia* non prevedono classi o proprietà che lo consentono. Faremo perciò uso di una estensione di *lemon* chiamata *polyLemon* (Khan et al 2015) che permette di rappresentare nel modello la struttura gerarchica dei sensi di ogni voce di un dizionario.

Proposta di un modello formale per il caso di studio

In questo articolo ci concentriamo sulle attestazioni delle forme lessicali, che sono fornite come prova testuale per ogni senso delle voce. Per poter fare ciò, è necessario dotare il modello di classi e proprietà che permettano di collegare il lessico ai corpora di riferimento. Partendo dal contesto sopra

3 Una sua versione riveduta, denominata ONTOLEX, è stata recentemente pubblicata dal W3C.
4 https://lov.okfn.org/dataset/lov/

descritto e dai requisiti definiti precedentemente, in questo lavoro viene presentato un modello che estende *lemonDia*, che permette di rappresentare le attestazioni delle forme lessicali in un corpus e di giustificare, tramite la letteratura secondaria, l'associazione del senso a tali forme.

In figura 1(d) viene mostrata l'estensione proposta. Ogni senso o psenso[5] è associato con un elemento della classe *Attestation* tramite la relazione *hasAttestation*, che rappresenta l'attestazione di uno specifico senso[6].

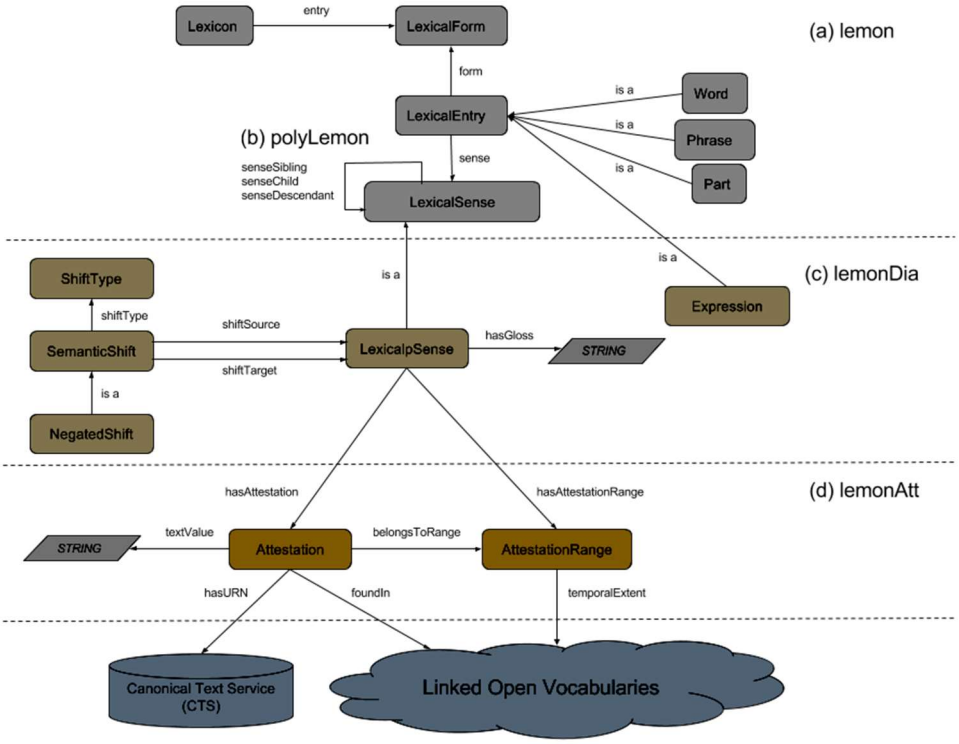


Figura 1. (a) modello *lemon* (b) *polylemon* (c) estensione diacronica (d) estensione citazionale

Tali elementi hanno una proprietà *explanation* che permette di specificare l'eventuale glossa, una proprietà *textValue* che permette di rappresentare, come stringa, il contesto della forma lessicale e una proprietà *hasURN* che permette di specificare, se esiste, l'urn dell'architettura CITE-CTS[7] (Canonical Text Service) che identifica, univocamente e globalmente, il frammento di testo relativo all'attestazione. Inoltre è possibile associare l'elemento attestazione con il record bibliografico dell'opera a cui esso appartiene tramite la proprietà *foundIn*. Infine ogni senso può essere associato con un elemento della classe *AttestationRange* che permette di specificare un intervallo temporale (Khan 2014) (chiuso o aperto) tramite l'utilizzo del vocabolario OWL-Time (Hobbs and Pan 2004).

A titolo di esempio, mostriamo come il caso d'uso riportato all'inizio del contributo può essere rappresentato nel nostro modello. La figura 2 mostra come i relativi datasets, le entrate parziali di

5 Un psenso rappresenta un significato di un'entrata lessicale in un certo intervallo di tempo, come descritto in (Khan 2014).

6 È possibile codificare le attestazioni di un'entrata lessicale alla stregua di quanto viene fatto nell'ambito degli studi di opere frammentarie attraverso le "quotation" e il "text reuse" (Büchler et al 2014). Nello specifico, si rappresentano le attestazioni nel modello come "porzioni di testo citato" collegate alle fonti primarie che veicolano il "frammento" in oggetto attraverso il paradigma dei LOD e un sistema di identificazione globale ed univoco. Si veda (Berti 2011) e (Berti 2012).

7 L'architettura CITE, sviluppata nel progetto Homer Multitext, fornisce, da un lato, una modalità di identificazione standard e valida semanticamente per mezzo di URN, dall'altro lato, un protocollo di recupero dei passaggi testuali all'interno di oggetti citabili chiamato CTS (Smith and Weaver 2009). Gli URN, essendo un tipo di URI, possono essere inclusi all'interno della descrizione RDF della risorsa lessicografica e quindi perfettamente aderenti alle specifiche LOD.

Treccani e di Battaglia, sono rappresentate tramite *lemonDia*. Ambedue i dizionari riportano i due significati della parola "riprovare" come entrate separate (omonimi riprovare$_1$ e riprovare$_2$) aventi lo stesso significato nei due dizionari. La figura 3 riporta l'entrata della voce riprovare$_2$ della Treccani.

Il modello permette di rappresentare la struttura gerarchica dell'entrata di riprovare$_2$ della Treccani tramite la proprietà *senseSibling*, come mostrato in figura 2. Il senso di interesse è 1b. ed è codificato come un psense, S_{21b}, che riporta la glossa "Dimostrare falso, confutare" e un'attestazione dal III Canto del *Paradiso* della Divina Commedia. In particolare tale attestazione è collegata all'URN CTS che identifica la specifica citazione nel testo. L'attestazione è collegata anche a un record che rappresenta il Paradiso (il record proviene dal dataset RDF della Library of Congress[8]), che fornisce anche una data relativa all'anno di composizione che è utilizzata dal modello come data dell'elemento attestazione.

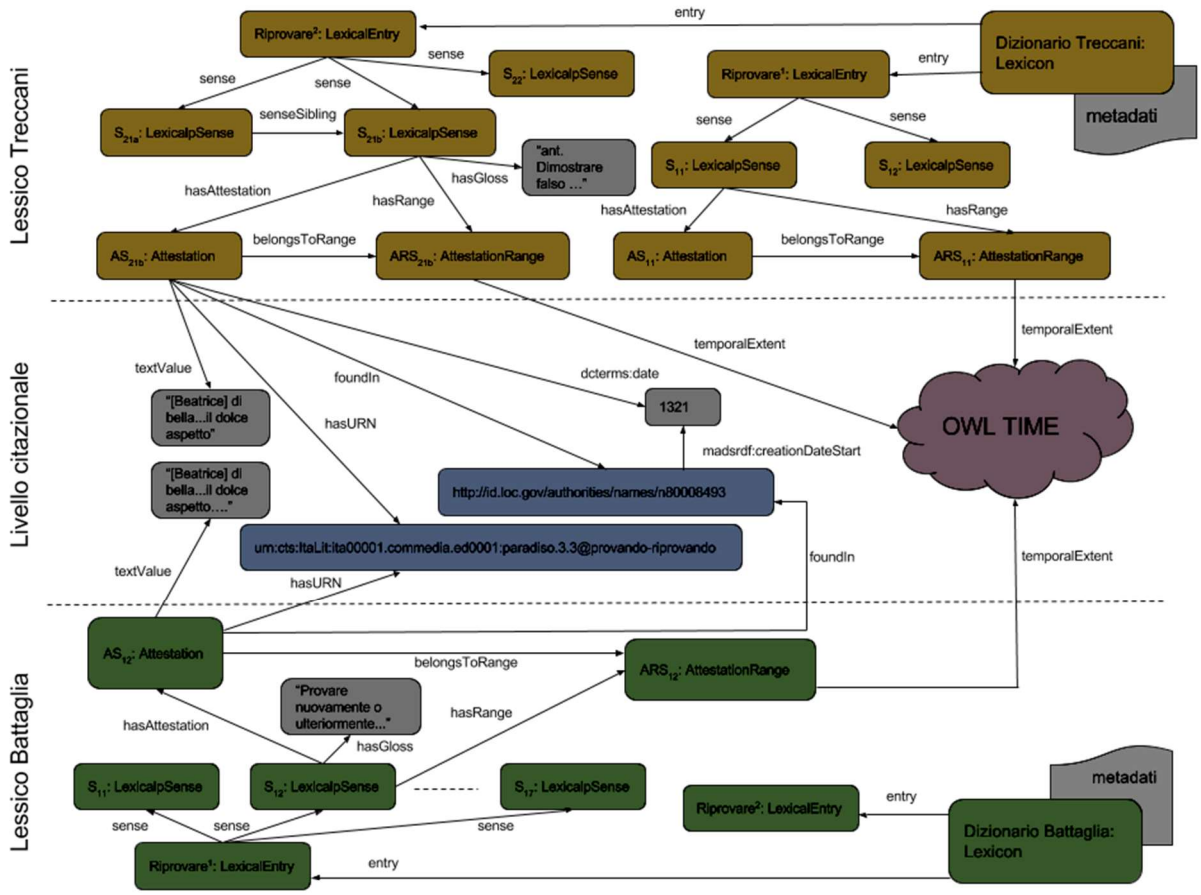


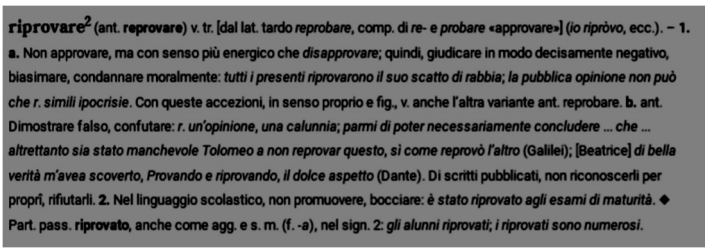
Figura 2. Modello proposto. Rappresentazione delle citazioni: esempio "provare e riprovare".

In questo caso, come in molti altri, c'è la necessità di fare riferimento alla letteratura secondaria per determinare le evidenze a supporto di ogni ipotesi. Anche in questo senso, il paradigma dei LOD fornisce un formato ideale per rappresentare questo tipo di situazione grazie sia alla possibilità di rappresentare formalmente tali fenomeni, sia alla disponibilità di vocabolari già esistenti. La Figura 4 mostra questa caratteristica del modello proposto. In questo caso è possibile rappresentare che l'articolo di Tucciarone[9] in "Quaderni di Storia" rifiuta l'interpretazione dell'attestazione di Dante

8 http://id.loc.gov/authorities/names/

9 Per la descrizione di schede e riferimenti bibliografici è stato utilizzato il vocabolario ontologico denominato Bibliographic Ontology (BIBO) (http://bibliontology.com/specification). BIBO fa uso dei termini Dublin Core per la definizione e la pubblicazione linked data della descrizione di documenti. Una versione più accurata per la descrizione catalografica delle attestazioni farà uso del modello FRBR (Le Boeuf 2005) e di alcuni moduli delle

data dal dizionario di Battaglia, a supporto del significato rappresentato dal psenso S_{12}. Inoltre è possibile specificare che tale articolo supporta invece l'interpretazione della Treccani.
Le relazioni che rappresentano tali fenomeni, rispettivamente *refutes* e *confirms*, appartengono al vocabolario Linked Science Core Vocabulary (LSC) che definisce uno schema grazie al quale è possibile descrivere relazioni temporali, spaziali e altro, tra oggetti scientifici.



riprovare[2] (ant. **reprovare**) v. tr. [dal lat. tardo *reprobare*, comp. di *re-* e *probare* «approvare»] (*io ripròvo*, ecc.). – **1. a.** Non approvare, ma con senso più energico che *disapprovare*; quindi, giudicare in modo decisamente negativo, biasimare, condannare moralmente: *tutti i presenti riprovarono il suo scatto di rabbia; la pubblica opinione non può che r. simili ipocrisie*. Con queste accezioni, in senso proprio e fig., v. anche l'altra variante ant. reprobare. **b.** ant. Dimostrare falso, confutare: *r. un'opinione, una calunnia; parmi di poter necessariamente concludere ... che ... altrettanto sia stato manchevole Tolomeo a non reprovar questo, sì come reprovò l'altro* (Galilei); [Beatrice] *di bella verità m'avea scoverto, Provando e riprovando, il dolce aspetto* (Dante). Di scritti pubblicati, non riconoscerli per propri, rifiutarli. **2.** Nel linguaggio scolastico, non promuovere, bocciare: *è stato riprovato agli esami di maturità*. ◆ Part. pass. **riprovato**, anche come agg. e s. m. (f. -a), nel sign. 2: *gli alunni riprovati; i riprovati sono numerosi*.

Figura 3. Entrata della voce riprovare$_2$ di Treccani.

In base al vocabolario LSC, le opportune classi del modello sono state tipizzate in accordo con i campi di esistenza di *refutes* e *confirms*: l'oggetto dell'interpretazione è diventato l'ipotesi (classe lsc:*Hypothesis*) e i sensi che lo confutano o lo supportano diventano rappresentazioni di oggetti scientifici (classe lsc:*Research*), in quanto assumono una connotazione più di ricerca lessicografica che di una rappresentazione dell'uso di una parola.

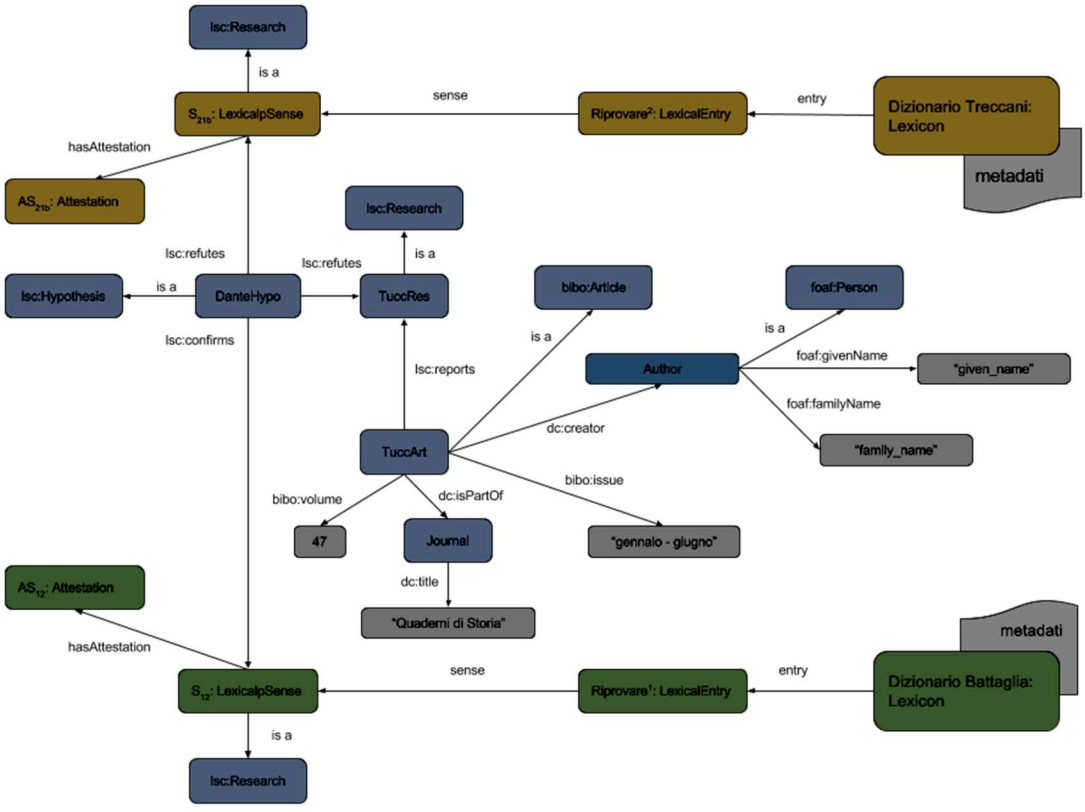


Figura 4. Rappresentazione della letteratura secondaria: "provare e riprovare".

SPAR ontologies (Peroni 2014).

Conclusioni

La lessicografia italiana, che vanta una storia molto gloriosa, sta raggiungendo la maturità anche nel mondo digitale, per qualità e numero di opere disponibili, dalle diverse edizioni del Vocabolario della Crusca (http://www.lessicografia.it), al Tommaseo-Bellini (http://www.tommaseobellini.it), al Vocabolario Treccani (http://www.treccani.it/vocabolario). Abbiamo cercato di dimostrare l'importanza di collegare le risorse lessicografiche non solo con le risorse testuali citate, ma anche con le risorse bibliografiche (possibilmente in *full text*) che criticano e interpretano tali risorse. Se un vocabolario senza citazioni è un corpo senz'anima, una citazione senza interpretazioni è un'anima senza spirito.

Riferimenti Bibliografici

Berti, Monica. 2011. "Citazioni E Dinamiche Testuali. L'intertestualità E La Storiografia Greca Frammentaria." In *Tradizione E Trasmissione Degli Storici Greci Frammentari II. Atti Del Terzo Workshop Internazionale*, 439–58. Roma.

Berti, Monica. 2012. "Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres." *Ancient Society* 43: 269–88.

Büchler, Marco, Philip R Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. "Towards a Historical Text Re-Use Detection." In *Text Mining*, 221–38. Springer International Publishing.

Ciotti, Fabio. 2012. "Web Semantico, Linked Data E Studi Letterari: Verso Una Nuova Convergenza." *Quaderni DigiLab* 2 (1): 243–76.

Ciotti, Fabio, and Francesca Tomasi. 2016. "Formal Ontologies, Linked Data, and TEI Semantics." *Journal of the Text Encoding Initiative*, no. 9: 1–23. doi:10.4000/jtei.1480.

Daquino, Marilena, and Francesca Tomasi. 2014. "Ontological Approaches to Information Description and Extraction in the Cultural Heritage Domain." In *AIUCD2014 Proceedings*, 8. Bologna: ACM.

Hobbs, Jerry R., and Feng Pan. 2004. "An Ontology of Time for the Semantic Web." *ACM Transactions on Asian Language Processing*, Issue on Temporal Information Processing, 3 (1): 66–85.

Khan, Fahad, Andrea Bellandi, and Monica Monachini. 2016. "Tools and Instruments for Building and Querying Diachronic Computational Lexica." In *LT4DH2016 Proceedings*, 164–71. Osaka.

Khan, Fahad, Federico Boschetti, and Francesca Frontini. 2014. "Using Lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources." In *LDL2014 Proceedings*, 50–54. Reykjavik.

Khan, Fahad, Javier E. Díaz-Vera, and Monica Monachini. 2016. "Representing Polysemy and Diachronic Lexico-Semantic Data on the Semantic Web." In *SWASH2016 Proceedings*, 37–45. Heraklion, Greece.

Le Boeuf, Patrick. 2005. *Functional Requirements for Bibliographic Records (FRBR): Hype or Cure-All?* New York: The Haworth Information Press.

Peroni, Silvio. 2014. "The Semantic Publishing and Referencing Ontologies." *Semantic Web Technologies and Legal Scholarly Publishing*, 121–93.

Smith, D Neel, and Gabriel Weaver. 2009. "Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture." *Text Mining Services*, 129–39.

Tuccione, Aldo. 1998. "Provando E Riprovando: Beatrice E L'accademia Del Cimento." *Quaderni Di Storia* 47: 103–12.

Linked Open Data per l'analisi dei dati e lo sviluppo della ricerca sulle vittime della Shoah in Italia

Laura Brazzo, Fondazione CDEC, laurabrazzo@cdec.it
Silvia Mazzini, Regesta.exe, smazzini@regesta.com

Introduzione

Il progetto della Fondazione Centro di Documentazione Ebraica Contemporanea per la pubblicazione in Linked Open Data dei dati sulle vittime della Shoah in Italia rappresenta per molti aspetti un momento di svolta per la ricerca su questo tema.

Tale pubblicazione, eseguita secondo i paradigmi dei Linked Open Data e sfruttando le tecnologie standard proposte dal W3C per il Semantic Web, consente oggi di svolgere analisi sui dati che prima erano riservate ai soli autori/produttori della banca dati.

A complemento di ciò vanno sottolineati i principali benefici derivanti da questo tipo di pubblicazione: le attività di reasoning e interlinking – naturale sviluppo della pubblicazione LOD – per l'arricchimento dell'informazione relativa ai dati esposti - altrimenti demandata a lunghe e onerose ricerche manuali; la riconciliazione dei dati, per un'efficace attività di data-cleaning sui dati stessi.

Sviluppo e motivazioni del progetto LOD sulle Vittime della Shoah in Italia

Le informazioni sulle vittime della Shoah, sono state pubblicate nel 1991 nel volume "Il Libro della memoria" e nel 2011 sul sito www.nomidellashoah.it.

In entrambi i casi le informazioni rese pubbliche hanno avuto come principale scopo la commemorazione delle vittime; diversamente l'accesso alla banca dati, con la completezza dell'informazione relativa alle oltre 8000 vittime, è sempre rimasta riservata al produttore della banca dati stessa, ovvero la Fondazione CDEC.

L'apertura delle informazioni in un formato strutturato, grazie anche all'elaborazione di una ontologia di dominio (Shoah Ontology) atta a descrivere in maniera formale il processo di persecuzione e deportazione (http://dati.cdec.it/lod/shoah/reference-document.html) degli ebrei - dall'arresto fino al destino finale - permette oggi a chiunque di elaborare in proprio l'analisi delle informazioni sulla deportazione degli ebrei dall'Italia – analisi prima riservata al solo produttore della banca dati, a meno di una trascrizione manuale delle informazioni tratte dalle due fonti citate sopra.

Questo è quanto avvenuto, per esempio, per il progetto sugli arresti degli ebrei in Italia sviluppato nel 2012 da Alberto Giordano e Simone Gigliotti nell'ambito dello Spatial History Project promosso dalla Standford University (http://web.stanford.edu/group/spatialhistory/cgi-bin/site/viz.php?id=383&project_id=0). In questo caso le informazioni tratte dal "Libro della Memoria" sono state trascritte e strutturate al fine di creare una mappa dinamica degli arresti di ebrei avvenuti in Italia dopo l'8 settembre 1943 fino al marzo 1945 e fornendo statistiche relative a genere, età, percentuali di decesso. Di questo meritorio lavoro, va tuttavia sottolineato un limite, ovvero la sua non-ripetibilità, ossia non è possibile riprodurre quella mappa se non rifacendo ex-novo la trascrizione delle informazioni provenienti dalla fonte originaria (Il Libro della Memoria).

Ciò non toglie affatto autorevolezza al lavoro svolto, ma lo fa ricadere in quella sfera di attività in cui l'informazione (in questo caso, l'applicazione realizzata) non è disponibile ad altri che al suo autore. Di fatto, ci troviamo di fronte, di nuovo, ad una tradizionale interpretazione dei dati, con la sola differenza della forma in cui l'informazione viene restituita e che l'utente può solo recepire come risultato ultimo di un processo di analisi che gli rimane precluso.

Query SPARQL, interlinking, visualizzazione dei dati

La pubblicazione in formato LOD dei dati utilizzati per il progetto Spatial History consente ora a chiunque sia in grado di interrogare un end-point SPARQL, di ripetere l'analisi svolta da Giordano e Gigliotti. La granularità dell'informazione fornita permette inoltre l'allineamento dei dati estratti con altri dataset provenienti da altri data provider. Un caso, semplice, in questo senso potrebbe essere, per esempio, l'incrocio delle informazioni relative ai sopravvissuti e la loro produzione letteraria nel dopoguerra utilizzando i dataset di Worldcat o di Viaf.org, tale da permettere, per esempio, l'analisi dei picchi di produzione sulla base degli anni di pubblicazione, o sul genere.

Più complesso al momento risulta l'interlinking con analoghi dataset relativi alle vittime della Shoah poichè tali dati non sono ancora disponibili in formato RDF. Considerata la tipologia, i diversi criteri storiografici utilizzati dai vari enti ed istituti di ricerca nella realizzazione delle banche dati dedicate alle vittime della Shoah, l'allineamento dei dati sulle vittime consentirebbe uno sviluppo importante degli studi nella direzione della "individuazione" delle vittime, anziché della semplice "nominazione" come è attualmente. A dimostrazione delle opportunità di ampliamento della conoscenza che si aprirebbero in questo campo si intende presentare per l'occasione un test di allineamento ad hoc, fra due subset di dati provenienti della Fondazione CDEC e dal Memoriale della Shoah di Parigi.

La presentazione includerà anche l'esposizione dei risultati dell'allineamento dei primi due subset di dati rilasciati in formato LOD dalla Fondazione CDEC, quello relativo alle vittime della Shoah e quello nuovo relativo ai partigiani ebrei operanti in Italia fra il 1943 e il 1945. Di questo primo test-bed verranno messe in luce le possibilità di interconnessione come anche le inconsistenze emerse. Gli esempi di query SPARQL contestualmente forniti serviranno ad illustrare le molteplici possibilità di indagine (e di visualizzazione grafica dei dati) che la modalità di pubblicazione LOD offre oggi alla ricerca.

Riferimenti Bibliografici

Anderson, S. & Blanke, T. (2012). Taking the long view: from e-Science humanities to humanities digital ecosystems. *Historical Social Research*. 37 (3), 147–164. [online]. Available from: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-378350 (Accessed 24 August 2015).

Beorn, W. et al. (2009). Geographies of the Holocaust. *Geographical Review*. 99 (4), 563–574. [online]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1931-0846.2009.tb00447.x/abstract (Accessed 19 October 2015).

Blanke, T. & Kristel, C. (2013). Integrating Holocaust Research. *International Journal of Humanities and Arts Computing*. [Online] 7 (1-2), 41–57. [online]. Available from: http://www.euppublishing.com/doi/abs/10.3366/ijhac.2013.0080 (Accessed 18 October 2015).

Cohen, D. J. (2004) Digital history: the raw and the cooked. *Rethinking History*. [Online] 8 (2), 337–340. [online]. Available from: http://www.tandfonline.com/doi/abs/10.1080/13642520410001683996

Lefkovits, E. (2004). Eternalizing the names of Holocaust victims. The Jerusalem Post. 23 November. Retrieved from: http://www.lexisnexis.com.pros.lib.unimi.it/lnacui2api/api/version1/getDocCui?lni=53Y6-YJH1-F12G-D4J7&csi=10911&hl=t&hv=t&hnsd=f&hns=t&hgn=t&oc=00240&perma=true (Accessed 11 October 2015).

Faro, L. M. C. (2015). The Digital Monument to the Jewish Community in the Netherlands: a meaningful, ritual place for commemoration. *New Review of Hypermedia and Multimedia*. [Online] 21 (1-2), 165–184. [online]. Available from: http://www.tandfonline.com/doi/full/10.1080/13614568.2014.983556

Klarsfeld, S. (1978), Le mémorial de la déportation des juifs de France. Paris.

Knowles A. K., Cole T., Giordano A. (2014). Geographies of the Holocaust, Bloomington: Indiana University Press
Picciotto Fargion, L. (1991) *Il libro della memoria: gli ebrei deportati dall'Italia (1943-1945)*. Milano: Mursia.
Speck, R., Blanke, T., Kristel, C. et alii (2014) "The Past and the Future of Holocaust Research: From Disparate Sources to an Integrated European Holocaust Research Infrastructure".
Retrieved from http://arxiv.org/ftp/arxiv/papers/1405/1405.2407.pdf
Coyle, K. (2013) "Library linked data: an evolution".
Retrieved from http://leo.cineca.it/index.php/jlis/article/view/5443/7889
Edelstein, J. et alii (2013). "Linked Open Data for Cultural Heritage: Evolution of an Information Technology".
Retrieved from http://academiccommons.columbia.edu/catalog/ac%3A168445
Summers, E., Salo, D. (2013). "Linking Things on the Web: A Pragmatic Examination of Linked Data for Libraries, Museums and Archives." Library of Congress.
Retrieved from http://arxiv.org/ftp/arxiv/papers/1302/1302.4591.pdf
Halpine, H. et alii (2009). "When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data"
Glaser, H., Halpine, H. (2012). "The Linked Data Strategy for Global Identity".
Retrieved from http://eprints.soton.ac.uk/333924/3.hasCoversheetVersion/IC-16-02-Lnkd.pdf

SigNet – a network of Hellenistic sealings & archives Digital tools and methodologies for big data mining, cross-media quantitative and qualitative analysis, museum engagement and citizen science

Stefano G. Caneva, Marie Curie Research Fellow of the University of Padua, Italy,
stefano.caneva@unipd.it
Branko F. van Oppen de Ruiter, Allard Pierson Museum, Amsterdam, The Netherlands,
b.f.vanoppen@uva.nl

Introducing SigNet

SigNet is a transatlantic consortium for the digital study and dissemination of innovative research of miniature objects such as sealings and coins from the Hellenistic Eastern Mediterranean[1]. The wide geographical provenance of the evidence (the Aegean, the Near East and Egypt) allows for a broad-scale study of the administrative and cultural history of the regions conquered by Alexander the Great and held under Greco-Macedonian political control from *ca*. 330 to 30 BCE (fig. 1). The project aims to tackle methodological issues at the level of both research and dissemination by developing and testing digital tools for making data mining faster and more efficient, so as to enhance quantitative and qualitative analysis, and to promote citizen's engagement and develop interactive museum experiences. The following research areas and issues are involved:

- iconographic and stylistic analysis of seals and sealings;
- agents and processes involved in production and distribution of seals and sealings;
- geographical distribution of iconographic and technical features, within and across regional workshops;
- archival practice: administration and business, public and private;
- definition of royal portraiture and propaganda, its circulation and reception;
- religious and cultural environment of seals' owners and users.

Close collaboration with the American Numismatic Society will ensure that the study of seals and sealings is carried out with a sound cross-media approach and that the digital development of the SigNet tools and platform benefits from the expertise of the ANS in the field of Linked Open Data (LOD) for Digital Humanities.

The Nature of the Evidence

The hoards of sealings included in the SigNet project were found in the Greek island of Delos[2], the temple of Horus at Edfu as well as Elephantine in southern Egypt[3], Tel Kedesh in Israel near Tyre[4], and Seleucia-on-the-Tigris in Mesopotamia[5]. Following Alexander the Great's campaigns from Macedon to India, a substantial change in administrative practices occurred across the conquered

1 The project is currently under evaluation for a Digging into Data call for proposals: http://diggingintodata.org. For an introductory overview of the project goals and activities, see: Caneva – van Oppen 2016.
2 Boussac 1992; Brun 2010.
3 Plantzos 2011.
4 Ariel – Naveh 2003; Herbert – Berlin 2003; Herbert 2003-2004.
5 Bollati – Messina – Mollo 2004; Messina 2005.

regions. While Hellenistic states were characterized by a plurality of cultures and political systems, one common feature they shared was a commitment to record keeping and complex bureaucratic practices. States, including kings, confederacies and city-states, produced hundreds of thousands of documents written mostly on papyri in Greek, whereas others were written on parchment or cuneiform tablets and in various local languages. These documents were sealed with impressions from signet rings or stamps, and stored in various types of archives. Such archives considerably varied in function and size, ranging from the large official archive at Seleucia-on-the-Tigris, Babylonia (25,000+ sealings) to the small private archive in Elephantine, Egypt (35 sealings) (*cf.* fig. 2).

While most of these records are lost, the many thousands of preserved sealings shed light on the size of these archives and on the administrative and archival procedures in use across the Hellenistic world[6]. Moreover, seals and sealings transmit a distinctive iconography, which was chosen and disseminated by a variety of agents, from private individuals to the ruling dynasts and their officers, passing through cities and other regional institutions. As a consequence, these miniature objects provide priceless insights into the choices individual and institutional agents made in presenting their (often mixed) identity and status, as well as into the royal manipulation of local and imperial iconography[7]. More broadly speaking, they contribute to our understanding of the circulation of cultural and religious imagery and of their possible cross-regional interactions within the Hellenistic world[8]. When studied systematically rather than in isolation, the combined corpora of Hellenistic sealings allow scholars to explore the nature of archives and their content, as well as to compare archival practices across distant regions and different strata of the various societies.

SigNet Teams

The SigNet consortium consists of research teams, GLAM institutions and individual scholars with multi-disciplinary expertise. The main research groups are based in three countries, namely France, the Netherlands and the United States of America[9]. The project partners collectively own a large data-set of more than 100,000 objects, between seal impressions and coins, part of which has already been associated with high-definition 2D and 3D images and metadata. All-encompassing studies of the material have thus far been hampered by the fact that the separate repositories are far afield and do not correspond or communicate easily with each other. As the long experience of Europeana has now shown, bringing together digital assets from different cultural heritage institutions implies dealing with the standardization of metadata models and licensing policies for digital data on the internet[10]. The collaborative framework of the SigNet consortium provides an added value for the selection and development of common vocabularies, semantics and ontologies. While each partner institution is responsible for storing and managing its datasets, of which it remains the sole owner, the participatory definition of shared standards for ontology and metadata management will enable the alignment between the various managing systems in use and ensure interoperability between datasets within SigNet as well as with other major Digital Humanities initiatives.

Specific technical tasks are held by the French sub-teams, the American Numismatic Society and the Dutch team, the latter being supervised by the Allard Pierson Museum. The French sub-teams *ArScAn* and *HiSoMa* will supervise the process of definition of the shared SigNet ontology. As a

6 Boussac – Invernizzi 1996; Brosius 2003; Coqueugniot 2013; Faraguna 2013.

7 Kyrieleis 2015.

8 Plantzos 1999; Herbert 2008; Iossif 2015.

9 The French team is led by Dr. Marie-Françoise Boussac and is subdivided between *ArScAn* (UMR 7041) in Paris-Nanterre and *HiSoMa* (UMR5189) in Lyon. Dr. Boussac is assisted at *HiSoMa* by Dr. Annette Peignard and at Centro Scavi di Torino by Dr. Vito Messina. The American team is led by Dr. Sharon Herbert at University of Michigan in Ann Arbor, and is assisted by Dr. Peter van Alfen at the American Numismatic Association in New York. The Dutch team, based at the Allard Pierson Museum, is led by Dr. Wim Hupperetz and managed by Dr. Branko van Oppen.

10 See on this point the Europeana Data Model Documentation (http://pro.europeana.eu/page/edm-documentation) and the updated discussion by Charles – Isaac 2015.

leading institution in the online management of digital numismatic assets, the ANS will provide advice on the linking of the different datasets and the development of individual websites; it will also monitor their interconnectivity and implement the SigNet web tools and applications. Moreover, image-recognition software is being developed by the ANS to allow cross-media stylistic and technical queries in the gathered repositories as well as fine-grained stylometry analyses of workshops and artistic hands to which sets of objects can be ascribed. Statistical analysis, which is coordinated by the Dutch team, will considerably benefit from the development of the above-mentioned SigNet digital tools.

SigNet Development

The digital development of the SigNet project can be divided into three sections: (1) ontology engineering; (2) image-recognition software development; and (3) web-portal construction.

The first task is related to the definition of a common ontology of concepts, categories, themes and representations applying to all the diverse datasets of the hoards of sealings and to the ANS collection of coins[11]. Supervised jointly by the teams based at *ArScAn* and *HiSoMa* in tandem with the other teams, this crucial task will ensure that data mining, metadata management and cross-repository referencing will be effectively machine-readable. Drawing on an open access research model, SigNet aims to build upon successful models and to adapt them to the needs of sigillography and numismatic research. All software and digital tools developed within the course of the project will be managed following the principles of the open source web development and henceforth made accessible for other developers to contribute to through a gitHub repository.

The existing metadata describing the seals and sealings, extracted from the individual databases, will be reused and serialized in an XML format, conforming to the NUDS schema, which was developed by the ANS for the Numishare platform. NUDS is an ontology engine for numismatics, following the example offered by common metadata standards such as Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), and Encoded Archival Context-Corporate, Personal, Family Names (EAC-CPF). Numishare is an open-source collection of scripts and style sheets (mainly XSLT, Javascript, CSS, and XForms)[12]. It binds together four other open-source web applications into a cohesive architecture: eXist XML database; Apache Cocoon (serialization of the NUDS/XML records into HTML, RDF, KML, and other web standards), Apache Solr for faceted search and browse and Orbeon XForms for editing XML and publishing data into the Solr search index.

For better interoperability, the image data will be documented with metadata following the Nakala Data Model, already used by the French team via the IT facility infrastructure of the *Très Grande Infrastructure de Recherche (TGIR) Huma-Num*.[13] This model was designed to reuse existing web data models such as: Dublin Core Terms, FOAF, SKOS, ORE and RDFS. The inscriptions on the objects will be transcribed according to the TEI/EpiDoc schema.[14] This policy will also enhance future collaboration with the initiatives of the Stoa Consortium and the newly-founded International Association for Digital Epigraphy (IDEA).

The second section of the digital development concerns the creation of a new image-recognition tool which will enable cross-repository data mining at a very fine-grained level of detail. This state-of-the-art computer technology will establish iconographic and stylometric links between sealings as well as between coins and sealings across the various linked repositories. The efficacy of the image-recognition tool is ensured by the fact that the underpinning digitization of the material is performed through high definition 2D and 3D scanning techniques, which are able to reveal stylistic and artistic

11 *Cf.* Szabados 2014.
12 http://numishare.blogspot.fr.
13 http://www.huma-num.fr.
14 http://www.stoa.org.

details difficult if not impossible to observe by the human eye (*cf.* figs. 3-4). Together, HD scanning and image-recognition software allow researchers to match similar types, recognize fragments from complete images, and even associate artifacts produced at the same workshop or by the same hand. Such computer-aided queries will considerably speed up cross-referential analysis, as the process of establishing links between tens if not hundreds of thousands of items is painstakingly time-consuming – if not impossible – when done by individual scholars or even groups of experts.

The third task of digital development consists in the creation of an open-access platform providing a unique access to the various digital collections and tools. Free registration will provide users, both academic and non-academic, with an individual dashboard that enables simple access to the set of searching, storing and reuse tools developed by the SigNet Project. In this respect, the Pelagios Commons infrastructure provides a useful example of an effective infrastructure offering large combined tools, database and resources with maximum web community exchanges and visibility.[14] The SigNet portal will be built on top of a SPARQL endpoint that aggregates the datasets and will be hosted by an institutional infrastructure (*e.g.*, *Huma-Num*) or a private infrastructure to be determined during the development of the project. This over-arching SigNet website will make sharing and studying the source material (both sealings and coins) more efficient, by enhancing discoverability and opening up the digital assets not only to scientific research, but also to educational and museum organizations and individual amateurs.

The SigNet platform will follow the LOD principles adopted for the Numishare platform, through which the ANS already offers a proven and in-place ecosystem for sharing data in the numismatic field. The overall general data policy of the project is to have the datasets as widely and openly available as possible. Thus the open-access publication model of Numishare, which implements a CC-BY-NC 4.0 International (Attribution-NonCommercial 4.0 International) licensing policy for the pictures of coins, provides a model which will be followed as closely as possible, in dialogue with all the stakeholders.

SigNet Outreach

Current international trends in research policy identify community engagement as a strategic priority for research in humanities and the cultural heritage. In line with this policy, the SigNet consortium will also provide a variety of initiatives meant to reach out to the broader public of non-academic amateurs, cultural organizations, educational and GLAM institutions. This strategy is pursued at three interconnected levels: (1) the SigNet platform; (2) online initiatives in collaboration with non-academic educational institutions; and (3) museum presentations.

Free registration to the SigNet platform will provide users with an individual dashboard to store the results of their queries, collect statistic data and ask for permission to add and enrich metadata, select favorite items, combine them in a virtual exhibition and provide narratives for them. In case of small private collections of related material (such as signet rings, coins or gemstones), owners will have the chance to propose new material and to negotiate with the consortium the terms for the upload of the relevant digital assets. Registered users allowed to provide metadata and featured narratives include scholars, amateurs and educational groups (from schools, museums and universities). Future collaboration with Europeana partner projects could provide a strategic and mutually enriching solution to combine efforts and reduce time and human costs for the development of such user services.

Educational initiatives are meant to engage citizens in the augmentation and dissemination of knowledge about Hellenistic sealings and coins via crowdsourcing and citizen science projects. Schools and groups of non-specialists will be guided by volunteers and professionals from academic, educational and cultural heritage institutions in the process of metadata enrichment and in the construction of narratives for selected sets of evidence. Collaboration with Wikimedia is envisaged

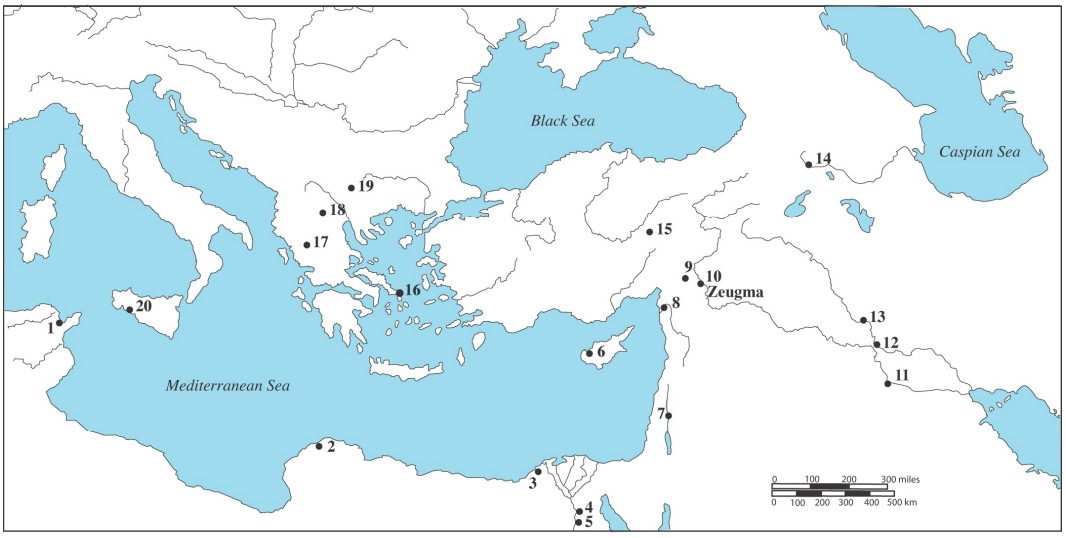
14 *Cf.* commons.pelagios.org

as a strategic priority to promote web-community outreach initiatives, especially through edit-a-thons, uploads of selected material on Wikimedia Commons and metadata enriching on Wikidata[15]. Employing a Wikipedian-in-residence is an interesting solution that will be examined during the development of the project and which could serve to select, test and spread good practices for museums.

Enhancing museum experience, finally, is another major goal which the SigNet consortium aims to achieve through the application of its developed digital tools. Miniature iconographical objects usually suffer from lack of engaging narratives explaining to visitors the technical and stylometric aspects of the creation of these objects as well as their iconographical significance in their historical context of circulation. In response to this issue, the SigNet partners want to let small objects tell big stories. The Digital Museum Lab of the Allard Pierson Museum will play a central role in this respect as its institutional function is to test and implement new approaches for public engagement via digital media. Digital hotspots can make the portal and image-recognition software available inside museums, thus combining the physical exhibition of artifacts with the possibility of rotating and zooming in on high-resolution 3D images of the exhibited objects. The data mining criteria of the SigNet software may allow visitors to enjoy a serendipitous virtual exhibition of the museum's collection, including non-showcased items, or potentially of the whole database of over 100,000 objects. Museum visitors may be allowed to follow and evaluate pre-selected virtual exhibitions enriched with narratives as well as to create and save their own virtual exhibition, thus contributing to the updating and improvement of the museum experience.

In the most ambitious vision shared by the consortium, therefore, the SigNet platform will contribute to the experimentation of new ways of conceiving and co-curating participated exhibitions whereby the selection of exhibits and the creation of thematic narratives can be shared by academic and museum experts, educational organizations and individual volunteers, in line with a methodology aiming to promote citizen science and public engagement initiatives in the field of (Digital) Humanities and the cultural heritage.

Figures



1 Carthage
2 Cyrene
3 Alexandria
4 Edfu
5 Elephantini
6 Nea Paphos
7 Kedesh
8 Antioch
9 Doliche
10 Zeugma
11 Uruk
12 Baghdad
13 Seleucia-on-the-Tigris
14 Artaxata
15 Caesarea of Cappadocia
16 Delos
17 Kallion
18 Pella
19 Titana
20 Selinus

fig. 1 – Map of hoards of Hellenistic sealings
(taken from Herbert 2003-2004, fig. 1)

15 https://www.wikidata.org.

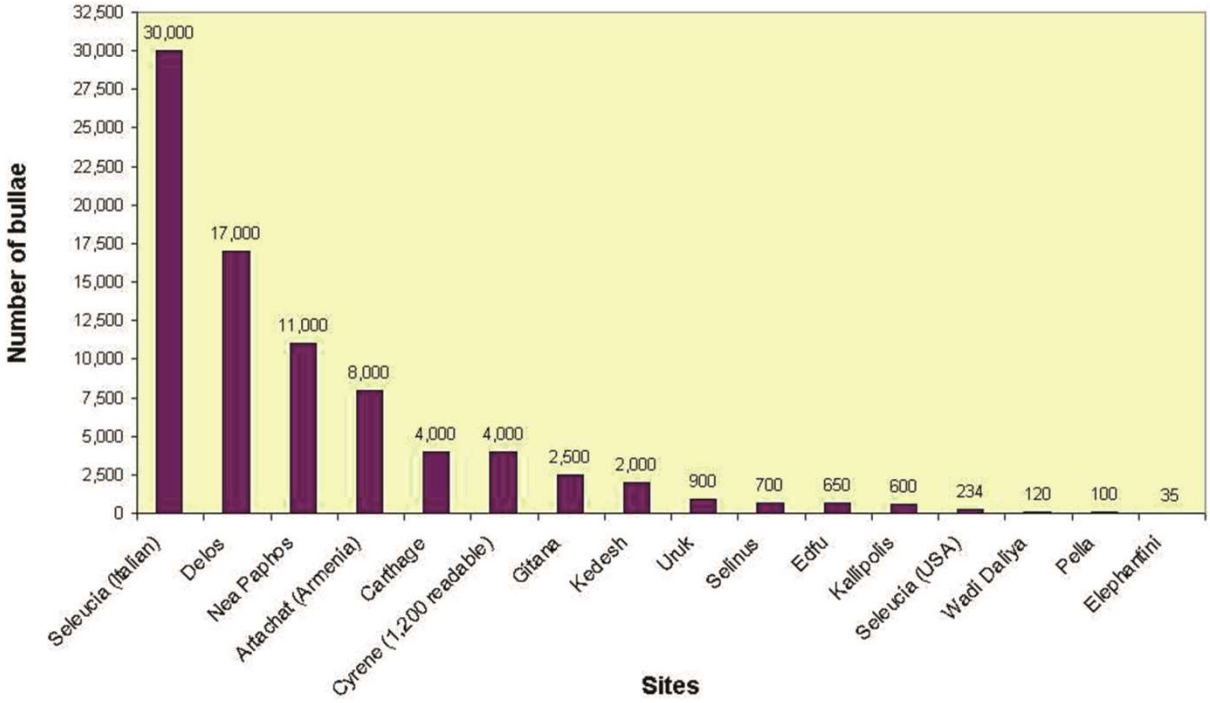


fig. 2 – Number of Hellenistic sealings per site
(taken from Herbert 2003-2004, fig. 2)

fig. 3 – Examples of 3D scans (performed by Moobles)
depicting (from left to right): Apollo, Ptolemy VIII and Cleopatra VII
(APM inv. nos. 8177.179, 278 + 056)

fig. 4 – Portrait comparison of sealing and coin of Cleopatra I
(APM inv. no. 8177-017 + British Museum inv. no. 1978,1021.1)

fig. 5– Portrait comparison of sealing and signet ring
perhaps depicting Ptolemy VIII Euergetes or Ptolemy IX Soter
(APM inv. no. 8177-025 + Fondation Gandur pour l'Art, Geneva, inv. no. GR-030;
courtesy of Dr. Robert S. Bianchi)

References

Ariel, Donald T., and Joseph Naveh. "Selected Inscribed Sealings from Kedesh in the Upper Galilee." *Bulletin of the American Schools of Oriental Research* 329 (2003): 61-80.

Bollati, Ariela, Vito Messina and Paolo Mollo. *Seleucia al Tigri: Le impronte di sigillo dagli archivi*, 3 vols., ed. by Antonio Invernizzi. Alessandria, Italy: Edizioni dell'Orso, 2004.

Boussac, Marie-Françoise, and Antonio Invernizzi (eds.). *Archives et Sceaux du Monde Hellénistique*, Bulletin de Correspondance Hellénique suppl. 29. Athens: École Française d'Athènes, 1996.

Boussac, Marie-Françoise. *Les Sceaux de Délos* I: *Sceaux Publics: Apollon, Hélios, Artémis, Hécate*. Paris: De Boccard, 1992.

Brosius, Maria (ed.). *Ancient Archives and Archival Traditions: Concepts of Record-Keeping in the Ancient World*, Oxford Studies in Ancient Documents 10. Oxford: Oxford University Press, 2003.

Brun, Hélène. "Les sceaux du Sarapieion C de Délos." *Bulletin de Correspondance Hellénique* 134 (2010): 195-221.

Caneva, Stefano G., and Branko F. van Oppen de Ruiter. "SigNet: A Digital Platform for Hellenistic Sealings and Archives." http://link.springer.com/chapter/10.1007/978-3-319-48974-2_25 In: *Digital Heritage – Progress in Cultural Heritage: Documentation, Preservation, and Protection*, Lecture Notes in Computer Science 10059, ed. by Marinos Ioannides http://link.springer.com/search?facet-creator=%22Marinos+Ioannides%22, pt. II: 222-231. Cham, Switzerland: Springer, 2016.
Charles, Valentine, and Antoine Isaac. "Enhancing the Europeana Data Model (EDM)." Europeana Foundation, 30 May 2015, http://pro.europeana.eu/files/Europeana_Professional/Publications/EDM_WhitePaper_17062015.pdf
Coqueugniot, Gaëlle. *Archives et bibliothèques du monde grec: Édifices et organisation, V^e siècle avant notre ère-I^er siècle de notre ère*. Oxford: Oxford University Press, 2013.
Faraguna, Michele (ed.). *Legal Documents in Ancient Societies* IV: *Archives and Archival Documents in Ancient Societies*. Trieste: Edizioni Università di Trieste, 2013.
Herbert, Sharon C. "The Hellenistic Archives from Tel Kadesh (Israel) and Seleucia-on-the-Tigris (Iraq)." *Bulletin of the University of Michigan, Museums of Art and Archaeology* 15 (2003-2004): 65-86.
_____. "The Missing Pieces: Miniature Reflections of the Hellenistic Artistic Landscape in the East." In *The Sculptural Environment of the Roman Near East: Reflections on Culture, Ideology, and Power*, Interdisciplinary Studies in Ancient Culture and Religion 9, ed. by Yaron Eliav *et al.*, 257-272. Louvain: Peeters, 2008.
_____, and Andrea Berlin. "A New Administrative Center for Persian and Hellenistic Galilee: Preliminary Report of the University of Michigan/University of Minnesota Excavation at Kedesh." *Bulletin of the American Schools of Oriental Research* 329 (2003): 13-59.
Iossif, Panagiotis P. "Seleucia on the Tigris under the Seleucids: 'Monetary' Pantheon vs. 'Glyptic' Pantheon". In *Des hommes aux dieux: processus d'héroïsation et de divinisation dans la Méditerranée hellénistique*, ed. by Stefano G. Caneva and Stéphanie Paul. *Mythos* 8 (2015): 35-53.
Kyrieleis, Helmut. *Hellenistische Herrscherporträts auf Siegelabdrücken aus Paphos*. Wiesbaden: Reichert Verlag, 2015.
Messina, Vito. "Da Uruk a Seleucia: Pratiche amministrativi e archivi della Babilonia seleucide." *Mesopotamia* 40 (2005), 125-144.
Plantzos, Dimitris. *Hellenistic Engraved Gems*. Oxford: Oxford University Press, 1999.
_____. "The Iconography of Assimilation: Isis as Royalty on Ptolemaic Seal- impressions." In: *More than Men, Less than Gods: Studies on Royal Cult and Imperial Worship*, proc. of the international colloquium of the Belgian School at Athens, November 1-2, 2007 (Studia Hellenistica 51), ed. by Panagiotis P. Iossif, Andrzej S. Chankowski and Catharine C. Lorber, 389-415. Louvain: Peeters, 2011.
Szabados, Anne-Violaine. "From the *LIMC* Vocabulary to LOD: Current and Expected Uses of the Multilingual Thesaurus *TheA*." In: *Information Technologies for Epigraphy and Cultural Heritage*, ed. by Silvia Orlandi *et al*., 51-67. Roma: Università La Sapienza Editrice, 2014.

KB-Driven Information Extraction to Enable Distant Reading of Museum Collections

Giovanni A. Cignoni, Progetto HMR, giovanni.cignoni@di.unipi.it
Enrico Meloni, Progetto HMR, enrico-meloni@hotmail.it

Short abstract (same as the one submitted in ConfTool)

Distant reading is based on the possibility to count data, to graph and to map them, to visualize the relations which are inside the data, to enable new reading perspectives by the use of technologies. By contrast, *close reading* accesses information by staying at a very fine detail level. Sometimes, we are bound to close reading because technologies, while available, are not applied.
Collections preserved in the museums represent a relevant part of our cultural heritage. Cataloguing is the traditional way used to document a collection. Traditionally, catalogues are lists of records: for each piece in the collection there is a record which holds information on the object. Thus, a catalogue is just a table, a very simple and flat data structure. Each collection has its own catalogue and, even if standards exists, they are often not applied.
The actual availability and usability of information in collections is impaired by the need of accessing each catalogue individually and browse it as a list of records. This can be considered a situation of mandatory close reading.
The paper discusses how it will be possible to enable distant reading of collections. Our proposed solution is based on a *knowledge base* and on *KB-driven information extraction*.
As a case study, we refer to a particular domain of cultural heritage: *history of information science and technology* (HIST). The information corpus about HIST is particularly suitable to distant reading methods. Moreover, being information technology pervasive of everyone life we need new ways of telling HIST: distant reading may help to engage people in the discovery of HIST and in the understanding of the science behind today informatics and in the comprehension of cultural and social phenomena generated by use and habit of these technologies.

Introduction

Today technologies make new reading approaches viable. *Distant reading* is based on the possibility to count data, to graph and to map them, to visualize the relations which are inside the data, thus enabling new reading perspectives (Moretti 2005). By contrast, *close reading*, is the way to access information by staying at a very fine detail level. There are cases in which we are bound to close reading because the technologies are available but not applied.
A relevant part of our cultural heritage is made of the collections preserved in museums. Cataloguing is the traditional way used to keep and to convey documentation about a collection. Catalogues are lists of records: for each piece in the collection there is a record which keeps information about the physicality of the object (the size, the materials, the state of conservation...), its origin (the author, the place where it was manufactured), its day by day management (its place in the museum exhibits o in the deposits, if it is on loan...).
As a data structure, a traditional museum catalogue is just a table, that is a very simple and flat data structure – we could say primitive. Moreover, each collection has its own catalogue and often, even if standards exist, it adopts its own schema for the data fields in the table.
The actual availability and usability of information about cultural heritage in collections is impaired by the need of accessing the catalogues individually and then by browsing them as lists of records. We could say that we are in a situation of *mandatory close reading*.

The paper discusses how it will be possible to enable distant reading of collections. We will refer to a particular domain of scientific and technological heritage: the *history of information science and technology* (HIST). Apart of being our field of historical interest, it is an interesting case study. First of all, the information corpus about HIST is particularly suitable to be investigated with distant reading methods. Moreover, being information technology pervasive of everyone life, there is interest – as well as need – for new ways of telling HIST: distant reading may help to engage people in the discovery of the many stories of HIST, in the understanding of the science behind today informatics and in the comprehension of the cultural and social phenomena generated by use and habit of these technologies.
Section 2 of the paper presents our specific domain and defines the requirements for enabling distant reading of HIST starting from the museum collections. Section 3 describes the solution we are proposing which is based on a *knowledge base* (KB) and on *KB-driven information extraction* (KB-DIE). Section 4 is devoted to discuss related approaches and methods and to analyse the feasibility of the implementation of our solution.

Distant reading of collections, the HIST case study

Information technologies are born in the Fifties and have quickly flourished in the following years. While they can be considered recent, they pervasively affect our everyday life, thus they are a proper cultural heritage of humanity – and not only as a scientific and technological matter.
As a natural consequence of the impact of informatics on the society, there is a growing curiosity about HIST: protagonists such Alan Turing or Steve Jobs have become pop icons celebrated in popular movies (Tyldum 2014, Stern 2013, Boyle 2015). At the same time, the interest in the conservation of HIST relics is raised. Important collections belong to museums generally devoted to science and technology, such as, among the best known, the *Science Museum*[1] of London, the *Deutsches Museum*[2] in Munchen, the *Conservatoire National des Arts et Métiers*[3] in Paris, the *Museum of Science and Industry*[4] in Manchester. There are also museums specifically dedicated to HIST like the *Computer History Museum*5 in Mountain View, the *Heinz Nixdorf Museumforum*[6] in Paderborn, or *The National Museum of Computing*[7] at Bletchley Park. In addition to museums there are collections belonging to enthusiasts, often organized in *retro-computing* clubs and associations. People involved are usually very cooperative, often also very competent on specific topics. While not open to the public on a regular basis, such collections represent an important contribute to research and preservation about HIST. In the following, collections have to be intended in a wide and participative way.

The starting point, from catalogues to KB

The idea of passing from many different, flat catalogues of collections to a unique shared KB has already been presented, as a general idea and as a feasible international project (Cignoni and Cossu 2016). A very basic prototype, namely *CHKB*[8], has been developed as part of several students' theses at the University of Pisa to investigate how a simple KB structure can be implemented and how users could populate it through a web interface. A KB has several advantages. It keeps the complexity of relations among the facts of the HIST domain. Being unique, the KB collects all the knowledge and

1 http://www.sciencemuseum.org.uk/ (last accessed November 13, 2016).
2 http://www.deutsches-museum.de/en/exhibitions/communication/computers/ (last accessed November 13, 2016).
3 http://www.cnam.fr/ (last accessed November 13, 2016).
4 http://msimanchester.org.uk/ (last accessed November 13, 2016).
5 http://www.computerhistory.org/ (last accessed November 13, 2016).
6 https://www.hnf.de/en/museum.html (last accessed November 13, 2016).
7 http://www.tnmoc.org/ (last accessed November 13, 2016).
8 http://hmr.di.unipi.it/CHKB_en.html (last accessed November 13, 2016).

makes it surfable at different levels, both for experts and researchers and for the general public. The KB is also an authoritative source: all the facts pass through a peer review revision process by an open group of experts which cooperates to assure a high level of reliability of the KB content. For the details, we refer to the cited works, here we focus on how such a KB can allow distant reading of HIST.

Distant reading of HIST by enriching the KB contents

Distant reading may greatly support the work of HIST scholars. The domain is characterized by many complex relations. From the perspective of relics conservation, the KB explicits the relation among the pieces in the collections and the product they are instances of. For example, from the KB it is possible to build the map of all the preserved *Apple][* telling which of them are on exhibition, which of them are in working conditions, which of them are periodically demonstrated.

From a more historical perspective, the Apple][was a product of *Apple*, it has a predecessor (the *Apple-1*) and a successor (the *Apple ///*), it has a designer (*Steve Wozniak*). Even more interesting in the KB are the relations concerning the software or hardware components. For instance, the Apple][uses the *MOS 6502* microprocessor like many other devices of the time which belong to different categories: *personal computers* (like the Apple ones as well as the *Commodore PET* series, or the less known *Ohio Scientific Challenger P* series), *home computers* (as the *Commodore Vic-20* or the *Acorn BBC Micro*) or *videogame consoles* (as the *Atari 2600* or the *Nintendo Entertainment System*)... the "genealogic tree" of the 6502 extracted from the KB helps to visualize a technology in terms of market products. There are a lot of interesting possibilities to count, graph, map and visualize the KB content and they are useful for a public that is wider than scholars.

Museums have a mission of culture preservation and diffusion: their catalogues, though originally addressed to researchers, are the core information used to design and set-up exhibitions. Publishing catalogues on the web[9] is part of an effort to engage a wider audience and to build public awareness and knowledge about HIST cultural heritage.

With respect to traditional "flat" catalogues, presenting views in the large may be an appealing way to capture the attention of the public and to stimulate its curiosity. Distant reading helps the historians to comprehend scientific and technological history as well as its cultural and social consequences and repercussions. In terms of visualization – such as the technological genealogic trees – distant reading also helps to convey HIST to the public, both as bare facts and as interpretations of the historians – as they can use the views to support their findings.

Moving in the large

Catalogues of collections are a primary source of information about HIST, but not the only one. There are many other sources: books, media, web pages. The web is a wide and very general container: many books and media are online, at least partially via Google Books or YouTube; there are pages belonging to institutions (some museums have their catalogues online); there are online newspapers and magazines with pages dedicated to technology and HIST is increasingly a topic of interest. There also are pages and blogs written by enthusiasts which, often, offer very valuable information (like deep technical knowledge) mixed with biased opinions (overstated appraisal of a brand or a particular model). Moreover, lot of knowledge is still in the memories of people who were part of the (recent) history of computer industry and is increasingly confided in posts on social networks – again on the web.

Distant reading is usually related to large amounts of data. To extend in the large the contents of our KB it is possible to enrich the information related to collections with additional facts about HIST by extracting them from the web. Browsing the web in search of information worth being inserted in the KB

9 see for instance the Computer History Museum http://www.computerhistory.org/collections/search/ (last accessed November 13, 2016)

has two other advantages: helps the HIST research and fosters a richer presentation of the history behind the pieces preserved in the museums.

Additional sources – not always reliable

Feeding the KB browsing the web raises the problem of accuracy and reliability of the information. Which is not a flaw of bloggers and enthusiasts only. Sources like newspapers and magazines should be trustworthy. Yet, the need for stunning news results in highly inaccurate information.

Recently, *La Repubblica*, one of the most important Italian newspapers, titled "The first electronic music track revived, created by Alan Turing" (Rivive la prima traccia musicale elettronica creata da Alan Turing, in the original Italian title[10]). Unfortunately, it was not the first music played by a digital computer and it was not created by Alan Turing. The true story[11] is that the first *recording* of a music played by a digital computer, which was already known, was digitally remastered. Turing was involved in the project at the University of Manchester that built the computer, but the author of the music program was Christopher Strachey. Probably, La Repubblica, as many others, copied the scoop from another (unreliable) source[12], in any case many people were induced to believe that Turing was also the first digital musician of the history.

The above example depicts a serious problem we have to cope with, especially if we want to feed the KB with information automatically extracted from the web: it is difficult to discern good information. Sometimes trustworthy sources, like renowned national newspapers, publish false news; moreover, replication of the same information on the web, which usually follows the news, does not work as an indicator of reliability.

Estimating the KB size

Following a *Fermi problem* approach, we can attempt an upper-bound estimation of the size of the knowledge about HIST. The main goal is to support the feasibility of our solution by giving evidence of a manageable size. We can use the number of *triples* as an unit of measure (and just as that), assuming that triple-store is both a well known paradigm and a quite mature technology with some reference benchmarks (see for instance Oracle 2016). We also consider the knowledge intrinsic in the HIST collections: they are our first interest and a huge concentration of information.

The size of the KB can be estimated by the following formula:

catalogue record fields × *collection size* × *collections for country* × *developed countries*

As the number of the catalogue record fields we can take as an example the *ICCD PST* record proposed by the Italian *Central Institute of Cataloguing and Documentation* for the technological and scientific heritage which count (in the last version) 327 fields (of course not all appliable to every object)[12]. To each field can roughly correspond a triple.

As the collection size we can use the 115233 records of the online catalogue of the already cited Computer History Museum – probably the largest one.

As the number of collections in a country we can take the recent survey[13] made by the Italian Association for Automatic Computing which registered 58 collections in Italy, including museums, retrocomputing associations and privately owned ones.

As the number of developed countries, among many candidates, we can use the number of members of the *Club de Paris*[14].

10 http://www.repubblica.it/tecnologia/2016/09/26/news/turing_musica_computer-148561122/ (last accessed November 13, 2016).
11 see for instance http://blogs.bl.uk/sound-and-vision/2016/09/restoring-the-first-recording-of-computer-music (last accessed November 13, 2016).
12 http://iccd.beniculturali.it/getFile.php?id=258 (last accessed January 8, 2017)
13 http://www.aicanet.it/dettaglio_evento/598532 (last accessed January 8, 2017)
14 http://www.clubdeparis.org/ (last accessed January 8, 2017)

Summing up:

$327 \times 115\,233 \times 58 \times 22 = 48\,081\,199\,716 \approx 5 \times 10^{10}$

Which is a quite large number of triples, yet twenty times smaller than the trillion limit of current technologies (see again, as an example, Oracle 2016). Furthermore, our number actully estimates the number of record fields in the case of flat independent catalogues: using a shared KB eliminates replication of information.

KB and KB-driven Information Extraction

As we said before, the first step for populating the KB is bringing together all the museums catalogues. This can be done via some importing procedure from collection catalogues whose results are validated through the mentioned peer review process; supplementary information can be added by editors and reviewers from personal knowledge and research.

Adding to the KB the great amount of information contained in the web can be done via *ontology based information extraction* (OBIE) methods. An OBIE system processes unstructured or semistructured natural language text through an ontology driven mechanism and outputs information structured with respect to the same ontology (Wimalasuriya et al, 2010). The ontology which structures our KB can be used to drive the *information extraction* (IE) process.

However, OBIE systems are not able to avoid all the pitfalls of bad information often found in the web – the above-mentioned musician Turing case can be an example that can deceive an OBIE system, as the fact "first computer music created by Alan Turing" is structurally correct.

To improve the reliability of OBIE, we propose KBDIE. We call a KBDIE system one where the KB plays a double role: the ontology on which the KB is structured is used as in a OBIE system, then the contents of the KB (which we assume not empty and validated) are used to check IE output. In practice, the KB contents are used as a reference in a sort of machine learning process.

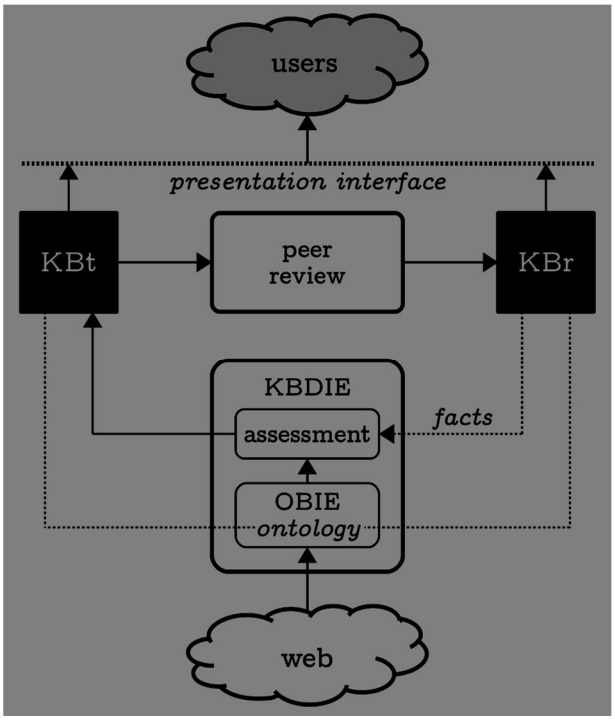


Fig. 1. The KB driven information extraction process

As shown in fig. 1, in our solution the KB is made of two KBs sharing the same ontology. A *reliable KB* (KBr) contains the validated facts that passed the peer review process. A *temporary KB* (KBt) contains only the automatically extracted facts.

The facts in KBr are used to check the IE output and to assess it in terms of *novelty* (that is IE found something that is not in KBr) and *reliability* (what IE found does not negate something which is in KBr). The assessment marks the facts in KBt with scores which help the reader in the interpretation of the KB content. In general, reliability may include other aspects like *temporal validity* and dependance on *personal opinions* – which are lecit. Being an historical knowledge base, time should not be of too much concern. Fans of a particular machine or brand – say for instance the Commodore 64 – may write on the web biased information. This is harder to spot unless there are evident contradictions with KBr facts – this is one of the reasons for having reviewers in the process.

Back to our musician Turing example, if the KBr contains the fact "first computer music created by Geoof Hill", which refers to the very first music created and played (but not recorded) on the Australian CSIRAC (P. Doornbush, 2005) and/or the fact "first computer music recorded by Christopher Strachey", which refers to the event in Manchester, then the assessment phase of KBDIE can detect a mismatch with the found fact "first computer music created by Alan Turing" and put it in KBt with a "suspicious" score.

KBt facts, even if assessed, cannot go directly in the KBr, as the latter must contain only reliable knowledge. As part of the peer revision process, reviewers examine the KBt facts and move in the KBr those which are considered reliable (after some editing if needed). In other words, automatically extracted information is treated like contents submitted for insertion in KBr as results of the research of keepers, curators and historians in general. To end our example, the fact "first computer music created by Alan Turing" will never go in KBr – unless we consider the idea of a "pit of shame" where bad examples of HIST popularization can be relegated.

Each fact in KBt has the URL of the web source it comes from. The URL can be used in the calculation of the reliability score (depending on previous finding from the same source) and can be used by the reviewers for an additional check. If the fact is deemed reliable and is moved in KBr, the URL is maintained both as a reference and as an indicator of the source trustworthiness.

To the users, KBt and KBr are seen as an unique KB. All the facts are accessible, the difference is in the score, those actually belonging to KBr are marked as "reliable" and, besides the URL of the source (possibly more than one), refer also the names of the reviewers.

Reviewers, as well as editors that contribute by submitting facts directly in KBt, do a very valuable job. Reviewers, being at the end of the process, will be provided with a big amount of information to validate. The automatic assessment is a valuable filter, but cannot guarantee the same level of confidence that an expert can offer. Yet, the work of reviewers should not be very taxing. We expect that they are keepers, curators and historians which use the KB for their usual work and research. They can use KBDIE to order custom searches as well as to submit to the KB the results of their findings. Furthermore, sharing and being in contact with colleagues is a benefit for everyone.

Related works

In many fields of research, history shows a shift from "little" to "big science". Our proposal stems from the idea of moving in the large the management of the knowledge about a particular sector of cultural heritage: the collections about HIST. The benefits regard the ability to look at that knowledge from different perspectives, as well as the enabling of cooperation and sharing processes among scholars – and collaboration plays a big role in research (N. Vermeulen et al., 2013).

In (G.A. Cignoni and G.A. Cossu, 2016) we already presented the general idea[15], discussing the differences with respect to traditional catalogues, like for instance the standards proposed by the Italian Central Institute of Cataloguing and Documentation. In the same work we also discussed the

15 The idea has been also presented and discussed at the AIUCD 2015 Conference (http://www.aiucd.it/digital-humanities-e-beni-culturali-quale-relazione-quarto-convegno-annuale-dellaiucd/, last accessed January 8, 2017) and in a workshop organized by AICA and the University of Verona (http://www.aicanet.it/dettaglio_evento/598532, last accessed January 8, 2017).

differences with respect to very general approaches of union of many catalogues such as, for instance, the *Europeana* project[16] – which adds standard metadata that may help the information extraction and assessment tools of our proposal.

In the proposed solution, like in all OBIE systems, the ontology which structures the KB is designed by experts, defining a priori the entities of the domain and their relations (D. Wimalasuriya, 2010). In the KBDIE method we use facts from the KBr as a guide for accuracy evaluation of the new information extracted from the web.

Machine learning methods are considered very effective to support IE: induction from data produces better results with less effort with respect to definition of formal rules for a logical reasoning system (V. Tresp et al., 2008). In facts, logical reasoning has not proven effective when applied to the web scale nor suitable to manage uncertain information, which is abundant on the web (A. Rettinger et al., 2012). In our case, machine learning is achieved through constantly improving the contents of KBr and by maintaining information about the trustworthiness of sources.

Several approaches focus on learning ontologies, that is using machine learning to extend and improve the ontologies, which is promising on some fields of application (J. Völker et al., 2008). However, in the KBDIE method we are not interested in obtaining a better ontology, as it is assumed that, given the HIST domain, it is possible for experts to design the ontology in advance and we expect it to be stable in time; moreover, being the KB structured on the ontology, changes in the ontology may imply a reorganization of KB contents, which in our case should be supervised.

In (H. Xie et al., 2016) is proposed a statistical evaluation algorithm, which uses a corpus dataset as a reference for evaluating KB triples. Being our triples like <Apple][; released ; 1977> we cannot rely on average values desumed by many repetitions of the same information with slightly different values.

A different approach is in (M. Wick et al., 2013) where sets of triples are scored depending on their inner coherence. This is an interesting direction for building the score, yet in our case supervision is needed: musician Turing would be erroneously considered valid because the news was repeated by many sources. Once that enough facts are stored in KBr, it is likely that using the knowledge in KBr the inaccurate news would be spotted as suspicious.

To improve scoring of found facts, we can use the category of the source: information extracted from a digital copy of a technical manual is more reliable than the one extracted from an advertising brochure. In (F. Sebastiani, 2002) some methods are described for categorising natural text using unsupervised machine learning. As far as it is limited to sources categorisation, this approach looks promising for our case too.

Being mainly interested in assuring the reliability of KBr contents, we generally prefere a supervised approach to accept new facts in KBr. Apart of the role of the KB as reference for HIST, KBr facts are used to assess the new facts in the KBDIE process. Inaccuracies in KBr will affect the assessment and will cause a general decrease of reliability of the whole system. As noted by (M. Wick et al., 2013), IE systems, however advanced, can offer a confidence level that is not comparable to the confidence of a human expert.

Acknowledgments

We wish to thank the many involved in previous fruitful discussions. Those involved in the development of the prototype, in particular Giuseppe Lettieri from the University of Pisa, and those currently involved in the proposal of an Horizon 2020 EU project on these themes: Giovanni A. Cossu, Norma Zanetti and Sauro Salvadori from Hyperborea srl, Franco Niccolucci from the University of Florence.

16 Europeana Collections, http://www.europeana.eu, (last accessed January 8, 2017)

References

Boyle, Danny. 2015. *Steve Jobs* (movie).

Cignoni, Giovanni A., and Giovanni A. Cossu. 2016. *The Global Virtual Museum of Information Science & Technology*. Proceedings of International Communities of Invention and Innovation, IFIP WG 9.7 Conference, New York, to be published in Advances in Information and Communication Technology, no. 491, Berlin: Springer.

Doornbusch, Paul. 2005. The Music of CSIRAC: the Australia's First Computer Music. Champaign: Common Ground.

Dou, Dejing, Hao Wang, and Haishan Liu. 2015. *Semantic Data Mining: A Survey of Ontology-based Approaches*. Proceedings of the IEEE 9th International Conference on Semantic Computing (ICSC), USA: IEEE.

Jänicke, Stefan, and Greta Franzini, Muhammad Faisal Cheema and Gerik Scheuermann. 2015. *On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges*. Proceedings of the EuroVis 2015 State-of-The-Art Reports.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Oracle. 2016. *Oracle Spatial and Graph: Benchmarking a Trillion Edges RDF Graph* (white paper).

Rettinger, Achim, Uta Lösch, Volker Tresp, Claudia D'Amato, and Nicola Fanizzi. 2012. *Mining the Semantic Web. Statistical learning for next generation knowledge bases*. Data Mining and Knowledge Discovery, vol. 24 no. 3, 613-662.

Sebastiani, Fabrizio. 2002. *Machine learning in automated text categorization*. ACM Computing Surveys (CSUR), 1-47. New York: ACM.

Stern, Joshua Micheal. 2013. *Jobs* (movie).

Tresp, Volker, Markus Bundschus, Achim Rettinger, and Yi Huang. 2008. *Towards Machine Learning on the Semantic Web*. In Uncertainty Reasoning for the Semantic Web I, edited by Paulo Cesar G. da Costa, Claudia D'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewcz, Matthias Nickles, Michael Pool, 282-314. Berlin: Springer.

Tyldum, Morten. 2014. *The Imitation Game* (movie).

Vermeulen, Niki, John N. Parker, and Bart Penders. 2013. *Understanding life together: A brief History of collaboration in biology.* Endeavour, vol 37 no.3, 162-171.

Völker, Johanna, Peeter Haase and Pascal Hitzler. 2008. *Learning Expressive Ontologies*. Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, 45-69. Amsterdam: IOS Press.

Wick, Michael, Sameer Singh, Ari Kobren and Andrew McCallum. 2013. *Assessing confidence of knowledge base content with an experimental study in entity resolution*. Proceedings of the 2013 workshop on Automated knowledge base construction, 13-18. New York: ACM.

Wimalasuriya, Daya C., and Dejing Dou. 2010. *Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches*. Journal of Information Science vol. 36 no. 3.

Wimalasuriya, Daya C., and Dejing Dou. 2010. *Components for information extraction: ontology-based information extractors and generic platforms*. CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management.

Xie, Haihua, Xiaoqing Lu, Zhi Tang, and Mao Ye. 2016. *A Methodology to Evaluate Triple Confidence and Detect Incorrect Triples in Knowledge Bases*. Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, 251-252. New York: ACM.

Sentiment Analysis d'autore: l'epistolario di Italo Svevo

Cristina Fenu, Biblioteca civica "A.Hortis" di Trieste, cristina.fenu@comune.trieste.it

Rendere visuale un museo letterario: *Sentiment Analysis* e la rappresentazione grafica dell'emotività di Italo Svevo

"Rendere visuale un museo letterario": è l'obiettivo del progetto *La città di carta. Museo virtuale della letteratura triestina* che ho realizzato come *project work* per la prima edizione del Master in *Digital Humanities* dell'Università Ca' Foscari di Venezia presentando, in collaborazione con il Museo sveviano di Trieste, il Dipartimento di Ingegneria e Architettura dell'Università degli studi di Trieste e il Dipartimento di Scienze Ambientali, Informatica e Statistica di Venezia un prototipo di museo letterario virtuale partendo dalle collezioni sveviane.

Un piccolo museo, quello Sveviano, compresso in uno spazio espositivo esiguo, ma potenzialmente esplosivo: un autentico ordigno degno di Svevo. Un museo in parte nascosto, con manoscritti inediti costretti al buio di scatole d'archivio, ma pieni di vita, di storie, di immagini. Raccontare queste storie, far vedere queste immagini, rendere visuale il museo letterario dello scrittore emblema della letteratura triestina del Novecento. Sì, ma come? Il presente contributo descrive i risultati preliminari della realizzazione *in fieri* di un *Archivio digitale Italo Svevo* che ha preso le mosse dall'applicazione di tecniche di *Sentiment Analysis* e *data mining* su parte del *corpus* dell'archivio Svevo, l'epistolario édito dello scrittore (Svevo 1965; Svevo 1966)[1].

Durante lo *stage* previsto dal Master ed effettuato presso il Machine Learning Lab (MaLeLab) del Dipartimento di Ingegneria e Architettura dell'Università di Trieste sotto il tutoraggio del dott. Eric Medvet, docente di *Machine Learning and Data Mining*, ho analizzato il *corpus* delle epistole di Svevo con tecniche di *Sentiment Analysis*. Lo scopo è ottenere una visualizzazione grafica della rete di relazioni personali e intellettuali dello scrittore, connotata per picchi di emotività positiva o negativa, e rendere visibile a colpo d'occhio, mediante grafi di immediata leggibilità anche per un pubblico di non specialisti, la personalità di Ettore Schmitz, facendo "entrare nella testa" dello scrittore i visitatori del Museo sveviano *in loco* e da remoto, tramite il sito web[2].

Sentiment Analysis d'autore: lo stato dell'arte

Fra i più interessanti contributi e *case-studies* di recente realizzazione o pubblicazione inerenti l'applicazione di *Sentiment Analysis* e di altre tecniche (*Network* e *Topic Analysis*) su *corpora* testuali d'autore ed epistolari, vanno senz'altro citati gli studi di Saif M. Mohammad, ricercatore in linguistica computazionale presso il National Research Council of Canada, concernenti la creazione di *Emotion Lexicons* (Mohammad and Turney, 2010) e *SA* applicata a testi narrativi e *email* in lingua inglese (Mohammad, 2011). Vanno menzionati anche i lavori di Matthew Jockers[3] che sul proprio

1 Il progetto *La città di carta* comprende due accessi "visuali" a un patrimonio documentale unico che connota la cultura triestina – e non solo – del Novecento: la realizzazione di un *Archivio digitale* – da me curato - e dello *Svevo Virtual Tour* mediante *software* di realtà aumentata, a cura mia e di Elena Griguol.

2 Quanto illustrato è visibile sul sito www.museosveviano.it/ar/, un'installazione Wordpress presso il server web ufficiale del Museo sveviano, in continuo aggiornamento col procedere delle analisi testuali.

3 Jockers è Associate Dean for Research and Global Engagement presso il College of Arts & Sciences e Associate Professor of English alla University of Nebraska; è anche Faculty Fellow presso il Center for Digital Research in the Humanities e direttore del Nebraska Literary Lab. Tiene un *blog* in cui rende disponibili i risultati delle sue ricerche in *computational text analysis* (http://www.matthewjockers.net/).

blog riporta i risultati dell'applicazione di *tool* e tecniche di *Sentiment Analysis* su *Portrait of the Artist as a Young Man* di James Joyce (Jockers, 2014b). Imprescindibile per imponenza e qualità del progetto nel campo della *Network Analysis* e della visualizzazione ed esplorazione interattiva di *corpora* di lettere è sicuramente *Mapping the Republic of Letters* (http://republicofletters.stanford.edu/). Per quel che riguarda più in generale gli strumenti di visualizzazione utilizzati nelle *Digital Humanities* nell'ambito del *distant reading* si fa infine riferimento alla recente panoramica tracciata in (Jänicke et al., 2015).

L'analisi testuale del carteggio sveviano con software R: procedure, sperimentazione, primi risultati

Le lettere che compongono l'epistolario édito di Italo Svevo costituiscono un *corpus* di 894 documenti i cui originali sono nella quasi totalità conservati in Museo sveviano. Si tratta di documenti già digitalizzati e descritti archivisticamente in un *database*. In preparazione alla *text analysis* ho proceduto a

1. convertire in formato testo (standard UTF8) i file *tiff* dei documenti digitalizzati tramite *software ocr FineReader* con lettura supervisionata e correzione degli errori di *output* in fase di conversione
2. uniformare le intestazioni delle singole lettere contenute in un unico file *txt*, premettendo a ciascuna l'indicazione dei seguenti attributi
   - mittente
   - luogo da cui scrive il mittente
   - destinatario
   - luogo in cui si trova il destinatario
   - data
   - lingue utilizzate
   - numero progressivo di lettera nell'epistolario édito utili alla successiva operazione di *export* in formato tabellare *csv*
3. importare la tabella *csv* – suddivisa in 12 variabili e 894 osservazioni – in ambiente di lavoro "R Studio" per sottoporre i testi dell'epistolario a analisi statistiche tramite software "R".

Presso MaLeLab dell'Università di Trieste ho iniziato a effettuare *Sentiment Analysis* sul *corpus* dell'epistolario sveviano sperimentando le potenzialità di *Syuzhet Package* - una delle librerie disponibili in ambiente "R"[4], scelta perché tarata specificatamente dal suo creatore Matthew Jockers per l'analisi di testi d'autore e di *plot* narrativi tramite la rilevazione di picchi emotivi nella scrittura. Inoltre *Syuzhet Package* comprende il lessico *NRC Word-Emotion Association Lexicon* (*EmoLex* http://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm) costituito da più di 14.000 lemmi associati per aree semantiche a otto emozioni-base (rabbia, paura, aspettativa, fiducia, sorpresa, tristezza, gioia, disgusto) pertinenti la polarità *sentiment* positivo-negativo (Mohammad and Turney, 2010).
Inizialmente mi sono concentrata sul solo carteggio Svevo-Joyce, un *corpus* minimo, di appena dieci lettere, in lingua prevalentemente inglese e con qualche frase in italiano e in dialetto triestino. *Syuzhet Package* elabora estrazioni di *sentiment* basandosi esclusivamente su *token* di lingua inglese, fornendo in questo caso risultati utili circa la rilevazione del *sentiment* data la presenza sporadica di espressioni non anglofone nei testi analizzati.

Ma in realtà Ettore Schmitz nelle sue lettere usa contemporaneamente l'italiano, il tedesco, il francese e l'inglese, assieme a moltissime espressioni dialettali triestine e più di qualche lemma latino

4 *Syuzhet Package* comprende anche più metodi di estrazione di *sentiment* e di visualizzazione della temperatura emotiva dei testi tramite una molteplicità di grafi (https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html).

e russo, senza attenersi esclusivamente agli usi linguistici dei suoi interlocutori. Per poter procedere con la *Sentiment Analysis* sul *corpus* dell'epistolario sveviano è stato perciò necessario individuare un lessico multilingue. Dal luglio 2015 *EmoLex* è stato implementato con un lessico plurilingue che consiste nella traduzione del vocabolario inglese in una ventina di idiomi tramite algoritmo *Google Translate,* come riferisce Mohammad nel suo *blog* (http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm). Tuttavia questo lessico plurilingue non è mai stato implementato né testato in alcuna libreria di *SA* di "R", *Syuzhet Package* compreso. Per poter sperimentare un primo utilizzo della versione multilingue di *EmoLex*, lo *staff* del MaLeLab ha realizzato un'estensione di *Syuzhet Package* da un codice suggeritomi dallo stesso Jockers. Sono così riuscita a estrarre i valori di *sentiment* connotanti ciascuna delle 894 lettere dell'epistolario sveviano per le otto emozioni-base, ciascuna validata per lingua.

Successivamente ho verificato come la qualità dei risultati della *Sentiment Analysis* può venir influenzata effettuando il preprocessamento del testo noto come *stemming*, che consiste nel ricondurre ogni parola alla sua radice morfologica, deprivandola quindi del suffisso di declinazione e coniugazione: in particolare, ho applicato lo stemmer *Snowball* sia ai *token* ricavati dalle lettere che agli elementi del vocabolario *EmoLex*.

Per ognuna delle varianti di *Sentiment Analysis* provata, segnatamente quelle con o senza *stemming*, ho esaminato a campione le lettere che costituiscono il carteggio édito tra Italo Svevo e Eugenio Montale - 62 documenti in tutto - verificando che il *sentiment* assegnato fosse coerente con quello rilevabile leggendo ciascuna missiva.

Ho così potuto constatare come la *Sentiment Analysis* basata esclusivamente su lessico non dia risultati eccellenti dal punto di vista quantitativo quando applicata a un dominio linguistico e comunicativo estremamente peculiare, come quello da me analizzato. L'approccio multilingue, per di più all'epistolario di un autore che usa con grande disinvoltura gli aspetti di polarità a fini artistici - come nel caso dell'ironia sveviana - rende problematica la *text analysis* basata su lessico applicata alle lettere di Italo Svevo se ci si attiene ad una valutazione squisitamente quantitativa dei dati rilevati dall'algoritmo, spesso molto scarni e a volte falsati, specie nella variante *SA* con *stemming*[5]. Va segnalato che sulla *Irony Detection* si concentra molta letteratura scientifica degli ultimi anni relativa alla *SA*, ma prevale l'attenzione sui *social* mentre latita ancora quella dedicata all'ambito più segnatamente letterario: si vedano a riguardo alcuni recenti contributi (Reyse et al., 2011; Reyse et al., 2013). Tuttavia lo scopo del lavoro che intendo perseguire non è una sperimentazione spinta di *SA*, né tantomeno la definizione di un lessico italiano letterario per la *Sentiment Analysis*, ma rendere visuale il *sentiment* di un carteggio che giace inscatolato in faldoni, o il *plot* dei racconti sveviani, e rilevare, lungo la linea del tempo tratteggiata dalla produzione letteraria dello scrittore, l'andamento del *sentiment* di un autore che si è sentito fallito. Una semplificazione qualitativa, dunque, ma pienamente funzionale agli obiettivi dichiarati di questo progetto e del tutto perseguibile grazie alla visualizzazione in grafi dei risultati delle *text analysis* condotte, anche alla riprova del vaglio – ovvero della lettura dei documenti già analizzati dall'algoritmo.

"Rendere visuale" il carteggio di Italo Svevo: alcuni esempi[6]

Per visualizzare l'epistolario sveviano già un semplice diagramma alluvionale (fig. 1), una rappresentazione grafica del flusso delle lettere scritte da Ettore Schmitz ordinato in senso cronologico, consente di scorrere la lista dei suoi interlocutori - per lo più familiari e amici del giro delle conoscenze triestine - e rende immediatamente percepibile l'isolamento dello scrittore nel

5 I risultati relativi alla *SA* con e senza *stemming* sul *corpus* del carteggio Svevo-Montale sono disponibili sul sito web www.museosvevian.it/ar.

6 Le visualizzazioni portate come esempi sono state realizzate tramite i *software RAW* (http://app.raw.densitydesign.org/#%2F) e *Palladio* (http://hdlab.stanford.edu/palladio/).

contesto intellettuale europeo contemporaneo sino al 1925, l'anno della “resurrezione di Lazzaro” operata da James Joyce su Svevo per tramite della critica francese.

Solo a partire da quella data il carteggio acquista un respiro internazionale e le dimensioni di *network*, del dialogo tra intellettuali.

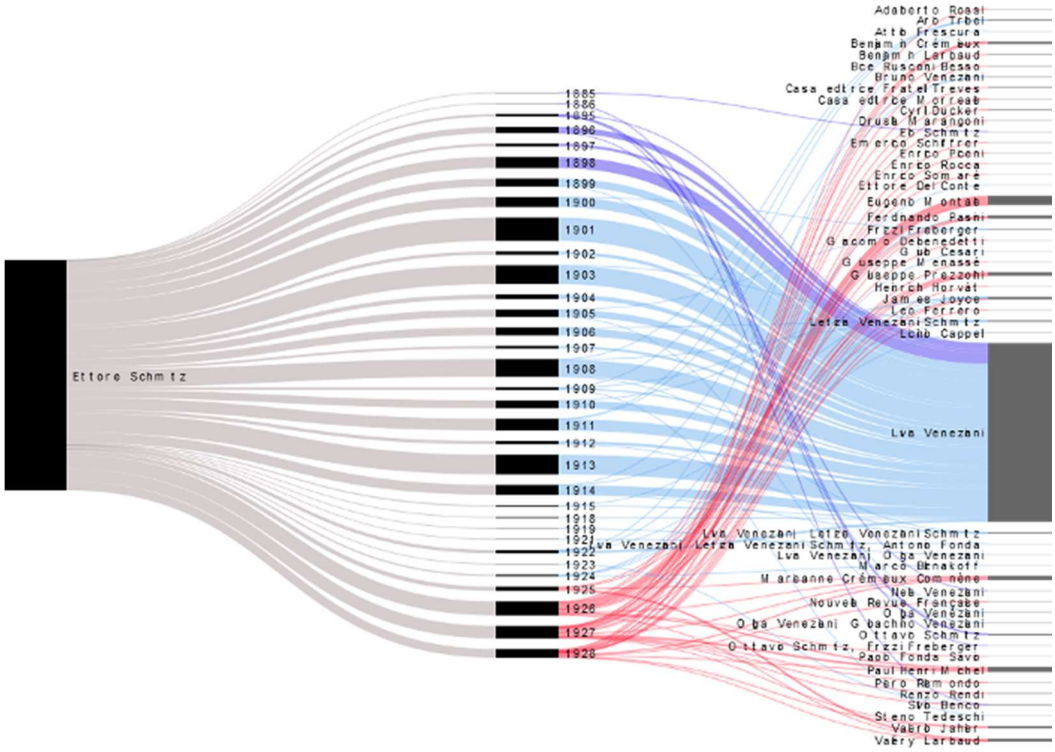


fig. 1. diagramma alluvionale del carteggio di Italo Svevo ottenuto da *data frame csv* tramite *tool* RAW

L’uso del colore associato alla visualizzazione grafica rafforza il significato immediatamente leggibile del grafo

- il colore blu evidenzia il flusso delle lettere scritte da Svevo sino al 1898, anno in cui viene pubblicato il suo secondo romanzo *Senilità*
- l’azzurro tinge il carteggio sveviano fino al 1925 e si spande sui 27 anni di silenzio della critica e dello scrittore, che dopo l’insuccesso di *Senilità* dichiara di aver rinunciato al vizio della letteratura
- il rosso rappresenta il “tramonto d’oro” di Italo Svevo - per prendere a prestito le parole di Saba - e colora i tre anni di gloria vissuti in vecchiaia dallo scrittore della *Coscienza di Zeno* (1923), rendendo visibile a colpo d’occhio anche l’ampliarsi dei contatti con autori e critici suoi contemporanei e estimatori.

Rendere visuale l’archivio sveviano significa anche consentire agli utenti di esplorare in maniera interattiva le lettere partendo sempre da un grafo che sintetizzi visivamente la struttura del carteggio stesso così come rilevata dalla *Sentiment Analysis*, ovvero un grafo a nodi stellato (fig. 2).

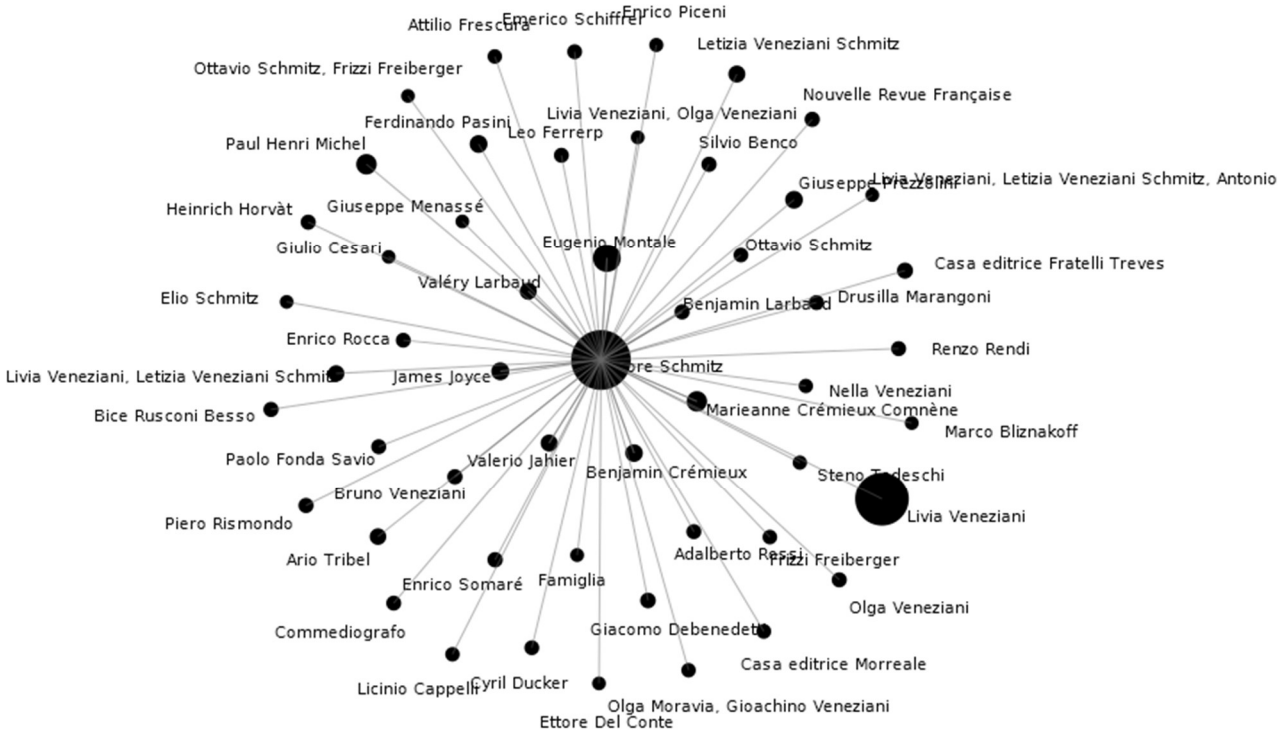


fig. 2: grafo a nodi stellato dell'epistolario sveviano ottenuto tramite tool *Palladio*

Anche in questo caso, a fini esemplificativi, ho scelto un solo *corpus* particolarmente significativo, il carteggio tra Svevo e Eugenio Montale. Il faldone contrassegnato dalla segnatura archivistica Museo Sveviano FS Corr. A 76 contiene il carteggio tra lo scrittore e il poeta appena trentenne autore nel novembre-dicembre del 1925 dell'*Omaggio a Italo Svevo* pubblicato nella rivista *L'esame*, la scintilla che, con due mesi di anticipo sulla critica francese, innesca il "caso Svevo" in Italia e in Europa. Sul sito del progetto *La città di carta* l'intero carteggio Svevo-Montale è sfogliabile per immagini scorrendo la *gallery* delle lettere che lo compongono, cui si accede cliccando sul nodo del grafo contrassegnato dall'etichetta "Eugenio Montale". Sempre sul sito sono disponibili i risultati della *SA* condotta su ciascuna missiva del medesimo carteggio.

Per portare ancora un esempio dell'intento comunicativo-visuale che riassume le finalità del progetto, nella figura 3 viene rappresentato il *sentiment* della prima lettera di Svevo a Montale datata 17 febbraio 1926. L'immagine rende percepibile a colpo d'occhio il sentimento di gratitudine dello scrittore triestino nei confronti del suo primo estimatore illustre in Italia. Nel grafo il *sentiment* rilevato dall'algoritmo sulla base della classificazione delle otto emozioni impostata secondo i criteri di *NRC Word-Emotion Association Lexicon* è evidenziato cromaticamente e per aree dalla superficie corrispondente alle percentuali rilevate dal *software*.

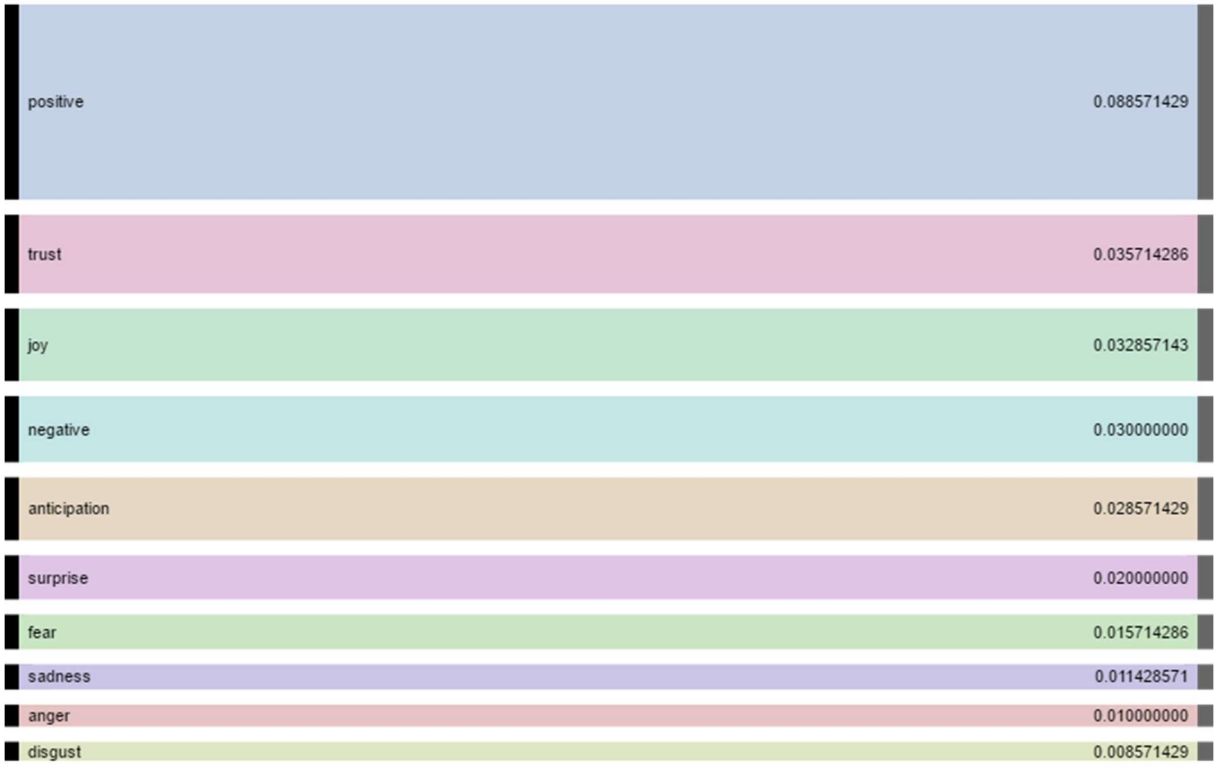


fig. 3: rappresentazione del *sentiment* della lettera di Svevo a Montale datata 17 febbraio 1926 tramite *tool* RAW

L'impatto visivo dà l'immediata percezione del contenuto emotivo di uno dei documenti più importanti dell'epistolario sveviano - e della storia della letteratura italiana del Novecento: una resa visuale che fornisce un punto di accesso diverso al testo della missiva, anticipandone e corroborandone il senso.

Future work: *Topic Modeling* dell'epistolario e della narrativa di Italo Svevo

Prossima tappa della sperimentazione delle tecniche di *text mining* sul *corpus* archivistico sveviano conservato nel Museo di Trieste sarà porre in relazione la rilevazione del *sentiment* con la *timeline* della produzione narrativa dello scrittore, analizzando l'epistolario e la narrativa di Italo Svevo con tecniche di *Topic Modeling* tramite algoritmo *LDA - Latent Dirichlet Allocation* - con classificazione non supervisionata. Per una formalizzazione delle problematiche inerenti le tecniche di *Topic Modeling* faccio riferimento a considerazioni e metodologie formulate in (Blai, 2012).

L'obiettivo rimane costante, dare rappresentazione visivamente immediata della temperatura emotiva della scrittura sveviana associata alle sue tematiche più ricorrenti con un approccio diacronico. La rappresentazione visiva delle reazioni di Svevo ai fiaschi editoriali delle sue due prime prove di romanziere e al successo della *Coscienza di Zeno* potrà essere rilevabile, infatti, anche quantificando la presenza/assenza di determinate parole-chiave nel *corpus* delle lettere, ad esempio l'intensificarsi dell'argomento "musica" e delle attestazioni della parola "violino", il palliativo all'astinenza forzata dallo scrivere romanzi, *leitmotiv* che connota l'epistolario sveviano dal 1898 fino agli anni Venti del Novecento.

Riferimenti Bibliografici

Blai, David M. “Probabilistic Topic Models. Surveying a suite of algorithms that offer a solution to managing large document archives”, *Magazine Communications of the ACM*, 4 (2012): 77-84. accessed November 25 2016, doi: 10.1145/2133806.2133826.

Jänicke, Stefan and Franzini, Greta and Cheema, Muhammad Faisal and Scheuermann, Gerik . 2015. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges” in *Eurographics Conference on Visualization (EuroVis)*, edited by R. Borgo and F. Ganovelli and I. Viola, 83-103. Geneve: The Eurographics Association.

Jockers, Matthew L.. 2014a. *Text Analysis with R for Students of Literature*, New York: Springer

Jockers, Matthew L.. 2014b. “A Novel Method of Detecting Plot” *Metthew L. Jockers blog*, June 5. http://www.matthewjockers.net//?s=A+Novel+Method+of+Detecting+Plot&search=Go

Mohammad, Saif M. 2011. “From Once Upon a Time to Happily Ever After: Tracking Emotions in Mail and Books”, in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105-114. Portland: Association for Computational Linguistics.

Mohammad ,Saif M. and Turney, Peter D. 2010. “Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon”. in *Proceedings of the NAACLHLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* , 26-34. Stroudsburg: ACL.

Reyse, Antonio and Rosso, Paolo and Buscaldi, Davide “From humor recognition to irony detection: The figurative language of social media” *Data & Knowledge Engineering*, 74 (2011): 1-12, accessed November 26, 2016, doi:10.1016/j.datak.2012.02.005

Reyse, Antonio and Rosso, Paolo and Veale, Tony “A multidimensional approach for detecting irony in Twitter” *Language Resources and Evaluation*, 47 (2013): 239–268, accessed November 26, 2016, doi: 10.1007/s10579-012-9196-x

Svevo, Italo. 1965. *Carteggio con James Joyce, Eugenio Montale, Valéry Larbaud, Benjamin Crémieux, Marie Anne Comnène, Valerio Jahier*, compiler Bruno Maier, Milano: dall’Oglio

Svevo. 1966. *Epistolario*, compiler Bruno Maier, Milano: dall’Oglio.

Quantitative syntactic approaches to Spanish Poetry. A preliminary Study on Fernando de Herrera's Poetic Works

Laura Hernández Lorenzo, University of Seville, lhernandez1@us.es

Distant reading and macroanalysis in the Humanities

Digital Humanities and new digital techniques have led to new approaches in Literature Studies, following the perspectives of "macroanalysis" by Matthew Jockers (Jockers, 2013) and "distant reading" by Franco Moretti (Moretti, 2007). Moreover, these new tools are extremely useful for advances in research thanks to new possibilities of working with data. According to Franco Moretti (Moretti, 2007), distant reading emerges as an opposite to close reading and as a way to bring more objectivity in Literature Studies. Instead of analysing one text or fragment in detail, Moretti defends the use of a more abstract methodology which enables the researcher to work with huge amount of texts in order to see patterns and relations between them. On the other hand, Matthew Jockers claims that "what we have today in terms of literary and textual material and computational power represents a moment in revolution in the way we study the literary record […] large scale text analysis, text mining, 'macroanalysis', offers an important and necessary way of contextualizing our study of individual works of literature" (Jockers, 2013). Although this new methodology is still rejected by some scholars, provoking a polemic debate on newspapers and media –to name a recent example, the publication of "'It's like hitting a painting with a fish'. Can computer analysis tell us something new about literature?" in *The Guardian* last 15th September-, it is a reality as a part of the revolution of Big Data and Digital Humanities.

Quantitative approaches to Literary texts: applying the methodology to Fernando de Herrera's works

This paper, which is part of a broader research project, has the aim of applying quantitative research to one of the most fascinating problems of textual transmission and authorship in Spanish Literature, which affects to the poetic works of Fernando de Herrera (1534-1597). In spite of being one of the greatest poets of Spanish Golden Age, the controversy about his works has provoked more questions and made more difficult to establish Herrera's part in the evolution of the poetic language at the time.

The debate started when the famous writer Francisco de Quevedo noted big and significant differences between the works which The Divine published in life, titled *Some Works* (1582), and those published by his admirer, the painter and poet himself Francisco Pacheco, titled *Verses* (1619). Since then, and especially from the start of the XXth century, academics and experts on Herrera's poetry have discussed about the style and authenticity of *Verses* for decades without reaching to an agreement (Cuevas, 1985). On the one hand, erudite Italian academics such as Oreste Macrí (Macrí, 1972), Salvatore Battaglia (Battaglia, 1954) or Inoria Pepe Sarno (Pepe Sarno, 1982) defended Herrera's authorship and the probability of an evolution of Herrera's style towards Barroque and the poetry of Góngora. On the other hand, recognised Spanish philologists, especially José Manuel Blecua (Blecua, 1958) and Cristóbal Cuevas, totally rejected this possibility and suspect from a possible hand from another writer on the text, presumably Pacheco's. Although some good points have been made, experts recognise the difficulty of getting to definite conclusions without using an approach which allows the researcher work with the entire corpora.

Therefore, close reading has been proven to be insufficient in facing this question whereas there is a chance for quantitative and computational approaches to show not only how they can change our perceptive of Literary History, but also their interest in shedding new light when traditional close reading or Philology can't do it. In addition, quantitative and computational approaches have been applied specially to the study of narrative, but there are less research applying it to poetry, because of the subjectivity of this poetic genre which complicates computational approaches to it. This paper adds research to this field, even more neglected in Spanish Studies, where we would like to highlight the quantitative metrical approach to Spanish Sonnets by Navarro Colorado (Navarro-Colorado, 2015).

First the text was digitalised through Blecua's authorised edition (Herrera, 1975) using OCR software. After that, the text was revised and corrected, maintaining original orthography, because of its interest as another indication of variation between the two books, but creating a copy in which the orthography was standardized to actual Spanish standard orthography. Then the two corpora of Herrera's works were contrasted. This task was realised using our DIY Java software, LitCon, a corpus tool specifically developed to work with poetry.

Our first approach was from the methodologies of corpus and computational Stylistics, using concordances, keywords and function lists, in addition to a new function which contrast words from the two corpus and presents common and different words between them. These results have been presented at international conferences on Corpus Linguistics and Digital Humanities and have underlined some interesting stylistic differences between the two corpora, mainly considerate lexical differences – although 94% of *Some Works*' words appear in *Verses*, more than a half of words in this last corpus do not appear in *Some Works*. Vocabulary in *Some Works* was discovered to be more balance, less violent and hyperbolic in contrast with *Verses*. Moreover, in *Some Works* there is a big presence of the exterior world and the woman, whereas we find in *Verses* much more subjectivity and words related with the speaker's suffering.

After these experiments, an analysis with Part of Speech tagged-corpora is being undertaken. The two corpora in the standardised version have been morphologically tagged using Freeling PoS tagger databases in LitCon (http://nlp.lsi.upc.edu/freeling/node/1). LitCon was programmed to extract words in the corpora and then search them in Freeling databases and morphologically tag them according to it. As Herrera's poetry presents words from XVI century which don't exist in actual Spanish, we use the version for old Spanish of Freeling (Sánchez-Marco, Boleda, & Padró, 2011). Moreover, when a word is not in Freeling databases, Freeling (Padró & Stanilovsky, 2012) has an algorithm which applies the most likely tag, which we then have revised and introduce in a customised database of Freeling. The tagged-corpora is currently being revised and corrected and we will present at the Conference our first results in these approximation.

We will focus specially in noun-adjective against adjective-noun patterns, both in quantitative and qualitative terms to see if they are interesting similarities or differences between the use of morphosyntactic patterns in the undoubted works of Herrera (*Some Works,* 18810 words) and the problematic *Verses* (71457 words). In order to do that, a new function has been developed in Litcon which extracts combination of tags and shows the examples in the text.

This quantitative morphosyntactic analysis will be the previous step to future analysis using Stylometry and non-traditional authorship attribution techniques, giving us preliminary information about morphosyntactic patterns in both texts, with a special focus on Part of Speech relations.

The questions we will try to answer with this study will focus on three aspects: a) Is it possible to observe differences in morphosyntactic patterns between *Some Works* and *Verses*? b) Has really *Verses* a more complex syntax as Literary Criticism has suggested? c) To what extent quantitative approaches can shed new light facing unresolved questions in Literature Studies?

References

Battaglia, S. (1954). Per il testo di Fernando de Herrera. *Filologia Romanza*, (I), 51–88.

Blecua, J. M. (1958). De nuevo sobre los textos poéticos de Herrera. *Boletín de La Real Academia Española*, *XXXVIII*, 377–408.

Cuevas, C. (1985). No Title. In *Poesía castellana original completa* (pp. 87–99). Madrid: Cátedra.

Herrera, F. de. (1975). *Obra poética*. (J. M. Blecua, Ed.). Madrid: Boletín de la Real Academia Española.

Jockers, M. L. (2013). *Macroanalysis: digital method and literary history*. Urbana: University of Illinois Press.

Macrí, O. (1972). *Fernando de Herrera*. Madrid: Gredos.

Moretti, F. (2007). *La literatura vista desde lejos*. Barcelona: Marbot Ediciones.

Navarro-Colorado, B. (2015). A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. *4th Workshop on Computational Linguistics for Literature (CLfL 2015)*, 105–113.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3 . 0 : Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, 2473–2479.

Pepe Sarno, I. (1982). Se non Herrera, chi? Varianti e metamorfosi nei sonetti di F. de H. *Studi Ispanici*, 33–69.

Sánchez-Marco, C., Boleda, G., & Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. *5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, (June), 1–9.

Beyond the tree – a theoretical model of contamination and a software to generate multilingual stemmata

Armin Hoenen, CEDIFOR, Universität Frankfurt, hoenen@em.uni-frankfurt.de

Graph theoretical model for contamination

This article features two different subsections in the realm of stemmatology, where the first is an attempt at accommodating an important phenomenon of stemmatology in graph theoretical terms. We seek to contribute a possible theoretical model for a hitherto theoretically only marginally elaborated phenomenon: contamination. The second part presents a practical counterpart to gain multilingual stemmata by means of a software called **Mu**ltilingual-**Ste**mma-**G**enerator - MuSteG.

In this first part, we want to first focus on contamination in textual criticism and gain a graph theoretical perspective on it. Contamination roughly refers to texts being not transmitted (copied) strictly vertically. With this respect, (Buzzoni et al. 2016) recently showed that the application of certain digital methodology should take into consideration the type of tradition (as should classical non-digital textual criticism). Their points of reference are the terms "open tradition" versus "closed tradition" as coined by (Pasquali 1952), where “closed” largely means uncontaminated in the sense that only one vorlage (vertical transmission) is used in such traditions when compiling a copy, whereas in an open tradition variously contaminated copies (horizontal transmission) complicate their later genealogical assessment. For genealogical modelling however, the most widespread, ideal, simple and typical structure is – the tree. In graph theoretical terms, a stemmatological tree can be defined as a set of nodes symbolizing manuscripts, a set of archs symbolizing copy processes and a distinguished node *root* referring to the authorial original or archetype. Those trees are further characterized by having directed edges (from vorlage to copy) and being acyclic (which indirectly follows from the fact that any node other than root has exactly one parent). Such a model is not capable of representing horizontal transmission, but this does not mean that graph theory doesn’t offer other means of modelling. However, only the fewest publications like that by Flight (1992) have elaborated models other than the tree in philological context, although a vast number of traditions is affected by contamination. Flight presents so-called “Greg graphs” (the name refers to the philologist W.W. Greg, comp. (Greg 1927)), which are graphs that possess labeled nodes representing extant manuscripts and unlabeled nodes representing reconstructed, hypothetical ones, they may contain cycles. Other than in philological context, there are graph types which are likewise conceptually applicable to model stemmatological genealogical relationships with and without contamination: *multiple edge class graphs*, as for instance defined by (Rocklin & Pinar 2013) and *multi layer graphs*, as defined in (Boden et al. 2012), which can be used to represent different qualities of edges or nodes.

Although these models possess a large expressive power and inspired the current paper, here we want to explore yet another approach which importantly maintains the traditional tree while at the same time representing contamination by secondary, ternary, etc. copies. In order to achieve this, the oridnary tree defined as a set $\{V,E,r\}$[1] is expanded by a second set E’ of edges, which are constrained only by their members not being allowed to pertain to E and E’ at the same time. Any edge in E’ is a secondary edge originating only in vorlagen, which have been consulted in case that the primary vorlage has been corrupt. Thus the source node of edges in E is the primary vorlage, the source of edges in E’ are secondary

1 V stands for vertices, a term used interchangeably with the term nodes, E is a set of edges (archs) between two nodes (V x V) and r is the designated root node.

vorlagen and since no manuscript can be vorlage and secondary vorlage at the same time E and E' must nessessarily be completely disjoint.

Such a "contamination tree" preserves the original tree as a subset of its superset {V,E,E',r}. If we model a contamination tree in this way, we can ask, for a tree of n nodes, how many contamination edges are maximally possible. For a tree, the number of edges is, as is well known n-1. Thinking of the possible place to insert the first contamination edge of E', it appears there are n *x* (n-1) (possible source times possible target positions) principal positions for such an edge among the n nodes. However, only older can theoretically contaminate younger edges, which is why n *x* (n-1)/2 would be the appropriate number of insertible edges. Now subtracting the existing n-1 edges of E, we arrive at a formula, which explains at how many places a first contamination edge could be inserted: [n *x* (n-1)/2] - n-1. Now, for any number z of contamination edges to be inserted, the formula for possible placements derives from this formula for the first edge, but is the product of all places for the first with all places for the second …with all places of the z-th edge. Obviously for the second contamination edge, we have the same possible number of places as for the first minus the one place, where the first is located, for the third the same number of possibilities apart from where the first and second are placed and so forth. Doing some basic arithmetics one arrives at a variation of:

$$(1)\ \Pi_{i=1..z}\ [(n^2-3n+2)/2] -(i-1),\ i,z,n \in \mathbb{N},$$

where n is the number of nodes in the tree, z the number of edges to be inserted and i an incremental operator for the ordinality of the to be inserted edge. Ultimately in this way, we will know how many trees including contamination are possible at all by multiplying with the number of all possible trees for n nodes, which may not be as straightforward to achieve with other models of genealogical relationships of manuscripts circumspanning contamination. Given some historical parameters such as the probability of contamination (size of z) one can now arrive at an approximate assessment of how many heriditary processes characterize the tradition at hand confronting various degrees of openness. The model is simplistic in modelling contamination with only one additional edge set E'. Different types of contamination such as exemplar shift, addition of text in the margins and so on could all be modelled by their own more differentiated additional edge sets E'', E''' etc. with restrictions individually defined – depending on each tradition -between those and E (of the type an "exemplar shift edge" and a "contamination edge" cannnot coincide). Only contamination by way of oral transmission and text strata of single manuscripts would then require additional theoretical modelling (for instance through secondary node sets or the use of labels).

Translation and a software for multi-lingual stemmatology (MuSteG)

In antiquity and the middle ages, translation was another important process rarely theoretically modelled in the context of textual criticism. Translating a work or author of renown may have at times been economically especially feasible and thus widely applied for under more two reasons. Firstly, one would not have to establish prestige and attraction of a work oneself and secondly, although being a creative process, translation may often take less time than the composition of a text ex novo. Hence, any serious theory of stemmatology should be able to accommodate translation as a process, also because for texts such as the Bible, translation was an indispensable vehicle. Unsurprisingly, accommodations of translation are not innumerous in classical stemmatology (practically though not necessarily theoretically) and manifest themselves in various stemmata in articles or prefaces/appendices of editions, such as in (Vancamp 2010). This article proposes a model of a multilingual stemmatic forest, which functions not unlike the above model of contamination. Each tradition possesses one or many genuine

trees in each language, which may or may not coincide with redactions. Whenever any of the manuscripts is translated, such a manuscript is termed "seed" and the resulting translation becomes root of a new tree in another language. It is needless to say, that by this token processes of contamination become much more (one is almost tempted to say infinitely more) complex than in a monolingual scenario.

Putting this model into practice, we elaborate a software which will produce such multilingual stemmatic forests. The principle problem of multilingual stemmatology is, that texts of different languages cannot be compared word by word, letter by letter, since their linguistic appearance is different in both languages through formal, linguistically required deviations (of the languages rather than the texts). Word order often changes, certain function words or morphemes may be added or deleted or transformed and the words of the two languages even for very closely related varieties are not likely to be all cognate. For a sentence in the source language more than one appropriate translation may exist while any sentence copied within the same language can have only one single correct unaltered copy form. Thus aligning corresponding elements between two languages is not straightforward, which makes it extremely difficult and subtle to judge on genealogically informative deviations departing from word and letter levels.

However, instead of comparing words and letters, one can use other more general and language independent levels of information such as the sequences of chapters. We use this information in comparing manuscripts in various languages of the work *Physiologus,* a medieval Christian compendium of animals, the original of which was written in Ancient Greek. First, we extracted 25 chapter sequences each representing one single manuscript from editions, faximiles and the literature (Carmody 1953, Frank 1971, Gippert 1995, Lauchert 1889, Mirandola 2001, Sbordone 1936). The sequences were encoded in a way, that the chapter on the same animal always was represented by the same number. Hence, "wale – antelope – onager" and "antelope – onager –wale" could be encoded as 1 -2-3 and 2-3-1. Marking the start and end, we inserted two symbols at the beginning and end of each sequence. Then we computed the bigram sequence similarity between any pair of manuscripts regardless of their language. Bigram sequence similarity is simply the number of shared bigrams divided by the number of bigrams present in both manuscripts. This lead to a pairwise distance matrix (1-similarity). This matrix may immediately be used with phylogenetic software (for instance one could use R and the algorithm Neighbour Joining (NJ) as implemented in the package 'ape') to obtain a phylogentic tree for all manuscripts in all languages. A similar methodology has been used monolingually for chapters of Chaucer (*Canterbury Tales*) in (Spencer et al. 2003).

However, here, we present an approach, which in our view is adapted for multilingual transmission since it allows the incorporation of philological rules depending on the tradition. In order to obtain a multilingual stemmatic forest as described above, a software has been implemented in the Java programming language, which produces editable LaTeX code. The approach can be characterized as iterative clustering, which is not hierarchical, but produces multifurcations.

Starting from the pairwise matrix of ditances, we cluster together all manuscripts, which share a certain percentage of bigrams (for instance 50%, which is then the distance threshold that has to be set by the user in advance). As in phylogenetic networks there can be incompatible groups resulting from this cluster step. This is the case if for instance a manuscript M1 deviates less than 50% from a manuscript M2, but more from a manuscript M3, which itself however is less than 50% different from M2. In that case, we follow a similar strategy as (Huson 2002) and merge incompatible groups (M1,M2 and M3 become one group). The next step is to apply philological knowledge on the tradition involved. If in a clustered group for the Physiologus for instance, there is only one Ancient Greek manuscript, it will be considered the cluster root. If in a group there is no Ancient Greek but only one Latin manuscript, the latter will be considered root. In other cases, the algorithm postulates a lost Ancient Greek root for the group. In this way, extant manuscripts can end up as internodes or even root. Additionally, with rules such as this, transmission specific knowledge can be used during stemma generation and more importantly varied among different traditions. The third step consists of updating the matrix, which is a

step similar to the concurrent step in NJ. First, all clusters/groups are identified, named and a new pairwise distance matrix between them is set up. Each cluster is solely represented by its root. All pairwise distances are then retrieved from the old matrix (last step). If for instance for one of the groups a new root has been postulated, the average of distances of all group members with all of the second group in the pair is considered. If an extant manuscript has been chosen as root, its previous distances remain compulsory. After this step, the next round of clustering will be executed. The process terminates when after a clusterstep either there is only one group left comprising all manuscripts or when there remain only groups with one single member, all of which would be connected to a hypothetical root node.

The MuSteG software produces LaTeX source code displaying the result of the iterative clustering, when converted to pdf. See Figure 1 for illustration. Any edge entailing a shift of language points to a seed in the theoretical sense explained above. The forest here comprises thus 22 trees, testimony to the importance and spread of the Physiologus. Obviously, these simple rules produce a forest, which is not completely in line with results of traditionally hypothesized relationships, but carries remarkable structural similarities. Contamination can be inserted by the editorially informed user in the LaTeX source code, which corresponds to the structure imposed by the library 'forest'. MuSteG is available as a Java applet from the author, where the user can specify the location of the input sequences, his/her own cluster root resolution rules and the cluster distance threshold.

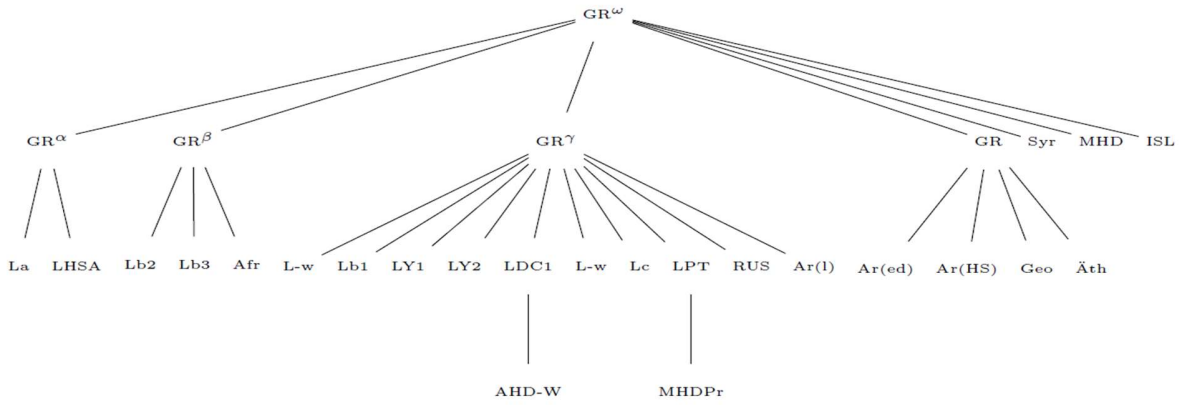


Figure 1: Multilingual stemmatic forest for 25 sequences of the Physiologus. Hypothetical manuscripts marked by lower case Greek letters in superscript. Afr = Old French, AHD = Old High German, Ar = Armenian, Geo = Georgian, GR= Ancient Greek, ISL = Islandic, L = Latin, MHD = Middle High German, RUS = Russian, Syr = Syriac, Äth = Ethiopian

Riferimenti Bibliografici

Boden, B., Günnemann, S., Hoffmann, H., & Seidl, T. "Mining coherent subgraphs in multi-layer graphs with edge labels. " *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (2012): 1258-1266.

Buzzoni, M., Burgio, E., Modena, M., & Simion, S. "Open versus Closed Recensions (Pasquali): Pros and Cons of Some Methods for Computer-Assisted Stemmatology." *Digital Scholarship in the Humanities* 31(3) (2016), 652-669.
Carmody, F. J. *Physiologus: The Very Ancient Book of Beasts, Plants, and Stones; Translated from Greek and Other Languages*. Book Club of California, 1953.
Flight, C. "Stemmatic Theory and the Analysis of Complicated Traditions." *Manuscripta* 36(1) (1992), 37-52.
Frank, L. *Die Physiologus-Literatur des englischen Mittelalters und die Tradition*. Fotoschnelldruck R. Köhler, 1971.
Gippert, J. "TITUS. Das Projekt eines indogermanistischen Thesaurus" ("TITUS. The project of an Indo-European thesaurus") *LDV-Forum* 12/2 (1995), 35-47.
Greg, W. W. *The calculus of variants: an essay on textual criticism*. Oxford: Clarendon Press, 1927.
Huson, D. H., Rupp, R., and Scornavacca, C. *Phylogenetic networks*. Cambridge University Press, 2010.
Lauchert, F. *Geschichte des Physiologus*. Slatkine reprints, 1974 (Orig. 1889)
Mirandola, L. *Chimere divine. Storia del Fisiologo tra mondo latino e slavo*, CLUEB, 2001.
Pasquali, G. *Storia della tradizione e critica del testo*. Firenze: Le Monnier, 1952.
Rocklin, M., & Pinar, A. "On clustering on graphs with multiple edge types. " *Internet Mathematics*, 9(1) (2013), 82-112.
Saitou, N., & Nei, M. "The neighbor-joining method: a new method for reconstructing phylogenetic trees. " *Molecular biology and evolution*, *4*(4) (1987), 406-425.
Sbordone, F. *Physiologus*. Georg Olms Verlag, 1936.
Spencer, M., Bordalejo, B. Wang, L-S., Barbrook, A.C., Mooney, L.R., Robinson, P., Warnow, T and Howe, C.J.. "Analyzing the order of items in manuscripts of The Canterbury Tales." *Computers and the Humanities* 37, no. 1 (2003): 97-109.
Vancamp, B. *Untersuchungen zur handschriftlichen Überlieferung von Platons" Menon".* Franz Steiner Verlag, 2010.

Finding reasons for modifications in historical manuscripts

David Lassner, TU Berlin, david.lassner@campus.tu-berlin.de

Problem

The idea that writing makes its way from the authors first draft manuscript to the intended reader without any detours or modifications is often inaccurate and oversimplified. In general, the author or a close person performs corrections and stylistic modifications in subsequent iterations. Additionally, there may be an editor or even an official censor who perform censorship of too private or too extreme parts of the document. The different versions of a document generated by these correction layers often become intransparent in printed versions of the document, while manuscripts are more likely to display traces of how the document has been modified to its current state. The digital scholarly edition "Letters and texts. Intellectual Berlin around 1800" (Baillot 2016, IB in the following) combines genetic edition and entity annotation. The corpus encompasses literary and scholarly testimonies by a group of people, who influenced the intellectual Berlin between Enlightenment and Romanticism. The genetic encoding gives precise information regarding deletions and additions in the manuscript text. However, the reason for these modifications is not encoded. Three main domains for reasons why to modify such a document as a letter in the intellectual context of the time around 1800 have been identified:

1. Correction of mistakes
2. Stylistic modification
3. Moral censorship based on the topic

This paper proposes an unsupervised machine learning approach, which assigns the according reason to every modification. The method focuses on dealing with content related modifications (2. and 3.) assuming that corrections of mistakes (1.) have been filtered out beforehand. It is thus a specialized extension to topic models for dealing with modifications in historical manuscripts and can be employed to automatically classify the modifications by reasons. In practice this could, apart from analyses on unknown historical data sets, be used for evaluation of expert classification. Ultimately I am aiming to increase the accessibility to manuscripts by providing a structure for the modifications which gives additional information and facilitates entry for a broader readership. Furthermore the proposed method may be applied on different editorial problems, which I will discuss in the Outlook section.

Method

As brought up in the Problem section, the proposed method focuses on stylistic and moral censorship reasons, based on the assumption that these two types of reasons relate to the topic of the modification. I convey a generative topic model, that is based on Latent Dirichlet Allocation (D. Blei, Ng, and Jordan 2003) and is able to take into account the structural information of modifications. There exists a wide range of topic models that customize LDA and many of these take into account additional structural information. To replace the Bag-of-words approach by introducing structural information about the word order is a major field of LDA research (Gruber, Rosen-Zvi, and Weiss 2007; Wallach 2006). Moreover there exists a lot of research on topic hierarchies (D. M. Blei, Griffiths, and Jordan 2010; Paisley et al.

2015). LDA has also been modified to work with graph-structured documents (Xuan et al. 2015). However I am not aware of any literature that shows how to model modification reasons in a corpus of natural language.

Figure 1 illustrates the conceptual functioning of the method from left to right. As input on the left, a collection of documents is given. The documents have parts marked as modified. The generative model in the center infers reasons by taking into account all text, inside and outside the modifications. Every reason may stand for a stylistic, or a certain moral censorship reason (e.g. political, religious). On the right side, the model outputs a reason-modification assignment.

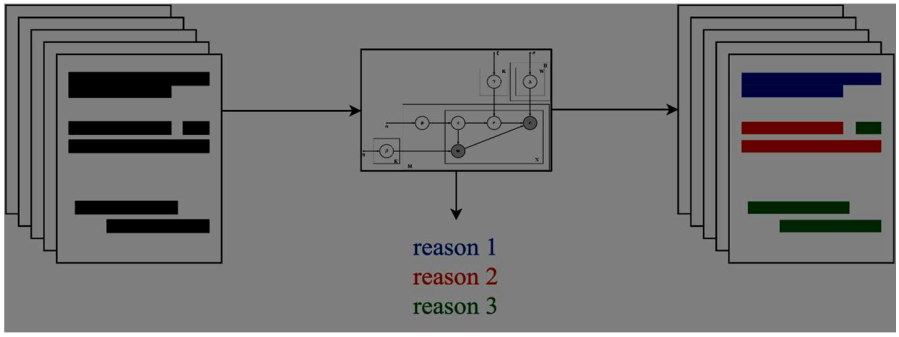


Figure 1: The generative model in the center receives input documents with modifications. It outputs reasons for modification and a reason assignment to each modification.

In addition to the LDA latent variables I introduce a topic-reason variable γ, a word-reason-modification tendency λ and a token-reason assignment *r*. The complete model in plate notation is shown in Figure 2. *c* (observed) models whether a token has been modified. For every topic, γ holds a distribution over reasons, which may cause a modification. For most modifications this distribution should be sparse, for example if a censor crosses out a sentence that discusses the financial situation of the author, the topic and the reason for censorship would be identical. On the contrary a stylistic modification wouldn't always have one or two clear corresponding topics.

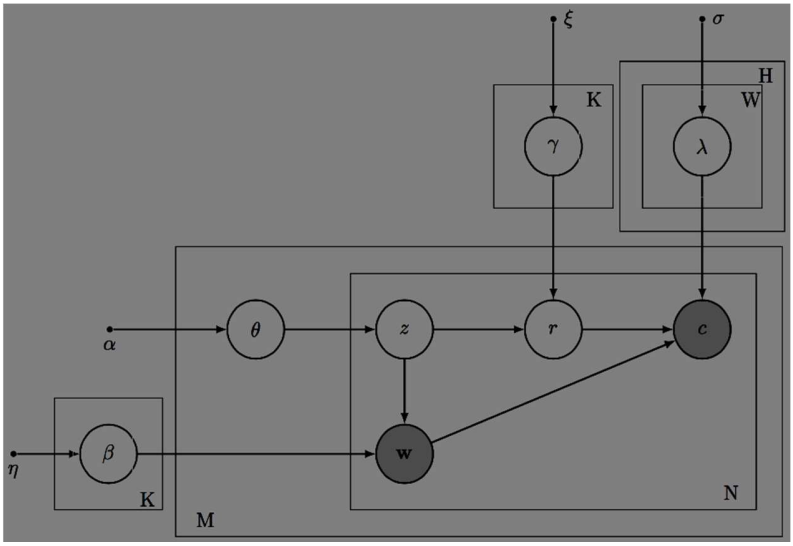


Figure 2: Plate notation of the model. The left four circled variables represent LDA, the right ones the modification part.

For every token and every reason, λ holds a distribution over two states, namely whether the token tends to be modified for this reason. There may be token, that are representative for a topic, but they nonetheless do not tend to be modified. The categorical variable *r* represents the reason assignment at that position.

The latent variables can be iteratively approximated using Variational Inference (Bishop 2006; Zhao 2013).

Intermediate results on toy data

To evaluate the characteristics of this method experiments with artificial toy data can be employed. As an example we show a series of experiments on how sensitive the model reacts to different sizes of the data set. For the three latent variables γ, β and λ we tried to estimate the true value by keeping most of the other latent variables fixed.

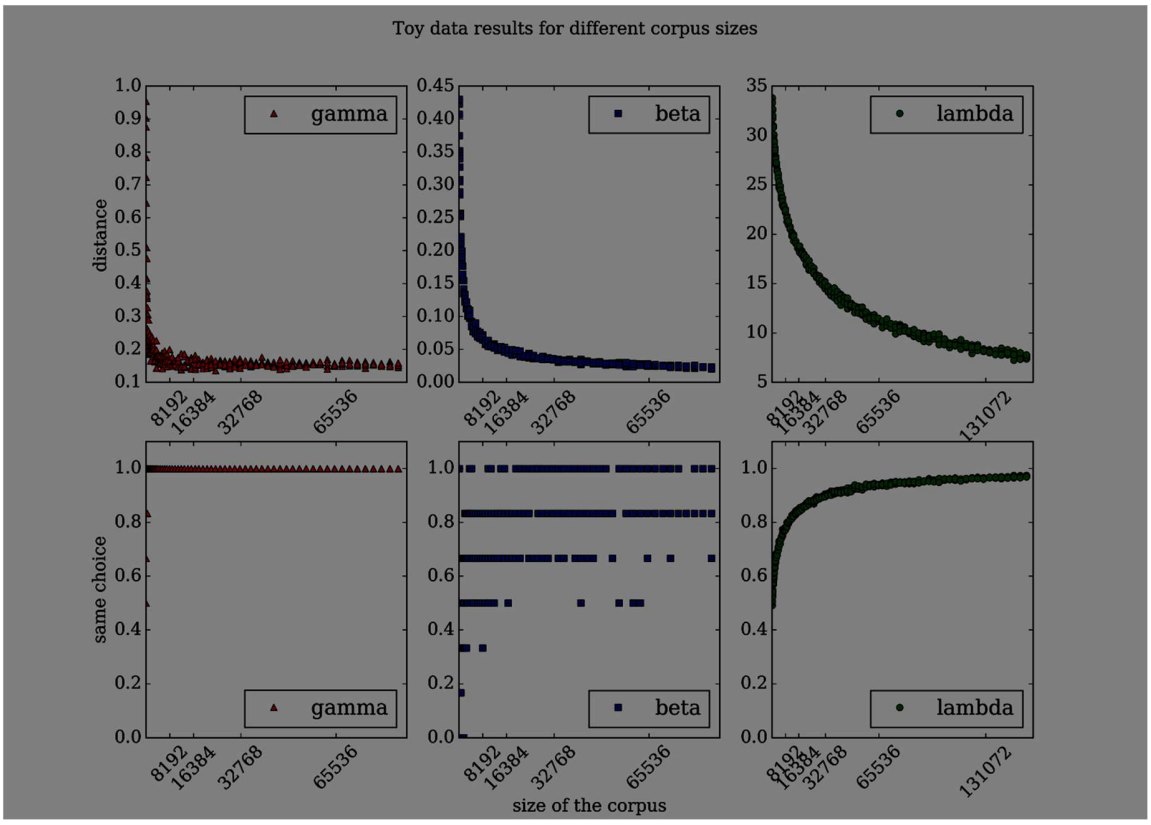


Figure 3: Accuracy of prediction for different corpus sizes. Number of words in dictionary (W): 420, number of topics (K): 6, number of reasons (H): 6.

The first row of panels in Figure 3 shows the distance between the latent variable and the expected value of its variational parameter. The second row of panels shows the ratio of correctly identified maxima in the expected values of the variational parameters. For example if the expected value of the variational parameter of λ would lean to modify exactly the same words as the ground truth, the value

would be equal to one. That means that in the first row a smaller value is better, in the second row however a value close to one is desired.

Since λ has the largest number of degrees of freedom it converges slowest - contrary to γ, which has the smallest number of degrees of freedom. Note that the rightmost column even shows a larger range of corpus sizes for λ.

Intermediate results on IB data set

To apply this method on the IB data set, some preprocessing steps are necessary. Apart from standard natural language preprocessing, one has to filter out all corrections of mistakes.

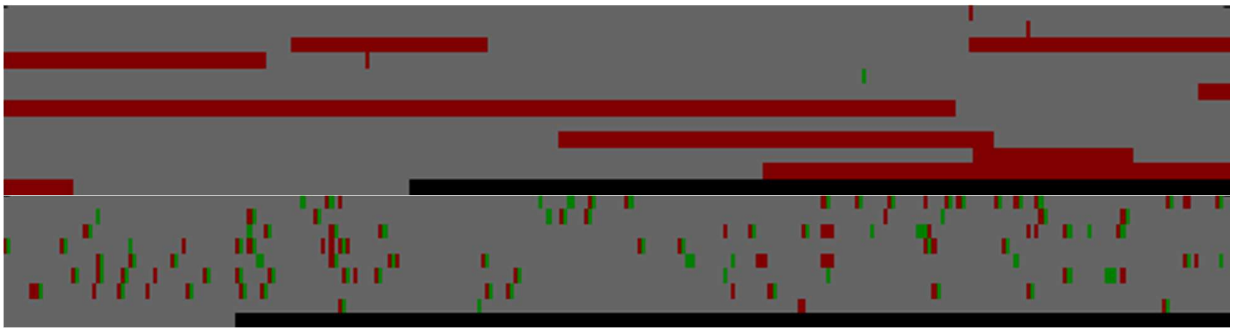


Figure 4: First panel: Letter 4, Chamisso to de La Foye contains small corrections. Below: Letter 14, Dorothea Tieck to Uechtritz contains larger modifications. Deletions (red), additions (green).

The visualization in Figure 4 reveals a great variety in the structure of the modifications. The figure shows the state of all tokens of two letters from the IB data set. The upper letter contains a lot of small changes, where often a green (added part) and red (deleted part) occur as a combination. The letter below contains a lot of longer deleted parts, concluding, that the letter above contains corrections of mistakes, whereby the lower contains modifications related to the topic. These insights lead to a method for distinction between corrections of mistakes and content related modifications based on the size, distribution and co-occurrence with additions.

Outlook

In the near future, I will undertake more experiments with toy data and move swiftly to applying the method on the IB data set. To do so, I will incrementally increase the number of modification reasons. The results will be made accessible as part of the IB corpus, making the permeability between editorial and algorithmic work more visible and accessible to all interested DH communities for reuse.
In a further step, I would like to look into different applications of this method. A promising idea would be to look into different editions of the same text and consider each difference as a modification.

Bibliographic References

Baillot, Anne, ed. 2016. *Letters and Texts. Intellectual Berlin around 1800*. Berlin: Humboldt-Universität zu Berlin. http://www.berliner-intellektuelle.eu/. Please visit the web page for an up-to-date version.
Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer Science+Business Media, LLC.
Blei, David M., Thomas L. Griffiths, and Michael I. Jordan. 2010. “The Nested Chinese Restaurant

Process and Bayesian Nonparametric Inference of Topic Hierarchies." *J. ACM* 57 (2): 1–30.
Blei, David, Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation." *JMLR*, January.
Gruber, Amit, Michael Rosen-Zvi, and Yair Weiss. 2007. "Hidden Topic Markov Models." In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. JMLR.
Paisley, John, Chong Wang, David M Blei, and Michael I Jordan. 2015. "Nested Hierarchical Dirichlet Processes." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2).
Wallach, Hanna. 2006. "Topic Modeling: Beyond Bag-of-Words." In *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York: ACM.
Xuan, Junyu, Jie Lu, Guangquan Zhang, and Xiangfeng Luo. 2015. "Topic Model for Graph Mining." *IEEE Transactions on Cybernetics* 45 (12).
Zhao, Wayne Xin. 2013. *Varitional Methods for Latent Dirichlet Allocation*. http://net.pku.edu.cn/~zhaoxin/vEMLDA.pdf (accessed on 27th of November 2016).

Global Citizem Scholars: Mobilizing the public to assist with paleographic transcription of medieval manuscripts using Massive Open Online Couses (MOOCs)

Roger Louis Martínez-Dávila, CONEX Marie Curie Fellow,
Universidad Carlos III de Madrid (España)
University of Colorado-Colorado Springs (USA)

Overview

Global Citizen Scholars is a Massive Open Online Course (MOOC)-based global initiative that motivates a public from over 140 nations to learn about medieval European interreligious history, acquire 13th-19th century paleography expertise, and transcribe original Castilian manuscripts to expedite original scholarly research. In this short paper, I highlight the process, methods, and outcomes of this citizen science, or crowdsourcing, teaching and research endeavor.

Using the coursera.org and edX.org MOOC platforms, I have reached a global audience of approximately 18,000 students in three iterations of my series of *Deciphering Secrets* MOOCs (www.decipheringsecrets.com) from Summer 2014 through Summer 2016. Presently, I am expanding my MOOC offerings via a 3-year Marie Curie Fellowship (2015-2018) to include manuscript investigations for four medieval and early modern Spanish cities. My students participated in 7- and 12-week MOOCS that investigated Jewish, Christian, and Muslim interrelationships in 14th and 15th century Spain. Using an original paleographic method that I developed for online audiences, known as *SILReST*, students rapidly acquired essential character recognition and transcription skills so that could perform original manuscript analyses. The results from the project are encouraging as it may demonstrate a new way for scholars to support the public's deep desire to learn more about their cultural history, as well as to facilitate the rapid advancement of original manuscript studies.

Two MOOC Case Studies

For example, consider the results from my summer 2014 coursera.org MOOC, *Deciphering Secrets: Unlocking the Manuscripts of Medieval Spain*. Over 10,600 students from 143 countries enrolled in the course and over 2,500 students completed it (or 25% of the enrolled students). Approximately 75% held a bachelor's or advanced degree, 52% were women and 48% men, and all age groups (18-60+) were almost equally represented in the class. While a 25% course completion rate may seem relatively low for a 12-week course, its impact cannot be underestimated when put into relationship to my entire teaching career. For example, during my 12-year on-campus teaching career I calculated that I have instructed 1,000 students. Thus, I reached approximately 83 students per year. However, via the *MOOC* and in just over 3 months, I was able to teach 2,500 students – or the equivalent of 30 years of on-campus teaching.

In terms of teaching paleography in an online setting, I developed a course that trained mostly non-Spanish speakers to transcribe Spanish-language documents. Via indirect demographic methods (ethnicity and international location of student participants), I estimated that 75% of my students had limited or no-Spanish language knowledge. However, we determined that on average, our students could identify approximately 90% of the Spanish letters, numerals, and abbreviations they viewed on a manuscript page. The primary caveat of this finding is that students were interpreting an 1807

transcription of an original early 15th century manuscript. Clearly, this script is simpler to read, yet, in principle our method proved effective.

The direct research benefit of teaching paleography to students was that I enlisted them in a crowdsourcing initiative that involved transcribing the 19th century copy of the 600+ page *Book One (1399-1453)* of the *Capitulary Acts* of the Cathedral of Plasencia (Spain). Over the course of 3 weeks my students transcribed at least 90% of the entire text. Moreover, using empirical data from individual student paleographic quiz scores and student peer-evaluation of each transcription, I can assign "accuracy and reliability" scores to each student transcription. In this way, I can electronically reassemble *Book One* using the most reliable transcriptions. From these tangible results, I expect: (1) to transfer the manuscript data into our project databases for our online-3D world *Virtual Plasencia* (http://revealingcooperationandconflict.com), (2) to write traditional scholar studies, and (3) to prepare a critical edition of *Book One*.

More recently, during summer 2016, I delivered a new 7-week MOOC on the interreligious history of the critical Castilian royal city of Burgos, titled, *Deciphering Secrets: Unlocking the Manuscripts of Medieval Burgos (Spain)*. See: https://www.edx.org/course/deciphering-secrets-unlocking-uc3mx-hga-2-1x-0. This MOOC was targeted to a very specific audience of students intrigued by Jewish, Christian, and recent converts to Christianity. These students also understood they would attempt to learn very challenging medieval Spanish paleography. Even with limited time to advertise the course, over 1,659 students from 95 nations participated in the course.

What was most exciting about the course is that it demonstrated that everyday students performed high-level analysis (transcription) of very challenging medieval manuscripts from the 15th century. Most historians have difficulty with this transcription work, but our students understood 65% to 75% of the medieval handwriting they studied (quantitatively evaluated). This is exceptionally good and it shows that students of all types of backgrounds and all ages can perform this scholarly research.

Further, the students transcribed eight original 15th century documents as a group (cohort), which can now be gathered for a new book (collection) of historical manuscripts that pertain to Jewish, Christian, and Muslim interrelations. Over the course of the next two years, we will have the opportunity to transcribe a large number of new manuscripts -- hopefully several hundred pages of documents – as the course continues to repeat on edX.org.

Research and Pedagogical Objectives

Essentially, my research and pedagogical goals center on three fundamental issues – (1) how to advance humanistic scholarship in the most expeditious manner given limited human resources, (2) how to engage the broader public in central questions in the humanities (i.e., the nature of interreligious co-existence) as a means of advancing the importance of the humanities in public life, and (3) how to harness the benefits of rapidly evolving connective technologies (i.e. Internet, online courses, etc.). Thus, my intellectual and activist intent blends multiple goals that I believe can be best realized through teaching online MOOCs that attract a large, international student body.

Methodological Overview

The methodological approach that I have conceptualized for *Global Citizen Scholars* and teaching paleography is straightforward – at the most basic level handwriting scripts are symbols and humans are excellent symbol-recognition beings. Put simply, my approach argues that just about every human can identify, for example, different Mayan or Egyptian glyphs even though they cannot understand their meaning. Of course, I am concerned about student comprehension, so if a student understands Spanish and therefore what they transcribe, this is a bonus for the student. Additionally, I develop video course

lectures that help students contextualize the types of content that they encounter in the manuscripts they are transcribing.

My approach is not a cynical one that uses humans as machines; rather, it is fundamentally about engaging the broader public in humanistic inquiry. Not every member of the public has the privilege of pursuing scholarly work, however, every person has a right to be intellectually curious. Moreover, every human has the capability to recognize and interpret human symbolic language and they can channel this innate human skill to participate in the advancement of scholarship. The specific manner I teach MOOC students paleography involves the production of two to three weeks of video course lectures on manuscripts and paleography, the development of a customized paleographic manual with practice exercises, the creation of machine-graded paleography quizzes to assess student achievement, and peer-evaluated transcription assignments (using standardized rubrics and online discussion forums).

SILReST (Teaching Paleography to the General Public)

Using the online toolsets that are typically incorporated into MOOC course management systems (CMS), I can teach basic paleography to MOOC students in only two weeks. My approach to paleography is called SILReST. It is an initialism, or what you might think of as an acronym. Each letter in SILRsST represents one of six strategies. Briefly, these strategies are:

S. Strategy #1 - Scan the entire document before attempting to transcribe it. It is important for you to become familiar with how the scribe writes. Repeatedly scanning a document will help you eyes to become accustomed to the "hand" of the scribe.

I. Strategy #2 - Identify those letters, abbreviations, and numbers that you can immediately recognize. This is very straightforward, but it is the beginning of finding your way into the document. Finding easy-to-recognize letters will help appreciate how much you can already see and it will spur you along to uncover other letters and words.

L. Strategy #3 - Locate common words to (a) understand how the scribe connects their letters together and (b) recognize other alphabetical letters and numbers. This strategy helps you identify letters that are hard to recognize. If you see a common word, and you are flexible in terms of how it might be spelled, then you see many curious spellings of words you know. More importantly, you can find new letters using this strategy.

R. Strategy #4 - Recognize the abbreviations used in the document and if they vary within the document. Finding and marking your abbreviations makes your task easier because it reminds you some words on the page are not complete words at all. Rather, they are almost nonsensical connections of letters. Find the abbreviations so that your eyes and mind do not attempt to create words that do not exist on the page.

S. Strategy #5 - Search for English-Spanish cognates (those words that share similar meanings and spellings in English and Spanish) to identify more letters and connections. Cognates are helpful because you can work "backwards" into reading letters on the page. For example, if you know the word might be "jurisdiction" in English and therefore is "jurisdicion" in Spanish, then you can begin to identify hard to read letters within the word on the page.

T. Strategy #6 - Type or write your transcription and leave plenty of room to add edits. Creating a transcript will help you fill in the blanks as you work through those last, hard to read letters and words.

My SILReST method is now integrated into each of my upcoming MOOCs. These MOOCs often attract former students to re-enroll and thus their paleography proficiency is likely to improve with each course.

Conclusions

Although the MOOC-model of open-access education and research is certain to change, *Global Citizen Scholars* demonstrates there is a broad Internet audience that is interested in contributing to original scholarly research.

This project has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nº 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509) and Banco Santander.

Bibliographic References

Causer, Tim and Valerie Wallace. "Building A Volunteer Community: Results and Findings from *Transcribe Bentham*." *Digital Humanities Quarterly* 6:2 (2012): 1-84.

Manzo, Christina et. al. "By the People, For the People: Assessing the Value of Crowdsourced, User-Generated Metadata." *Digital Humanities Quarterly* 9:1 (2015): 1-27.

Poole, Alex H. "Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities." *Digital Humanities Quarterly* 7:2 (2013): 1-72.

Shillingsburg, Peter. "Reliable social scholarly editing." *Digital Scholarship in the Humanities* 21 (Sep 2016): fqw044; DOI: 10.1093/llc/fqw044

Van Zundert, Joris J. "The Case of the Bold Button: Social Shaping of Technology and the Digital Scholarly Edition." *Digital Scholarship in the Humanities* 7 (Mar 2016): fqw012; DOI: 10.1093/llc/fqw012

Annotazione tematica assistita vs topic modeling: un confronto sui testi poetici delle origini

Daniele Silvi, Università di Roma 'Tor Vergata', silvi@lettere.uniroma2.it
Fabio Ciotti, Università di Roma 'Tor Vergata', fabio.ciotti@uniroma2.it

Introduzione

L'annotazione tematica in ambito letterario (Segre 1985; Sollors 1993) è comunemente considerata una operazione critica che, anche se supportata da tecnologie informatiche e computazionali (Ciotti 2014), va essenzialmente eseguita in modo "manuale" poiché richiede il sostanziale intervento della capacità critica ed ermeneutica, nonché della competenza storico-letteraria, del ricercatore che la esegue. Recentemente tuttavia, nell'ambito del cosiddetto paradigma del *distant reading* (Moretti 2013b) si sono ampiamente diffuse alcune tecniche di analisi statistico/probabilistica che cercano di operazionalizzare (nel senso di Moretti 2013a) la fase di individuazione dei cluster tematici presenti in un corpus testuale. In particolare ci riferiamo alle tecniche di *topic modeling*, ovvero l'estrazione dei cluster lessicali che caratterizzano un insieme di testi, e l'analisi delle loro distribuzioni (Underwood, 2012a; Blei, 2013).

Il nostro paper si propone di effettuare un confronto tra i risultati della marcatura tematica effettuata nel contesto di un approccio computazionale semi-automatico e quelli prodotti dall'applicazione al medesimo corpus dell'algoritmo di topic modelling più diffuso LDA (Blei 2012). Il fine della sperimentazione (che dettaglieremo con tabelle e grafici) è sia quello di verificare l'efficacia, o meglio il grado di convergenza o divergenza tra i metodi automatici e quelli *human-centred*, sia quello di arricchire l'analisi tematica del corpus di riferimento, i testi poetici delle origini in lingua volgare.

L'annotazione tematica manuale

Il lavoro che presentiamo si inserisce nell'ambito del contributo apportato dall'unità di ricerca dell'Università di Roma "Tor Vergata" al progetto di ricerca PRIN "Memoria poetica e poesia della memoria. Ricorrenze lessicali e tematiche nella versificazione epigrafica e nel sistema letterario", coordinato dall'Università Cà Foscari di Venezia, il cui obiettivo era la annotazione tematica digitale della tradizione epigrafica latina (Pistellato, 2015).

Il lavoro condotto dalla nostra unità ha riguardato l'estensione dell'annotazione tematica alle forme brevi della poesia italiana delle "Origini" in lingua volgare (considerata anche a partire dai suoi labili e mobili confini con la coeva produzione latina). Nella prima fase del progetto sono stati reperiti, digitalizzati e tematizzati tutti i testi poetici di forma breve (intendendo con questa come estensione massima la forma "sonetto") del XIII secolo. Su un corpus di più di 60 autori del '200 e un totale di circa 500 testi, sono stati individuati oltre 500 temi, di cui un solido 30% con corpose attestazioni. Tutti questi dati saranno presentati con tabelle e grafici riassuntivi (di cui in questa sede forniamo un esempio).

L'operazione di annotazione dei testi inseriti è stata effettuata mediante uno strumento di annotazione che permetteva di assegnare uno o termini tematici a un verso, una strofa o un intero componimento, sulla base di un thesaurus tematico predefinito (mutuato dal progetto gemello – e precedente- Musisque Deoque), ma estensibile su suggerimento degli annotatori.

Il processo di annotazione è stato condotto da una equipe di esperti, coordinato da Daniele Silvi. La creazione di un gruppo di lavoro stabile si è resa necessaria per garantire un certo equilibrio sul fronte interpretativo di questa operazione. Ricordando infatti come ogni atto di annotazione rappresenti un atto interpretativo, il rischio di incorrere nella non omogeneità derivante dalla soggettività

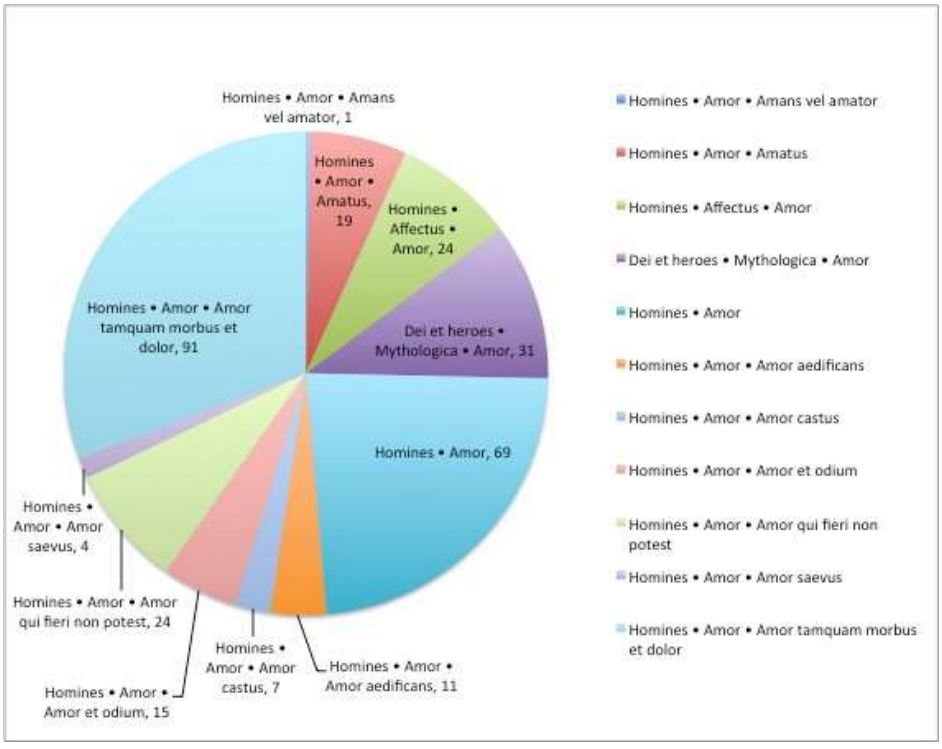
dell'esperienza culturale oltre che della percezione del testo, è stato parzialmente mitigato proprio dal confronto diretto tra i diversi membri del gruppo di lavoro.

I testi sono stati dunque suddivisi tra i membri ed analizzati a coppie; poi, ciascun testo annotato è stato revisionato dal coordinatore campionando una selezione di testi. Questa procedura si è resa necessaria al fine di garantire la maggiore uniformità di giudizio possibile all'interno della rosa di temi messi a disposizione dal temario predefinito adottato.

Dato il numero consistente dei testi e dei temi a essi associati che abbiamo elaborato, forniamo qui solo un esempio tra i più significativi. Si tratta della categoria che riguarda le diverse forme dell'amore che vengono descritte nella tradizione testuale:

- Homines • Amor • Amans vel amator
- Homines • Amor • Amatus
- Homines • Affectus • Amor
- Dei et heroes • Mythologica • Amor
- Homines • Amor
- Homines • Amor • Amor aedificans
- Homines • Amor • Amor castus
- Homines • Amor • Amor et odium
- Homines • Amor • Amor qui fieri non potest
- Homines • Amor • Amor saevus
- Homines • Amor • Amor tamquam morbus et dolor

I risultati dell'annotazione manuale, relativamente al corpus sopradescritto, sono riassunti nel seguente grafico, che ne indica la distribuzione.



L’analisi automatica con topic modeling

Lo stesso corpus di testi è stato sottoposto a un processo di analisi automatica basata su *topic modeling*. Esistono diversi algoritmi di topic modeling, ma ad oggi il più diffuso è quello noto come *Latent Dirichlet Allocation* (LDA), fondato su un approccio probabilistico bayesiano. In modo intuitivo possiamo dire che alla base di LDA vi è un semplicistico modello generativo del testo: quando un autore scrive un testo in prima battuta sceglie l’insieme degli argomenti (topic) di cui vuole parlare, poi determina la proporzione con cui ciascun argomento sarà presente e infine per ogni argomento mette nel testo un certo numero di parole che esprimono ciascun argomento. La cosa interessante di questo semplice metodo modello generativo è che può essere invertito: otteniamo in questo modo un algoritmo che è in grado di estrapolare i topic prevalenti in un insieme di documenti, ovvero la lista delle parole che co-occorrono con frequenza notevole e la loro distribuzione di probabilità. In termini tecnici si dice che in LDA un testo è una distribuzione di probabilità su un insieme di topic e un topic una distribuzione di probabilità su un insieme di parole.

Le liste ottenute vanno ovviamente interpretate, assegnando loro una determinata etichetta o categorizzazione. Nel nostro caso i testi del corpus sono stati sottoposti a un *tool* che implementa l’algoritmo LDA (Mallet, cfr. McCallum, A. K. 2002) seguendo diversi approcci di segmentazione e producendo diversi numeri di topic. Si è poi proceduto a verificare:

- la congruenza dei cluster ottenuti tramite LDA con l’albero del thesaurus
- la congruenza delle attribuzioni tematiche degli esperti con quelle individuate mediante LDA
- la possibilità di usare i termini del thesaurus per etichettare i diversi *topic* prodotti dall'algoritmo

I risultati della ricerca, che saranno esposti nella presentazione, sono piuttosto contraddittori, e sembrano indicare la scarsa efficacia degli algoritmo di topic modeling (e in particolare di LDA) per una applicazione tematologica in ambito letterario, come già suggerito da Ciotti (2016), sebbene ci si possano aspettare sviluppi interessanti da alcune varianti di LDA (in particolare Labeled LDA, Ramage at al 2009).

Riferimenti Bibliografici

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55.4: 77–84. Blei, D. (2013). Topic modeling and digital humanities. Journal of Digital Humanities, 2.1.

Ciotti F. (2014). Tematologia e metodi digitali: dal markup alle ontologie. In *I cantieri dell'Italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo. Atti del XVII congresso dell'ADI – Associazione degli Italianisti (Roma Sapienza, 18-21 settembre 2013)*, ed. Beatrice Alfonzetti. Roma: Adi editore.

Ciotti, F. (2016). What's in a Topic Model. I fondamenti del text mining negli studi letterari. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 149-151.

Dennett, D. C. (1990). The Interpretation of Texts, People and Other Artifacts. Philosophy and Phenomenological Research, 50.S : 177-194.

Eco, U. (1979). Lector in fabula: la cooperazione interpretativa nei testi narrativi. Milano: Bompiani.

Eco, U. (1990). I limiti dell’interpretazione. Milano: Bompiani.

Fahad Khan et al. (2016). Restructuring a Taxonomy of Literary Themes and Motifs for More Efficient Querying, MATLIT 4.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Moretti, F. (2013a). Operationalizing: Or, the Function of Measurement in Literary Theory. New Left Review 84: 103-119.

Moretti, F. (2013b). Distant Reading. London: Verso.

Pistellato, A. (a cura di), Memoria poetica e poesia della memoria. La versificazione epigrafica dall'antichità all'umanesimo, Atti dell'incontro di studi Venezia 3-4 maggio 2012, Edizioni Ca' Foscari, Venezia, 2015.
Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
Segre, C. (1985). Tema/motivo. In Segre, C. Avviamento all'analisi del testo letterario. Torino: Einaudi, pp. 331-356.

Trascrizione e analisi linguistica del sogno: oggetti concreti nei sogni di schizofrenici e depressi

Daniele Silvi, Università di Roma 'Tor Vergata', silvi@lettere.uniroma2.it
Marco Zanasi, Università di Roma 'Tor Vergata', marco.zanasi8@gmail.com
Sergio Pizziconi, pizziconi.sergio@gmail.com

Introduzione

L'idea di Marco Zanasi di codificare e di studiare le trascrizioni dei sogni dei suoi pazienti presso l'Ospedale Psichiatrico di Tor Vergata ha dato inizio ad una ricerca di squadra su un'ipotesi stimo- lante: C'è qualche correlazione tra la realizzazione linguistica della trascrizione di un sogno e la psico- patologia di cui il sognatore sta soffrendo? Finora, le analisi delle trascrizioni dei sogni si sono con- centrate principalmente sugli attanti, gli ambienti e i descrittori delle condizioni emotive, durante l'attività onirica. L'obiettivo della ricerca è verificare un insieme di funzioni linguistiche che possono essere significativamente correlate al tipo di psicopatologia su base statistica. Il sogno è un fenomeno psichico di precipua rilevanza per evidenziare alcune relazioni profonde tra corpo e mente. La prospet- tiva junghiana non vede differenze tra i significati latenti e quelli manifesti, pertanto l'apparente illeg- gibilità (o interpretazione) di un sogno è subordinata al linguaggio allusivo, allegorico ed opaco, uti- lizzato dalla narrativa onirica dell'inconscio. Un sogno, pertanto, deve essere considerato come un testo narrativo reale, al quale è possibile applicare l'analisi dello stile, narratologico e della struttura. Una trascrizione verbale dei dati onirici è, attualmente, il solo mezzo per rappresentare l'esperienza onirica. Naturalmente, la trasposizione di detta esperienza in un sistema linguistico preciso, si deve confrontare con fenomeni di traduzione intersemiotici: cosa si conserva e cosa si trasforma? È possibile identifi- care delle variabili che rimangono immutate in questo passaggio? Assumendo di sì, come possiamo collegare questi elementi con certezza all'attività psichica del soggetto studiato, filtrandola dalle strut- ture pertinenti al processo verbale utilizzato per riprodurla? La tesi sulla quale il progetto si basa, è che sia possibile, applicando l'analisi testuale al sogno considerato come espressione "grafica" di un lin- guaggio inconscio, estrarre importanti informazioni da un materiale apparentemente caotico e non strutturato. Il nostro obiettivo è dunque la creazione di un codice che consenta di leggere tali informa- zioni e non semplicemente una codifica che produca un testo annotato. A questo prodotto finale, la trascrizione del sogno, abbiamo poi applicator diverse tipologie di analisi, producendo tabelle e grafici che ne riportino I risultati. In questo paper intendiamo presentare i risultati della nostra ricerca sulla presenza, in questi racconti onirici, degli oggetti concreti: le cose che vengono menzionate dai pazienti in esame. In particolare, approfondendo questo punto, oggetto centrale del paper che presentiamo, se- guendo le suggestioni della grammatica funzionale di M.A.K. Halliday (1994), abbiamo indagato dif- ferenze e analogie nel modo i cui gli oggetti concreti sono inclusi nella struttura sintattica dei rapporti di sogni di schizofrenici e depressi. In particolare, abbiamo distinto la rappresentazione degli oggetti concreti all'interno dei processi espressi dai verbi in forma finita o non finita nel ruolo sintattico di (a) partecipanti o (b) circostanze.

La differenza, ad esempio, configura il ruolo sintattico di "autobus" menzionato in un sogno di una persona depressa nel ruolo di partecipante, strutturalmente il soggetto del processo nella frase: "l'autobus se ne va"; mentre è nel ruolo di circostanza a indicare il

luogo del processo verbale, nella frase: “mi trovavo su un autobus di quelli tipo aperti che sono da # da turismo no?”.

Abbiamo ridotto i ruoli di partecipante alle corrispondenti funzioni sintattiche di soggetto, oggetto diretto, oggetto indiretto, oggetto obliquo. La funzione circostanziale è stata invece ridotta ai tipi luogo, tempo, causa/fine, modo/maniera, condizione.

Il ruolo sintattico degli oggetti è stato quindi posto in correlazione con il tipo sintattico di processo distinguendo tra processi transitivi attivi, transitivi passivi, intransitivi, copulativi.

In questa porzione di sogno di persona cui è stata diagnosticata la schizofrenia si possono evidenziare diversi tipi di processo e di ruolo sintattico degli oggetti: “Ho sognato le farfallette per mettere le flebo mi sono sognato una farfalletta della flebo con un tubicino dal quale usciva tanto sangue ed era il mio sangue”. “Farfalletta” e “farfallette”, oggetti diretti strutturalmente, sono rappresentati come parteci- panti in due processi transitivi. “Flebo” è rappresentato con lo stesso ruolo nello stesso tipo di processo. “Sangue” è rappresentato come soggetto partecipante in un processo intransitivo (“usciva”) e in un processo copulativo (“era il mio”). “Tubicino” nella forma pronominale “dal quale” è rappresentato come elemento circostanziale a indicare il luogo di un processo intransitivo.

Tale matrice di dati arricchisce le correlazioni finora trovate nell'ambito del progetto Dream Coding tra forme linguistiche e tipo di psicopatologia diagnosticata ai soggetti che hanno raccontato i loro sogni per il gruppo di ricerca. L'idea originale di Marco Zanasi (Zanasi et al. 2005, 2008, 2010; Pizzi- coni, Zanasi e Silvi 2015) è che sia possibile affiancare ai normali protocolli diagnostici un'analisi linguistica dei racconti di sogni che individui la presenza di tratti linguistici che nelle indagini correnti e future sono stati correlati alle diverse psicopatologie.

Un uso terapeutico della TEI

In quanto finalizzato alla creazione di un codice per la lettura dei sogni, il nostro progetto di ricerca è stato denominato Dream coding. Si occupa di raccogliere, trascrivere e catalogare i sogni di un campoone di pazienti psichiatrici (raggruppato in due categorie: depressi bipolari e psicotici) e di controllo (soggetti sani). In seguito questi sogni vengono analizzati e valutati attraverso gli strumenti dell’informatica umanistica (codifica xml e analisi linguistica computazionale), con la finalità di indi- viduare delle correlazioni tra i due tipi di psicopatologie e particolari configurazioni di strutture, forme e significati. Nella presentazione saranno discussi i risultati dell'analisi di tempo, aspetto e modalità del sistema verbale nel corpus del progetto.

Il progetto nato nel 2008, si avvale di una equipe di ricercatori che raccolgono periodicamente, e tras- crivono (secondo delle norme specifiche, che verranno illustrate), i sogni dei pazienti precedentemente descritti, dividendoli in base al sesso, all’età, alla regione di provenienza ed al grado di istruzione. I raggruppamenti di patologie prese in esame, sono alcuni tra i più noti dal punto di vista clinico ed i pazienti afferenti sono di diagnosi accertata, in cura presso la struttura ospedaliera dell’Università di Roma Tor Vergata, dipartimento di Neuroscienze, clinica S. Alessandro.

I pazienti sono invitati a raccontare i loro sogni, che vengono registrati ed in seguito trascritti. In ques- ta trascrizione vengono riportati anche gli aspetti intonativi e le pause del discorso (non ci riferiamo alle normali pause dipendenti dalla sintassi del periodo, bensì a quell’attività di interruzione del dis- corso che emerge per il fatto che il paziente deve ricordare cosa ha sognato, o riflettere su ciò che sta per dire, ecc).

Per verificare la presenza o l’assenza, nel nostro studio, di un certo modo con cui una caratteristica specifica può presentarsi, abbiamo impiegato due ricercatori che alla cieca, ciascuno separatamente, hanno analizzato le relazioni di ogni singolo soggetto iscritto nei

due gruppi distinti: “schizofrenici” e “controlli”. Dai dati dell’analisi è emerso che i ricercatori sono in accordo sulla presenza di un tratto specifico nel 95.05% di tutte le trascrizioni. Per quanto riguarda il rimanente 4,95%, un terzo ricerca- tore, in modo indipendente, come detto sopra, ha proceduto ad un nuovo esame e ha deciso se includere o escludere ciascun caso dal nostro studio. L’indice di concordanza adottato è stato ottenuto quale complemento all’unità della media dei valori assoluti dei confronti ottenendo così delle entità positive minori o al massimo uguali all’unità.

In seguito, un secondo gruppo di ricercatori, specialisti di Informatica umanistica, si occupa di tradurre questa trascrizione in una codifica XML con marcatori TEI.

La prima fase di questa codifica è stata svolta applicando la release P5 della TEI, in particolare il tags- et denominato Interpretation and Analysis. Alcuni dei tag utilizzati sono: <interpGrp>, <interp>, <s>; tra gli attributi troviamo invece: ana, type, e id. Gli aspetti del discorso messi in evidenza in questa codifica di primo livello, sono: aspetti intonativi, morfosintassi del verbo, proforme, ripetizioni, sostituzioni, ellissi, metafore, pause.

La successive analisi, condotte attraverso l’analizzatore testuale TAPoR e l’applicazione di alcune leggi della Statistica testuale, mirano a individuare due tipi di tendenze statistiche:

1. la correlazione tra gruppi di psicopatologie e consistenza di singole scelte linguistiche, in partico- lare tempo, aspetto e modalità dei sintagmi verbali anche confrontate con le scelte del gruppo di con- trollo. Inoltre si rileverà la presenza di covarianze tra le categorie linguistiche;
2. coesione e coerenza interna ai singoli riporti di sogno.

L’analisi linguistica

Da un punto di vista linguistico deve essere subito rimarcato che non ci si deve attendere, fatta ec- cezione per particolari casi idiolettali, una varietà linguistica completamente estranea dal repertorio linguistico italiano. In un’ottica diasistemica (Weinreich, 1953) e adottando la prospettiva analitica della Varietätengrammatik (Klein, 1988; Dittmar, 1989; Klein and Dittmar 1979), si ricercano dei pro- fili espressivi che presentino una particolare correlazione con i gruppi sperimentali. Per quanto riguar- da il sistema verbale, ad esempio, se è chiaro che i tempi siano quelli presenti nel repertorio dell’italiano, abbiamo individuato distribuzioni di frequenza distintive nel loro uso. Sulla scelta del tempo grammaticale si è tenuto conto della differenza diatopicamente marcata nell’uso di passato re- moto (più frequente nei dialetti meridionali) e passato prossimo (più frequente nei dialetti centro-settentrionali) dell’indicativo. È questo un caso evidente di quell’azione di filtraggio, menzionata nei preliminari, di elementi espressivi che possono essere determinati dalla variazione sociolinguistica dell’italiano. Altre co-variabili di cui si è tenuto conto sono il livello di istruzione, sesso e età dei pa- zienti.

La scelta tra passato prossimo o passato remoto è anche significativa di come l’aspetto del verbo sia stato trattato. I diversi aspetti (incoativo, durativo, perfettivo, imperfettivo, abituativo…) possono es- sere veicolati o morfologicamente o lessicalmente. Questa è la prima variazione presa in considera- zione nell’annotazione dei testi dei sogni. Nel gruppo delle selezioni morfologiche il passato prossimo avrebbe la potenzialità sistemica di marcare l’aspetto perfettivo, ma per la sovrapposizione con la var- iazione sociolinguistica già descritta questa opzione deve essere controllata in modo differente.

Ulteriori filtraggi sono stati necessari anche per la modalità del verbo. Infatti, posta la distinzione in- iziale tra la sua espressione grammaticalizzata e lessicalizzata, nei casi in cui la modalità è espressa dal modo verbale possono esserci variazioni sociolinguistiche che annullano alcune differenze sintattico semantiche. Per cui, ad esempio, la differenza tra indicativo e congiuntivo che in italiano standard let- terario è significativa tende ad esserlo meno in base a differenze regionali e sociali dei parlanti.

I risultati ottenuti

Il lavoro, data la sua complessità, è stato strutturato in fasi successive, così riassumibili:

Raccolta dei dati: i pazienti che hanno già una diagnosi clinica definita vengono nominati di "Tipo A" e individuati con un metodo di schedatura che prevede la registrazione di dati socio-demografici e dello status diagnostico. Questa prima schedatura viene fatta per creare una raccolta di dati linguistici associabili a patologie conosciute, onde rilevare la presenza delle particolarità linguistiche associate ad una determinata patologia. La seconda parte della raccolta dei dati ha riguardato più propriamente la trascrizione, catalogazione e archiviazione dei sogni, come abbiamo già detto precedentemente.

Analisi dei sogni: Terminata la raccolta e l'archiviazione del materiale da analizzare (i sogni dei pa- zienti di Tipo A) si è passati ad una operazione di codifica del testo dei sogni, attraverso un linguaggio di marcatura, che ne permetterà sia l'analisi automatica che la pubblicazione o la stampa. Le caratteris- tiche del materiale che abbiamo sono le seguenti:

1. quello in esame è un caso di codifica del parlato: ciò involve tutte le problematiche affrontate su questo argomento da ampia tradizione per quanto riguarda sia la codifica che la normalizzazione lin- guistica;
2. la modellizzazione deve essere ben motivata e finalizzata, poiché una scelta diversa in questa fase può indirizzare la ricerca in una direzione piuttosto che in un'altra;
3. si deve prevedere la reale possibilità di dovere rivedere o integrare la codifica a seconda dei risul- tati che verranno in seguito all'analisi automatica dei testi, di conseguenza la struttura della codifica XML deve rimanere aperta e fluida.

Una volta stabilito il modello e applicata la marcatura, si sono analizzati i testi trascritti con un software che faccia ricerche statistiche e lessicali, come TAPoR. Abbiamo scelto di utilizzare questo strumento proprio perché è in grado di fare ricerche all'interno di tag XML.

Verifica del protocollo: il protocollo creato nella fase precedente dovrà poi essere verificato e cor- retto: per far questo lo si applicherà ai pazienti con una diagnosi clinica non ben definita (che chia- meremo di Tipo B); si partirà cioè dalle ipotesi linguistiche, verificando nei nuovi pazienti l'uso della lingua (ad un livello inizialmente solo statistico) e correlandolo a delle ipotesi di patologia.

Comparazione dei risultati: in quest'ultima fase i dati dei pazienti Tipo A e Tipo B vengono messi a confronto tra loro. In seguito a questo confronto verrà creata una griglia delle particolarità linguistiche più importanti rilevate che siano state associate (per alti riferimenti statistici) a determinate patologie. Si partirà dalle particolarità linguistiche che hanno un riscontro percentuale più alto rispetto ad altre ed infine verrà proposta una "diagnosi linguistica", da affiancare alla "diagnosi clinica" tradizionale.

È stato osservato che i pazienti, nel raccontare i loro sogni, inseriscono delle pause nel discorso. Questo è un elemento che abbiamo ritenuto importante da conservare nella codifica (e quindi nel modello). A tale scopo si è creata una griglia di conversione che ponga in relazione la durata della pausa (in secondi) che il paziente introduce, ed una simbologia di trascrizione (e di codifica). Il tutto è stato formalizzato nella "scala dei silenzi", riportata in figura 3:

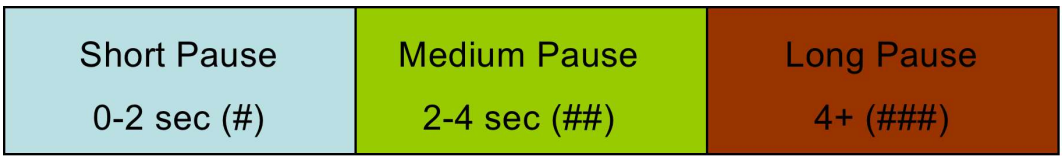


Fig. 3: Scala dei silenzi

In questa prima, propositiva, fase del progetto gli elementi di fatto codificati sono stati i seguenti:

- Intonation features (tag)
- Verb Morphosyntax (tag)
- Pro-forms and repetitions (tag with label of cross-reference)
- Metaphors (tag)
- Pauses (entities)

Per la visualizzazione dei testi così codificati si è scelto di operare una trasformazione XSLT, grazie alle strutture di indagine all'interno dei tag e degli attributi che questo linguaggio mette a disposizione. Si è creato un foglio di stile XSL che tenesse conto di tutti i fenomeni codificati, e quello che segue è un breve esempio di visualizzazione:

Mi ricordo di aver sognato, # ed era un sogno a colori, ## di cui però adesso mi sfugge la trama. qualcuno che mi gridava a viva voce "Trentotto! Trentotto!" me lo ripeteva e si arrabbiava perché io non # gli prestavo attenzione. ## ""Trentotto! Trentotto!" La cosa principale è che è buio e # sempre di notte quando. ## Ho la sensazione di soffoca', e di # impazzi'. ## E di solito o corro o come se corro se corro da qualcosa, non lo so, scappo da qualcuno. ## E # poi comunque in genere mi risveglio dopo # subito dopo. Niente, stavo, # sempre di notte, però su una bicicletta. Cercavo disperato una casa ma era tutto # non c'erano case dove stavo quindi cercavo erano tutte vie # vie, spiazzali e # correvo sempre con questa bici in cerca di una casa non so qualcosa ### e con questa bicicletta, stavo sola sempre.

Nell'esempio sopra riportato si sono messi in evidenza i cambiamenti di tempo verbale, individuati con diversi colori.

In fase di marcatura dei testi, sono stati individuati un novero di tempi verbali (riassunti di seguito) e di questi è stata studiata la distribuzione all'interno dei corpora e poi raffigurata in grafici a torta.

condizionale_passato, condizionale_presente, congiuntivo_imperfetto, congiuntivo_trapassato, fu- turo_semplice, gerundio_presente, imperfetto, infinito, non_specificato, participio_passato, passato_prossimo, passato_remoto, presente, trapassato_prossimo, Infinito passato.

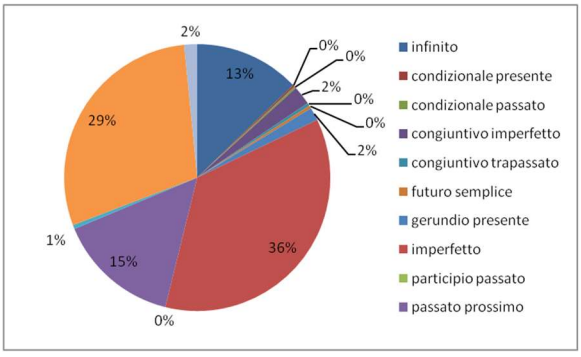


Grafico 1: tempi verbali nel corpus "Depressi bipolari"

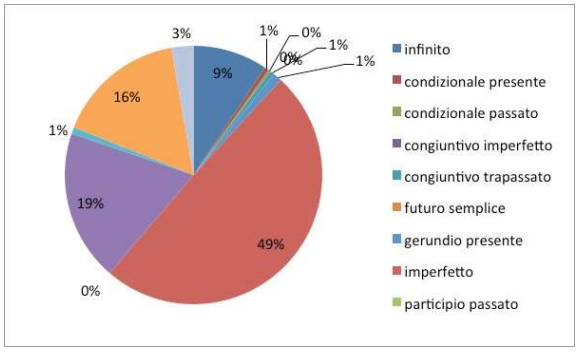


Grafico 2: tempi verbali nel corpus "Controlli"

Riferimenti bibliografici

Aserinsky, E. and Kleitman, N. (1953). Regularly occurring periods of eye motility and con-concomitant phe- nomena, during sleep. Science, 1 118: 273-4.
Chatman, S. (1978). Story and discourse, narrative structure in fiction and film. Neuropsychology Review, 13(2), 43–77.
Courtés, J. (1986). Le Conte populaire: poétique et mythologie. Paris: Presses Universitaires de France. Dittmar, N. (1989). Variatio delectat. Le basi della sociolinguistica. Galatina: Congedo.
Domhoff, G.W. and Schneider, A. (1998). New rationales and methods for quantitative dream research outside the laboratory. Sleep, 21, 398-404.
Dorus, E., Dorus, W., & Rechtschaffen, A. (1971). The incidence of novelty in dreams. Archives of General Psychiatry, 25, 364-368.
Foulkes, D. (1985). Dreaming: A Cognitive-Psychological Analysis. Hillsdale, NJ: Lawrence Erlbaum
Gee, J.P. (1998). Two styles of narrative construction and their linguistic and educational implications. In Chesh- ire, J. and Trudgill, P. (eds.) (1998). The Sociolinguistics Reader. Gender and Discourse. Vol. 2. London – New York: Arnold, pp. 295-314.
Gigliozzi G. (2003). Introduzione all'uso del computer negli studi letterari, New revised edition by Fabio Ciotti. Milano, Bruno Mondadori-Paravia.
Hall, C. (1966). Studies of dreams collected in the laboratory and at home. Institute of Dream Research Mono- graph Series (No. 1). Santa Cruz, CA: Privately printed.
Halliday, M. A. K. (1994). An Introduction to functional grammar. London-Melbourne-Auckland: Edward Ar- nold.
Pizziconi, S., Zanasi, M. & Silvi, D. (2015). Dream Coding: Re-writing dream reports as an object of textual analysis. In Zofia Ziemann and Wojciech Owczarski (eds.). Dreams, Phantasms and Memories. Gdańsk, Po- land: Wydawnictwo Uniwersytetu Gdańskiego (University of Gdańsk Press), pp.113-126.
Jung, C.G. (1945). L´essenza dei Sogni. Opere, Vol. 8. Torino: Boringhieri.

The emergence of Digital Humanities: An epistemological cartography of thematic issues in French academic journals

Quoc Tan Tran, GERIICO Laboratory, University of Lille 3, France,
quoctanvn@gmail.com

The growing importance of the *computational turn* has deeply affected the landscape of the social sciences and humanities (SSH). One of the most profound transformations caused by the development of digital technologies is the changes of the practice conditions and the production of knowledge (Berry 2011; Gold 2012; Liu 2009). In recent years, French academics working in the SSH have been devoting attention to "Digital Humanities" (DH), a novel territory that fosters collaboration, openness and enhancement of knowledge.

By assessing the emergence of those new practices, we can (re)discover a close link between technical culture and the culture of scholars. Since the culture related to computers appears highly technical (Guichard 2013), it is an epistemological challenge to provide critical assessment of the evolution, the structure and the dynamics of this new phenomenon. The study of this effect in the light of new epistemologies and paradigm shifts is promising in such a way that allows us to understand what is *fabricated* in the phenomenon, in terms of digital culture, both profanely and scholarly, and how much our digital practices are orchestrated (Kitchin 2014). The starting point of this work is to trace the emergence of DH, using corpus analysis, through special thematic issues of French academic journals. The next initiative is to find ways to illustrate, using cluster analysis, how digital mediation and networked collaboration influence the problematic(s) and the episteme(s) of disciplines commonly grouped under the label SSH: the adoption of emerging approaches and new research methods, editorial practices, and also combinatorial potentialities and questions posed by the uses of the "digital".

Corpus analysis

The aim of this task is to map main themes and emerging practices of French academic sphere, in terms of research, knowledge production and dissemination. The further objective is to locate, by a cartographic and cluster-based approach, the current state and scientific positioning of France research in the international context. The scope is not extended to the whole French-speaking landscape, which means that the research is bounded to the academic journals in France only, not in Belgium, Quebec, Switzerland, and the francophone African countries.

The database Cairn and the portal Revues.org offer the comprehensive collection of French language publications in the SSH. The material analyzed has been collected from nine academic journals (electronic and print) (see Table 1), which have special thematic issues dedicated to the topics of DH. It should be noted that four of them are published exclusively in electronic format, showing an innovation in the way that scientific information is communicated to the research community. The corpus is composed of 77 articles (in which 30 is co-authoring) in ten issues, whose themes varies from neo-structural sociology, critical theory, to transmedia and computer literacy (see Table 2). All of the issues had been published in a period of less than three years. It is noticeably short when we consider the publishing process in scholarly journals, but also remarkable in terms of an "emergence". The earliest

issue titled "Digital epistemologies of the SSH"[1] appeared in *Revue Sciences/Lettres* (n°2, 2014). The latest thematic issue is "DH and Information and Communication Sciences", which had been published in *Revue française des SIC* (n° 8) in April 2016.

Table 1 Distribution of the articles analyzed per newspaper

Journals	*Abbreviation*	*Type*	*N*
Cahiers du Numérique	CdN	P, E (c)*	14
Critique	C	P, E (c)	10
Multitudes	M	P, E (c)	4
Revue d'Anthropologie des Connaissances	RAdC	P, E (c)	5
Revue Française des SIC	RFdSIC	E (r)	8
Revue Sciences/Lettres	RSL	E (r)	12
Socio	S	P, E (r)	10
Tic & Société	TS	E (r)	8
Variations	V	E (r)	6
Total			77

*Notes: P = print; E = electronic; (c) = available in Cairn; (r) = available in Revues.org

It is astonishing when we take into account the journals' periodicity. While only *Critique* is published monthly, three are quarterly (*Cahiers du numérique*, *Multitudes*, *Revue d'Anthropologie des Connaissances*), one is tri-annual (*Socio)*, one is bi-annual (*Tic & Société*), and the other three are annual (*Revue françaises des SIC*, *Revue Sciences/Lettres*, and *Variations*). While most of the journals are in the fields of Information and Communication Sciences, there are some exceptions that might raise curiosity. For instance, the journal *Critique*, founded in 1946 by Georges Bataille, is a *revue* of general interest, with the orientation slanted towards literary and cultural analysis.

Table 2 Summary of 10 thematic issues in 9 journals

Issue	*Title*	*Themes*
Revue Sciences/Lettres (n°2, 2014)	Digital epistemologies of the SSH	epistemology, neo-structural sociology, anthropology, digital fictions
Tic & Société (vol. 7, n° 2)	Digital worlds: new research perspectives	digital research, qualitative and quantitative data, ethical issues of digital practices
Cahiers du numérique (vol. 10, n° 3)	The "delivered" humanities	transliteration, transmedia, delivery, network analysis
Cahiers du numérique (vol. 10, n° 4)	What arrangements for DH?	holistic approach, arrangement, digital traces, participative science, social simulation
Revue d'Anthropologie des Connaissances (vol. 8, n° 4)	The "delivered" humanities	book culture, re-invented reading
Socio (n° 4, 2015)	The digital turn... and after?	digital environment, computational turn, scientific culture, resources

1 All titles and terms originally in French have been translated by the author.

Multitudes (n° 59)	DH 3.0	DH manifesto, computational subjectivity, computer literacy, media archeology
Critique (n° 819-820)	Numbers and letters	transhumanism, digital materialism, mutations, algorithmic governance
Variations (n° 19)	Criticism and DH: For a materialistic approach of the "immaterial"	scientific production, social simulation, digital breaking, critical theory
Revue française des SIC (n° 8)	DH and Information and Communication Sciences	Uses of *dispositif*, discourse of the information society, empirical practice of digital, pluridisciplinarity, design turn

Cluster Analysis

Corpus-based clustering method is used to identify trends, patterns and the true ecological mechanisms. We aim to illustrate emerging topics and avenues of research, and tackle the questions of DH's disciplinary cohesion that have generated widespread discussion and debate among French academics. Based on the corpus of full-text articles, we extracted word co-occurrence counts over the corpus and map all co-occurrences into a two-dimensional cartography using VOSViewer. We then examined cluster networks of the whole corpus (see Fig. 1) and of each journal (see Fig. 2).

Fig. 1 shows the co-occurrence term map of all ten thematic issues. Each term or concept is represented by a circle. The circle's diameter and the label size represent the term's frequency. Its proximity to another term indicates the degree of correlation between the two concepts, and its color represents the group to which it conceptually belongs. Because the concepts are spatially interdependent in three-dimensional space and the program is limited to two-dimensional representation, some relationships cannot be obvious.

Figure 1 Co-occurrence term map of the whole corpus

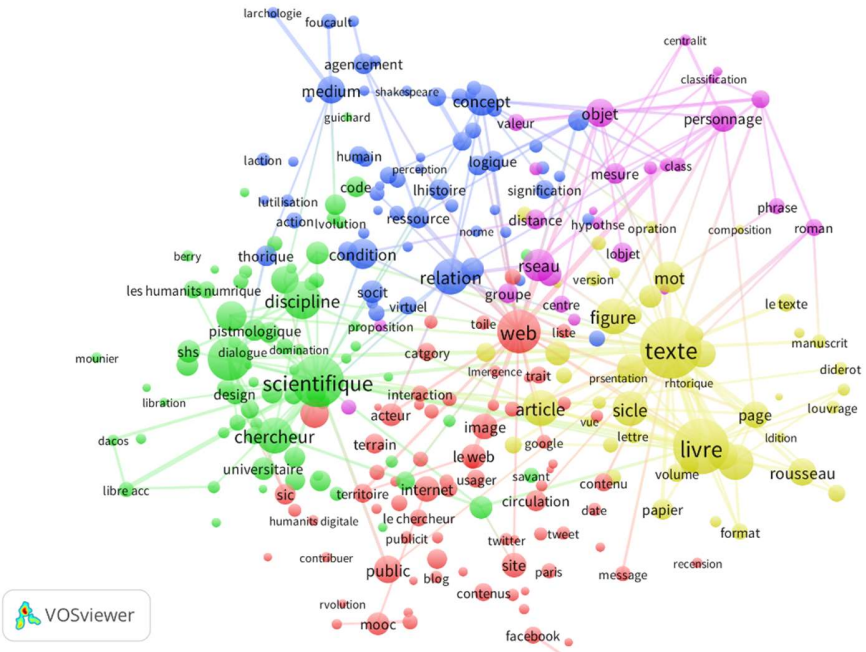
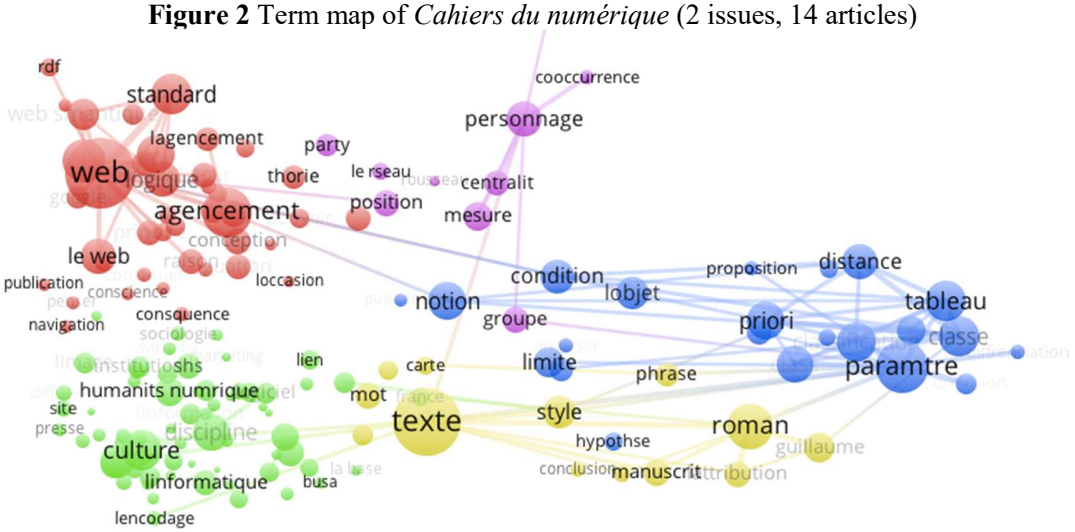


Figure 2 Term map of *Cahiers du numérique* (2 issues, 14 articles)



All the journals' term maps contain the main DH cluster (green), reflected by terms such as "*humanités numériques*", "digital humanities", "social sciences", and associated terms such as "science", "computer science", "research", etc. Regarding different dimensions of DH, *Socio* and *Revue Sciences/Lettres* put more emphasis on "research", while *Cahiers du Numérique* focuses on "digital culture", *Variations* on "scientific". In case of *Critique*, terms such as "tool", "treatment", "nature" and "theory" suggest that this journal concerns more with the issues of theoretical computing and algorithm studies. *Revue française des SIC* presents a sub-group of "design", which reflects the emphasis on the "design turn" of the DH.

Table 3 List of clusters with the most frequently used terms

Cluster	*Frequently used terms*
Cluster 1 (green)	scientific, discipline, digital humanities, epistemology, SSH, dialogue, design, academic, researcher, free access, Berry, code, domination, liberation, evolution
Cluster 2 (red)	web, image, public, MOOC, content, Internet, terrain, actor, site, Google, content, traffic, Facebook, interaction, emergence, size, ICT, message, review, blog
Cluster 3 (yellow)	text, book, century, page, article, figure, Rousseau, work, manuscript, Diderot, edition, letter, presentation, word, rhetoric, paper, volume, view, version, composition
Cluster 4 (blue)	concept, arrangement, medium, Shakespeare, concept, Foucault, archeology, logic, history, resource, condition, society, virtual, theoretical, human, action, meaning, use, norm, perception
Cluster 5 (pink)	network, center, group, object, measure, class, character, centrality, classification, sentence, novel, hypothesis, value, distance

Based on those term maps, we analyze each cluster by looking back the articles that contain the cluster's most popular keywords, and define the transversal themes of each cluster:

1. Epistemological issues posed by the digital humanities: the role of ICT as a tool to support and enhance interdisciplinary research within the SSH, the *making* of citizen participation, its community-based and collaborative essence. (Cluster 1)
2. Researchers' re-consideration of the role played by social media and new socio-political relationships, and the practices they make possible (Cluster 2)
3. Digital methods in the humanistic framework: Research and the creation facing the digitization of cultural heritage (Cluster 3)
4. Rethinking of the scientific ecology (Cluster 4)
5. DH's opportunities to broaden the spectrum of working and exchange languages, and to push the humanities into new territory of collaboration, openness, and experimentation (Cluster 5)

In examining the Cluster 1, we see how the *digital turn* has considerably transformed everyday life, shattering the relationship that individuals have with the world and leading them to reinvent their ways of interacting. The Cluster 2 indicates that the rise of new technologies has also had major scientific consequences: with the changing conditions of knowledge production and dissemination, the whole relationship to scientific research has been transformed. In the Kuhnian sense of a shift in the ontology of the positive sciences, the Cluster 5 suggests a humanistic understanding of technology, which consists of the "urgent inquiry into what is human about the *computational* humanities or social sciences" (Berry 2011, 9). Both the Cluster 3 and 4 reflect on the new knowledge, new uses, new postures and new paradigms that characterize social and human sciences research in the digital age, including the new circulation of knowledge results and the ability of researchers to make the best use of the data produced by these new tools. What is in debate is whether and how the work that digital humanists perform is scholarly and theoretical in scope (Cluster 4).

Future Work

In this paper, we attempt to trace the emergence of DH through ten special and thematic issues of French journals. The aim is to map the emerging practices and the dynamics of this phenomenon, in terms of research, knowledge production and knowledge dissemination, and the modalities that allow DH to develop new social and editorial assignments. To achieve the further objective of locating the current state of France research in the international context, and to move discussion forward from that point, there is a need for further bibliographic analysis. It is necessary to study the co-citation and co-authorship networks, and perform cluster analysis in both networks. The relative lack of international collaborations among French authors in our corpus causes the difficulty to track, analyze and visualize research using citation databases such as Google Scholar or Scopus. Efforts will be made to find an effective and French-customized co-citation method.

Bibliographic References

Berry, David M. 2011. "The Computational Turn: Thinking about the Digital Humanities." *Culture Machine* 12. http://culturemachine.net/index.php/cm/article/viewDownloadInterstitial/440/470.
Gold, Matthew K. 2012. *Debates in the Digital Humanities*. Univ of Minnesota Press.
Guichard, Éric. 2013. "L'internet et Les Épistémologies Des Sciences Humaines et Sociales." *Revue Sciences/Lettres* 2 (2). École normale supérieure. http://rsl.revues.org/389.
Kitchin, R. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1 (1): 1–12.

Liu, Alan. 2009. “Digital Humanities and Academic Change.” *English Language Notes*. https://liu.english.ucsb.edu/wp-includes/docs/writings/dh-and-academic-change-page-proofs.pdf.
Kilroe, P. (2000). The Dream as text, the dream as narrative. Dreaming, 10, 125–138.
Klein, W. (1988). Varietätengrammatik. In Ammon-Dittmar-Mattheier (eds), Sociolinguistics/Soziolinguistik, Vol. II. Berlin – New York: de Gruyter, pp. 997-1006.
Klein, W. and Dittmar, N. (1979). Developing Grammars. The Acquisition of German by Foreign Workers. New York – Heidelberg: Springer.
Kramer, M. (1993). Dream translation: an approach to understanding dreams. In Delanay, G. (ed.) (1993). New directions in dream interpretation. Albany, NY: SUNY Press, pp. 155-94.
Kramer, M., Kinney, L., & Scharf, M. (1983). Dream incorporation and dream function. In W. Koella (Ed.), Sleep 1982: The 6th European Congress on Sleep Research, Zurich, 1982. Basel, Switz: Karger. Pp.369-371.
Kramer, M., Roth, T. & Czaya, J. (1975). Dream development whithin a REM period. in P. LEVIN and W. KO- ELLA (Eds.), Sleep 1974: The 2nd European Congress on Sleep Research, Rome, 1974. Basel, Switz: Karger. pp. 406-408.
Lakoff, G. and Johnson, M. (1980). Metaphors We Live By. Chicago – London: The University of Chicago Press.
Meier, B. (1993). Speech and thinking in dreams. In C. Cavallero & D. Foulkes (Eds.), Dreaming as cognition.
New York: Harvester Wheatsheaf, pp. 58-76.
Moro, A. (2006). I confini di Babele. Il cervello e il mistero delle lingue impossibili. Milano: Longanesi.
Roffwarg, H., Dement, W., Muzio, J., & Fisher, C. (1962). Dream imagery: relationship to rapid eye movements of sleep. Archives of General Psychiatry, 7, 235-258.
de Saussure, F. (1922). Cours de linguistique générale. Paris: Edition Payot.
Snyder, F. (1970). The phenomenology of dreaming. In L. Madow & L. Snow (Eds.), The psychodynamic im- plications of the physiological studies on dreams. Springfield, IL: C.C. Thomas, pp. 124-151.
Strauch, I. (1969). Psychological Aspects of Dream Recall. Paper presented at the 19th International Congress of Psychology, London.
Taub, J., Kramer, M., Arand, D., & Jacobs, G. (1978). Nightmare dreams and nightmare confabulations. Com- prehensive Psychiatry, 19, 285-291.
Toolan, M.J. (1988). Narrative: a critical linguistic introduction. London – New York: Routledge.
Zanasi, M., De Persis, S., Caporali, M. and Siracusano, A. (2005). Dreams and Age. Perceptual And Motor Skills, 100: 925-38.
Zanasi, M., Pecorella, M., Chiaramonte, C., Niolu, C., & Siracusano, A. (2008). Dreams by persons with mood disorders. Psychological Reports, 103(2), 381–394.
Zanasi, M., Chiaramonte, C., Paoletti, G., Testoni, F., & Siracusano, A. (2010). Oneiric activity in anorexia.
Sleep and Hypnosis, 12(1/2), 1.
Zanasi, M., Calisti, F., Di Lorenzo, G., Valerio, G., & Siracusano, A. (2011). Oneiric activity in schizophrenia: Textual analysis of dream reports. Consciousness and Cognition, 20(2), 337-348. Conscious Cogn. 2011 Jun;20(2):337-48. doi: 10.1016/j.concog.2010.04.008. Epub 2010 May 15
Underwood, T. (2012a). Topic modeling made just simple enough. The Stone and the Shell. https://tedunderwood.wordpress.com/2012/04/07/topic-modeling-made-just-simple-enough/.
Sollors W. (1993). *The Return of Thematic Criticism.* Cambridge(MA)/London: Harvard University Press.

Visualizing Gender Balance in Conferences

Sytze Van Herck, KU Leuven, sytze.vanherck@student.kuleuven.be

As the graphic display of abstract information, data visualization serves two purposes, namely sense-making or data analysis and communication (Few 2013, 2229). As illustrated in the report of Malu A.C. Gatto on *Making Research Useful: Current Challenges and Good Practices in Data Visualisation*, "academics have often struggled to share their data with other actors and to disseminate their research findings to broader audiences." (Gatto 2015, 4). However, data visualization can truly advance research, not only as a communication tool towards a broader audience, but especially as a tool that allows pre-attentive processing of vast amounts of information. In short, "data visualization reduces knowledge gaps" (Gatto 2015, 5). Since conference data sets are often too large to process without the help of visualizations, I would like to introduce several techniques.

Michael Jensen identified some of the most important questions regarding digital scholarship more than a decade ago that remain relevant today. He asked Digital Humanities professionals: "how can we most appropriately support the creation and presentation of intellectually interesting material, maximize its communicative and pedagogical effectiveness, ensure its stability and continual engagement with the growing information universe, and enhance the reputations and careers of its creators and sustainers?" (Jensen 2004, 551). In order to engage with the audience or reader, the design of the project should not be overlooked, even though "necessity often dictates that we adopt and adapt tools and technologies that were originally developed for other needs and audiences." (Kirschenbaum 2004, 539).

I will illustrate the development of and uses for visualizations in Digital Humanities research based on my own visualizations of the gender balance in ten computer science conferences from 2000 until 2015. The dataset was created by Swati Agarwal et al. in the context of an article on *Women in computer science research: what is the bibliography data telling us?* (Agarwal et al. 2016) and can be accessed online via Mendeley Data.[65] In order to test data visualizations, several queries limited and structured the data further to create smaller subsets per year and per conference for network visualizations, which were combined to visualize authorship demographics and the evolution of the gender balance across conferences and over the years. The results of the queries had to be adapted to the format and structure accepted by either visualization software such as Tableau, or code from a visualization library such as Protovis and D3.js, which was done in Python. The iterative process of creating visualizations consisted of four prototypes. Two prototypes focused on a network visualization and two explored the evolution and general trends in authorship demographics. The first network visualization was based on an arc visualization in Protovis and showed every single author represented as a dot on a horizontal line, where the size of the dot or node represented the

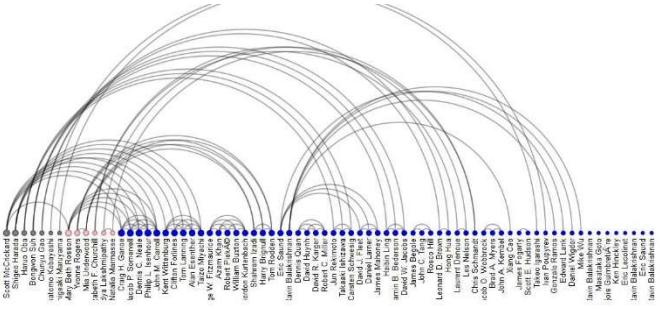


Figure 2: Protovis visualisation of the 2003 UIST conference.

[65] For more information regarding the dataset, please read the description of the dataset online: https://data.mendeley.com/datasets/3p9w84t5mr/1.

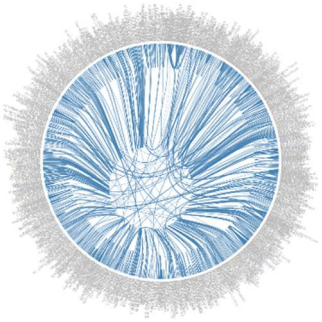


Figure 3 Chord Diagram of the 2005 CHI conference co-authorship network

number of co-authors, the color represented their gender and the lines connected authors working on the same paper. In the second network visualization authors were again represented individually as a dot, but this time arranged in the form of a circle and grouped by affiliation according to a hierarchical edge-bundling example from D3.js. The lines showed the relation between co-authors and when hovering over an author, the incoming and outgoing links to his or her co-authors highlighted. Unfortunately, network visualizations always run into the risk of cluttered screens, which is especially true for larger networks. However, added interactivity and filtering the data allows users to explore the results in a structured way and on their own terms.

Simple bar and line charts provided by Google Charts API demonstrating the evolution of gender balance in computer science conferences over time served their purpose in the first prototype on authorship demographics. Furthermore the first prototype also contained a map visualization of all the affiliations included in the dataset, created using the Google Maps API. These visualizations did not allow for interactive exploration, which is why the second prototype on authorship demographics used the software platform of Tableau.

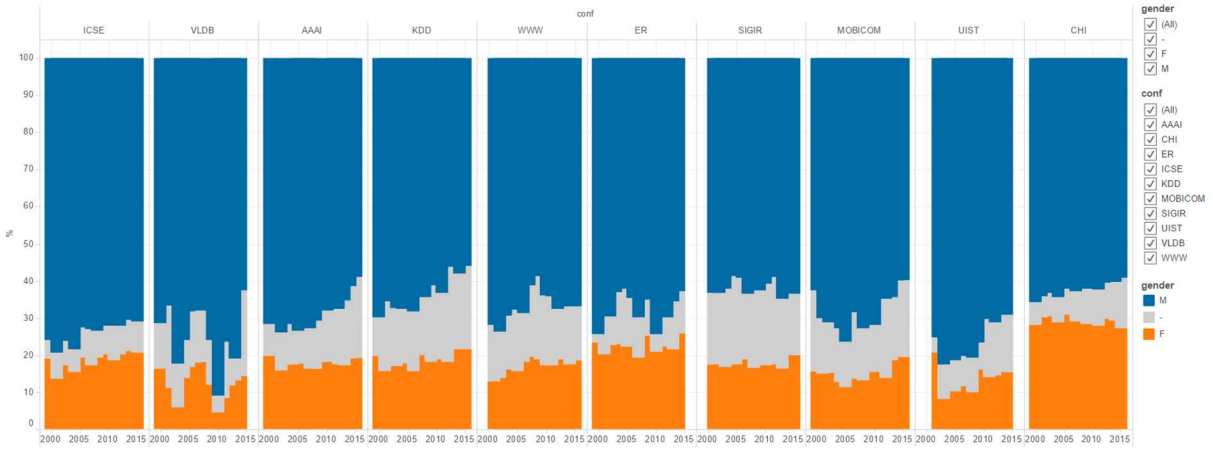


Figure 4 100% stacked bar chart of the gender balance in ten CS conferences from 2000 to 2016.

The first stacked bar chart of the second prototype demonstrating the evolution of gender balance over sixteen years was grouped per conference and color-coded according to gender and allowed hovering and filtering per conference, year and gender. In a tree map visualization showing the percentage of women per paper and grouped by conference, the same color-coding and filters were again adopted to allow the user to explore the data further. An added feature in Tableau is the details-on-demand, since hovering over data points often provides more detail. Finally to improve the previous map visualization, this time the percentage of women per

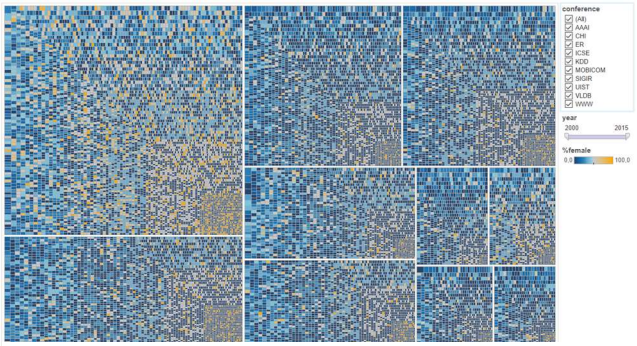


Figure 5 Treemap of the gender balance for each paper in ten CS conferences from 2000 to 2016.

country of their affiliation were displayed on a color scale and each country contained the total number of authors of all genders.

The existing dataset resolved any possible difficulties in accessing the data in this particular case, but the preparation of the dataset required substantial effort of its creators in structuring and enriching DBLP bibliography data with information on gender and affiliation (Agarwal et al. 2016). In order to visualize the data, free access to software or existing code is rarely guaranteed since this software is often commercial. Furthermore, if support for software or existing code ends, the visualization will likely disappear entirely. Regardless of challenges related to the visualization software, the main value of data visualization lies in both facilitating pre-attentive processing, as well as communicating results.

Norman's Action Cycle forms the basis of the interactivity of the visualizations (Norman 2013, 41). In order to form questions and find answers, the action cycle falls into two gulfs. First a user needs to form an intention, an action plan and he or she needs to execute this action during the gulf of execution. If this action has provoked a change in the world, for example the visualization shows an interesting pattern, then a gulf of evaluation follows. The perception of the visualization might lead to an interpretation which then needs to be evaluated again. Interactivity thus allows further exploration of the data by other researchers or the audience, while storytelling structures the relations between different visualizations and guides the audience through the research in a few clicks. The core value of these visualizations for the Digital Humanities therefore lies in accelerating data processing and raising possibilities for further research originally unimaginable.

Bibliographic References

Agarwal, Swati, Nitish Mittal, Rohan Katyal, Ashish Sureka and Denzil Correa. "Women in computer science research: what is the bibliography data telling us?" *ACM SIGCAS Computers and Society* 46:7-19. Accessed November 25, 2016. doi: 10.1145/2908216.2908218.

Few, Stephen. 2013. "Data Visualization for Human Perception." In *Encyclopedia of Human-Computer Interaction*, edited by Mads Soegaard and Rikke Friis Dam, 2229-2254. Aarhus: Interaction Design Foundation.

Gatto, Malu A.C., "Making Research Useful: Current Challenges and Good Practices in Data Visualisation." Reuters Institute for the Study of Journalism, May 2015.

Jensen, Michael. 2004. "Intermediation and its Malcontents: Validating Professionalism in the Age of Raw Dissemination." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell.

Kirschenbaum, Matthew G.. 2004. "'So the Colors Cover the Wires': Interface, Aesthetics, and Usability." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell.

Norman, Donald A.. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. New York: Basic Books.

PANEL

Reflections on reading history from a distance

Max Kemman*, University of Luxembourg, max.kemman@uni.lu
Mark Hill*, London School of Economics, m.j.hill@lse.ac.uk
John Regan*, University of Cambridge, jjr35@cam.ac.uk
Paul Nulty*, University of Cambridge, pgn26@cam.ac.uk
Peter de Bolla, University of Cambridge, pld20@cam.ac.uk
Pim Huijnen*, Utrecht University, p.huijnen@uu.nl
Tom Kenter, University of Amsterdam, tom.kenter@uva.nl
Daniele Guido*, University of Luxembourg, daniele.guido@uni.lu
*Presenting authors

Panel abstract

This panel aims to discuss how digital tools fit in to historical practices, and reflect on the interaction between digital and historical methods. While methods of computational textual analysis have been used in many other disciplines (Laver, Beniot, and Garry 2003, Grimmer and Stewart 2013, Schonhardt-Bailey 2006, Moretti 2013), historians (in particular, intellectual historians) have been suspicious of 'distant reading.' As Skinner argued, historical texts are not positive facts from which we can reconstruct empirical meaning (Skinner 2002). Nonetheless, there is an interest in these ideas and their application to historical studies. Thus, as a case study of the interaction between digital and historical methods, this panel will focus on how tools, and specifically distant reading techniques, may be used to investigate "concepts" in historical documents.

This panel follows recent investigations by the likes of Guldi and Armitage, who argue in *The History Manifesto* that, with the emergence of new digital materials, methods, and techniques, historians may be able to return to *longue durée* historical investigations (Guldi and Armitage 2014). However, how this is exactly to be done is not a triviality, and while most reflections have foregrounded the technical at the expense of the methodological, this panel hopes to investigate this relationship in more detail. That is, there is a shared concern that some digital investigations may introduce unexpected methodological problems and thus may ultimately have an impact on the accuracy of analysis and claimed conclusion. In essence, distant reading depends on counting occurrences of terms, but what these terms mean conceptually may change over time and context. When treating historical sources as big data and interpreting the outcome of distant reading methods, historians risk remaining ignorant to such *concept drift*. These issues have not been ignored (de Bolla 2013, Edelstein 2015, London 2016, Wang, Schlobach, and Klein 2011), but there remains a need for further serious methodological reflection, in addition to technological solutions, if we are to practice this sort of digital history.

To address these issues this panel brings together a number of researchers who are engaging with these problems from differing perspectives. This includes researchers who are at the cutting edge of the use of quantitative text analysis on historical documents, with John Regan and Peter de Bolla aiming to show how historically-sensitive distributional text analysis can facilitate the recognition of conceptual architectures, and Pim Huijnen and Tom Kenter focusing on continuities and changes in conceptual structures over time. Mark Hill looks at ways one may sidestep the traditional methodological problems quantitative analysis may be introducing by turning to niche historical investigations. Daniele Guido will reflect on his experience with both developing tools to be used by historians, and the potential pitfalls which may lurk. Finally, Max Kemman will reflect on how historians collaborate in digital history projects. The goal, then, is to bring together these practitioners, contrast their differing technological and historiographical approaches, and ultimately offer some further reflections on how digital history facilitates an interaction between digital technology and historical practices.

John Regan, Paul Nulty, & Peter de Bolla - What distributional concept analysis tells us about the philosophical concept of 'negative liberty': A case study in the shadow of Quentin Skinner

In his 1984 essay 'The idea of negative liberty', Quentin Skinner gives a historical account of two opposing ideas. One is 'negative liberty', in which the individual's social freedom is guaranteed only by the absence of limiting factors such as state intervention, responsibilities to one's communities, and other externalities. In this scheme, liberty can only be defined negatively, as Thomas Hobbes has it at the start of his chapter 'Of the liberty of subjects' from *Leviathan*: 'liberty or freedom signifieth (properly) the absence of opposition.' Skinner contrasts this with an ideal of liberty in which the operative factor is the virtue and value of public service. This is to say, that one is only consummately free when one acknowledges one's social responsibilities and when one carries out virtuous acts of public service. These contrasting ideas of liberty are named by Canadian philosopher Charles Taylor as the 'opportunity concept' and the 'positive exercise concept'. The former relies purely on the absence of constraint and prescribed social objectives, where in the latter the individual attains liberty by acting positively in the service of the state or community.

This presentation will include data from the dataset *Early English Books Online* (EEBO) and *Eighteenth Collections Online* (ECCO) in order trace the emergence, stability and evolution of the concept of liberty from the seventeenth century to the end of the eighteenth. The software developed by the Cambridge Concept Lab is capable of interrogating these datasets in sophisticated computational and statistical ways, massively enhancing the capabilities of these data resources. We shall share these data and demonstrate the innovative power of our methodology for understanding the history of philosophy.

Pim Huijnen & Tom Kenter - What we talk about when we talk about concepts. Applying distributional semantics on Dutch historical newspapers to trace conceptual change

Word embeddings – vector representations of words that embed words in a so-called semantic space where the vectors of semantically similar words lie close together – are increasingly used for semantic searches in large text corpora. Word vector distances can be used to build semantic networks of words. This closely resembles the notion of semantic fields that humanities scholars are familiar with.

We have previously shown how word embeddings, as produced by a popular implementation word2vec, can be used to trace concepts through time without the dependency of particular keywords (Kenter et al. 2015). However, there are two main challenges that come with the use of word embeddings to represent concepts and conceptual change for the study of history. Firstly: commensurability. The use of computational techniques like word2vec demands choices of practical or technical nature. How do we legitimize these choices in terms of conceptual theory? Secondly: dependency on data. Do the results of word embedding techniques provide insights into real conceptual change, or do they merely reflect arbitrary biases in the underlying data?

Both challenges illustrate the need for critical reflection now that advanced computational tools are adopted in historical scholarship. Based on concrete examples, we will show how we dealt with these challenges in our research.

Mark Hill - Niche Analysis: Historical Methods, Digital Humanities, and Smaller Data

This paper aims to question the relationship between intellectual history – in particular, the Cambridge School (Skinner) – and new tools and techniques in quantitative text analysis.

Specifically, it asks whether one can extract contextually relevant and historically interesting information from a digital corpus via methods of distant reading used on specific historical contexts. That is to say, while the paper does not claim that others have not engaged with these issues (de Bolla 2013), it does aim to offer thoughts on an engagement from a different angle: to investigate the use of these methods in niche historical areas, places, and time.

This approach may provide a number of benefits. First, by limiting itself historically and geographically the project mitigates the problems of decontextualized analysis. Second, by directing these tools towards a topic which one already has an understanding of, the problems associated with 'distant reading' become less pronounced – a researcher can more easily verify or falsify algorithmic outputs (Betti and van den Berg 2014). Third, by working with a limited dataset one is able to construct a more accurate corpus (Bullard 2013, Spedding 2011). Forth, it may allow us to take sources which have been read (and re-read) for centuries and extract new information. Finally, as the project's scale is limited its methodological lessons are more easily replicable, and therefore of use to other scholars.

While the potential outputs are certainly limited when compared to some other projects, the author claims that there may nonetheless be a role for the analysis of smaller sets of data.

Daniele Guido - When It Fails

There are many accessible, low cost or free to use web tools and services that enable humanities researchers to engage with complex analytical algorithms. Backed by the powerful Dbpedia ontology, identifying people, locations or institutions in different media has never been more easy. Tools like DBpedia Spotlight (Mendes et al. 2011), YAGO/AIDA (Hoffart et al. 2011), Textrazor (Van Erp, Rizzo, and Troncy 2013) or Babelfy (Moro, Cecconi, and Navigli 2014) allow researchers to effectively recognize named entities in different languages. Moreover, if those entities are somehow present in DBpedia, these tools can resolve ambiguity relying to DBpedia internal linking. This analysis activity has been made available for images as well, at least for people and groups recognition - with Image Feature Extraction (IEF) (Fang et al. 2015). Both categories of services have been tested and integrated into Histograph (histograph.eu), a web platform for multimedia collection exploration conceived with historians. Histograph enables automatic enriching and annotating of texts and images for different services and for different languages. However, when exploring the results of entity recognition activities, the multitude of homonyms, misspellings and ambiguities makes it clear that alongside automatic tools there is a need for side processes that help us correct the results and reduce noise. This activity of spotting and correcting errors is time consuming work for the researcher, but having unexpected and usually "wrong" results makes him/her start to disbelieve in the expression "the more the information, the better the result", resulting in *diminishing trust* of the tool. In Histograph we adopted a mixed approach, combining a technological solution with design.

On one side, we disambiguate the results by comparing between sources in different languages, obtaining different perspectives and contexts. On the other side, interactive visualizations - which elsewhere have been shown to help digital humanities scholars to evaluate and interpret complex datasets (Jänicke et al. 2016), enable exploration and the identification of *patterns of failures*. By thus combining technology and design we hope researchers accept the failure of the tool whilst they simultaneously "love the bomb".

Max Kemman - Digital History Projects as Boundary Objects

Digital history as a subfield of the digital humanities constitutes a form of *methodological interdisciplinarity*; using methods, concepts, or tools from other disciplines to try to improve historical research (Klein 2014). However, as this panel demonstrates, this is not a straightforward process of taking something from another discipline and implementing it in historical research. Instead, what we

see is a negotiation of practices to align the new methods with the scholarly values of the discipline (Kaltenbrunner 2015). This negotiation regularly takes place in the context of a research project, where participants with different backgrounds work together on a shared problem. Yet despite working on a shared problem, the individual participants may still have different research goals and incentives to enter the collaboration. Although the research project defines a common research problem, how this research problem is or should be approached differs between the different collaborators dependent of, among other factors, their disciplinary background. This paper will therefore analyze the research project as *boundary object*, i.e., as an object that maintains a common identity among the different participants, yet is shaped individually according to disciplinary needs (Star and Griesemer 1989, Star 2010). In order to investigate how participants shape the research project and align the project with their scholarly values, we will look at the individual *incentives* for collaboration, following research by Weedman on incentives for collaborations between earth scientists and computer scientists (Weedman 1998). For several digital history projects, we will discuss collaborators' reasons for joining the project, their individual goals with the project, and the expected effects of the participation after the project has ended.

This research is part of a PhD research on how the interdisciplinary interactions in digital history have methodological and epistemological consequences for the practice of historians (Kemman 2016). By untangling the individual interests in digital history projects, we aim to gain better insight into how digital history functions as a coordination of practices between historians and collaborators from other disciplinary backgrounds.

Bibliographic References

Betti, Arianna, and Hein van den Berg. 2014. "Modelling the History of Ideas." British Journal for the History of Philosophy 22 (4). Informa UK Limited: 812–35. doi:10.1080/09608788.2014.949217.

de Bolla, Peter. 2013. The Architecture of Concepts. Fordham University Press. doi:10.5422/fordham/9780823254385.001.0001.

Bullard, Paddy. 2013. "Digital Humanities and Electronic Resources in the Long Eighteenth Century." Literature Compass 10 (10). Wiley-Blackwell: 748–60. doi:10.1111/lic3.12085.

Edelstein, Dan. 2015. "Intellectual History And Digital Humanities." Modern Intellectual History 13 (01). Cambridge University Press (CUP): 237–46. doi:10.1017/s1479244314000833.

Van Erp, Marieke, Giuseppe Rizzo, and Raphaël Troncy. 2013. "Learning with the Web: Spotting Named Entities on the Intersection of NERD and Machine Learning.." In # MSM, 27–30.

Fang, Hao, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao et al. 2015. "From captions to visual concepts and back." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473-1482.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21 (3). Cambridge University Press (CUP): 267–97. doi:10.1093/pan/mps028.

Guldi, Jo, and David Armitage. 2014. The History Manifesto. Cambridge University Press (CUP). doi:10.1017/9781139923880.

Hoffart, Johannes, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. 2011. "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages." In WWW.

Jänicke, S, G Franzini, MF Cheema, and G Scheuermann. 2016. "Visual Text Analysis in Digital Humanities." In Computer Graphics Forum. Wiley Online Library.

Kaltenbrunner, Wolfgang. 2015. "Reflexive Inertia: Reinventing Scholarship through Digital Practices." PhD thesis, Leiden University.

Kemman, Max. 2016. "Dimensions of Digital History Collaborations." In DHBenelux. Belval, Luxembourg.

Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. "Ad Hoc Monitoring of

Vocabulary Shifts over Time." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 1191–1200.
Klein, Julie Thompson. 2014. Interdisciplining Digital Humanities: Boundary Work in an Emerging Field. Online. University of Michigan Press. doi:10.3998/dh.12869322.0001.001.
Laver, Michael, Kenneth Beniot, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." American Political Science Review 97 (02). Cambridge University Press (CUP). doi:10.1017/s0003055403000698.
London, Jennifer A. 2016. "Re-Imagining the Cambridge School in the Age of Digital Humanities." Annual Review of Political Science 19 (1). Annual Reviews: 351–73. doi:10.1146/annurev-polisci-061513-115924.
Mendes, Pablo N, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. "DBpedia Spotlight: Shedding Light on the Web of Documents." In Proceedings of the 7th International Conference on Semantic Systems, 1–8. ACM.
Moretti, Franco. 2013. Distant Reading. Verso Books.
Moro, Andrea, Francesco Cecconi, and Roberto Navigli. 2014. "Multilingual Word Sense Disambiguation and Entity Linking for Everybody." In Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272, 25–28. CEUR-WS.org.
Schonhardt-Bailey, Cheryl. 2006. From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective. Mit Press.
Skinner, Quentin. 1984. "The Idea of Negative Liberty: Philosophical and Historical Perspectives." Philosophy in History. Cambridge University Press Cambridge, 193–221.
Skinner, Quentin. 2002. Visions of Politics. Cambridge University Press (CUP). doi:10.1017/cbo9780511613784.
Spedding, Patrick. 2011. "The New Machine: Discovering the Limits of ECCO." Eighteenth-Century Studies 44 (4). Johns Hopkins University Press: 437–53. doi:10.1353/ecs.2011.0030.
Wang, Shenghui, Stefan Schlobach, and Michel Klein. 2011. "Concept Drift and How to Identify It." Web Semantics: Science, Services and Agents on the World Wide Web 9 (3): 247–65. doi:10.1016/j.websem.2011.05.003.
Skinner, Quentin. n.d. "Meaning and Understanding in the History of Ideas." In Visions of Politics, 57–
89. Cambridge University Press (CUP). doi:10.1017/ccol0521581052.004.
Star, S.L., and J. R. Griesemer. 1989. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social Studies of Science 19 (3): 387–420. doi:10.1177/030631289019003001.
Star, S.L. 2010. "This Is Not a Boundary Object: Reflections on the Origin of a Concept." Science, Technology & Human Values 35 (5): 601–17. doi:10.1177/0162243910377624.
Weedman, Judith. 1998. "The Structure of Incentive: Design and Client Roles in Application-Oriented Research." Science, Technology & Human Values 23 (3): 315–45. doi:10.1177/016224399802300303.

POSTERS

Processing Data on Fake Inscriptions: How to Build the New Epigraphic Database Falsae (EDF)

Lorenzo Calvelli, Ca' Foscari University of Venice, Principal Investigator for the Italian Research Project of National Interest «False testimonianze. Copie, contraffazioni, manipolazioni e abusi del documento epigrafico antico», lorenzoc@unive.it

A New Research Project on Epigraphic Forgeries

As a part of its ongoing commitment to finance basic research in Italy, the Italian Ministry of Education, University, and Research (MIUR) recently approved a list of Research Projects of National Interest (PRIN) which includes a three-year investigation on fake ancient inscriptions entitled «False testimonianze. Copie, contraffazioni, manipolazioni e abusi del documento epigrafico antico». Directed by Lorenzo Calvelli, tenured lecturer in Roman History and Latin Epigraphy at Ca' Foscari University of Venice, the project involves scholars from the universities of Bari, Bologna, Genoa, Macerata, Milan, Pisa, Rome, Trieste, Turin, Venice, and Verona with at least an additional five post-doctoral positions for young scholars. It will last from February 2017 to January 2020.

The project has 4 main goals: (1) to create an electronic archive of fake inscriptions called the EDF (Epigraphic Database Falsae); (2) to establish a regular schedule of work-in-progress meetings; (3) to organize an international conference on epigraphic forgeries whose proceedings will be published; and (4) to set up a temporary travelling exhibition dedicated to the subject.

The New Electronic Archive EDF (Epigraphic Database Falsae)

This poster addresses the preliminary steps related to the construction of the EDF, which is currently in an embryonic stage. The main reason why such a tool has not been created previously is the absence of a scholarly consensus as to what should be considered a fake inscription. Existing sections devoted to forgeries in the principle printed collections of ancient inscriptions group together large and ambiguous sets of documents which include intentional forgeries, copies of ancient inscriptions, and Medieval and Renaissance inscriptions that imitate classical models. Further confounding the question, intentional forgeries are most frequently written solely on paper (i.e. they are quoted in manuscripts or printed books), although they can also be incised on actual objects including ancient or only partially ancient materials as well as more recent artefacts.

In response, it has been decided that the EDF will compile data from multiple sources and concern all known typologies of fake inscriptions, including transcriptions, reproductions, and facsimiles of thousands of forgeries. Eventually, the electronic archive is meant to encompass all forged epigraphic texts that originate from the whole territory of Italy and, it is hoped, will extend its geographic range to include the rest of Europe and the Mediterranean, provided that partnerships with foreign and international institutions are created.

The Core Structure of EDF

Experts of the digitization of epigraphic documents are currently building the core of the EDF which will soon be accessible through a search engine. We aim to get feedback from the users of the World Wide Web at a very early stage of the project, in order to further improve the structure of the database.

Tentatively, the search engine will allow users to perform queries in the following fields:

1) textual typologies, distinguishing between actual forgeries (i.e. invented texts), post-classical

inscriptions, and copies of genuine ancient inscriptions. The latter will be subdivided into full copies, partial copies, and interpolated copies. In the case of Christian inscriptions it will also be possible to identify "Christianized" pagan monuments and later imitations of early Christian inscriptions;
2) the modes of transmission, distinguishing between forgeries written exclusively on paper and those that were actually incised on durable materials (stone, metal, etc.);
3) the methods of production, distinguishing between forgeries invented wholly from complete, partial, or interpolated copies of ancient inscriptions;
4) the intentions of the forgers, distinguishing between commercial forgeries (i.e. inscriptions produced with the intent to earn profit) and historical and documentary forgeries (i.e. documents fabricated with the intent of validating a certain historical statement, whether true or false);
5) the age when the forgery was created;
6) the identity of the forgers, certain or suspected, whenever possible;
7) the production site of actual forgeries;
8) the subsequent displacements of these forgeries;
9) their current location (whenever the objects are still traceable).

By combining the last three sets of data on an interactive map, it will be possible for users to visualize the locations of the workshops where forgeries were physically produced, to trace the routes of their dispersal, and to follow the steps that brought them to their current place of conservation. This kind of information, which is entirely absent in the indices of the main printed epigraphic corpora, will be made available through a GIS (Geographic Information System) innovative tool.

Interaction with existing online epigraphic resources

Visually, the EDF will be structured so as to be readily understandable and accessible in multiple languages in order to comply with the principles of the inclusive and intercultural approach to information promoted by the European Union. In the end, it will be an unrestricted online resource, freely accessible to multiple users: scholars from different countries, students, curators and visitors of local museums, national and international institutions committed to the protection and conservation of cultural heritage, and qualified professionals of the antiquities trade. The EDF will also be reachable through the new EAGLE shared portal (European network of Ancient Greek and Latin Epigraphy: www.eagle-network.eu), and it is planned that it will properly interact with the main existing online resources related to epigraphy, in particular the EDR (Epigraphic Database Roma: www.edr-edr.it) and the EDB (Epigraphic Database Bari: www.edb.uniba.it), which have as their aim the inclusion of all published Greek and Latin inscriptions (both pagan and Christian) from ancient Italy dating from before the 7th century AD.

In conclusion, the EDF will stimulate historical research by presenting previously neglected sources using up-to-date technologies. Digital data will be made available to scholars of different disciplines and will offer to the international academic community an enlarged documentary base for further research. Furthermore, the electronic archive will also facilitate the dissemination of knowledge by combining data on epigraphic forgeries with hyperlinks to other digital libraries and by incorporating already existing online data into its own records. Thanks to its unrestricted access and its intelligible structure, the database will also be accessible to non-specialist users, whose better comprehension of the world of forgery will also lead to a more critical awareness of the use of data and information in everyday life.

Selected Bibliography

Buonopane, Alfredo. 2014. "Il lato oscuro delle collezioni epigrafiche: falsi, copie, imitazioni." In Donati, *L'iscrizione e il suo doppio*, 291-314.
Carbonell Manils, Joan, Gimeno Pascual, Helena, and Moralejo Álvarez, José Luiz, eds. 2011. *El monumento epigráfico en contextos secundarios: procesos de reutilización, interpretación y falsificación*. Barcelona: Universitat Autònoma de Barcelona, Servei de Publicacions Bellaterra.
Donati, Angela, ed. 2014. *L'iscrizione e il suo doppio. Atti del Convegno Borghesi 2013*. Faenza: Fratelli Lega.
Orlandi, Silvia, Caldelli, Maria Letizia, and Gregori, Gian Luca. 2015. "Forgeries and Fakes." In *The Oxford Handbook of Roman Epigraphy*, edited by Christer Bruun and Jonathan Edmondson, 42-65. Oxford - New York: Oxford University Press.
Solin, Heikki. 2012. "Falsi epigrafici." In *L'officina epigrafica romana: in ricordo di Giancarlo Susini*, edited by Angela Donati and Gabriella Poma, 139-151. Faenza: Fratelli Lega.
Solin, Heikki. 2014. "Falsi epigrafici II." In Donati, *L'iscrizione e il suo doppio*, 227-242.
Vagenheim, Ginette. 2011. "La falsificazione epigrafica nell'Italia della seconda metà del Cinquecento. *Renovatio* ed *inventio* nelle *Antichità romane* attribuite a Pirro Ligorio." In Carbonell Manils, Gimeno Pascual, and Moralejo Álvarez, *El monumento epigráfico*, 217-226.

Il modello a microkernel di Omega nello sviluppo di strumenti per lo studio dei testi: dagli ADT alle API

Angelo Mario Del Grosso, ILC-CNR Pisa, angelo.delgrosso@ilc.cnr.it
Emiliano Giovannetti, ILC-CNR Pisa, emiliano.giovannetti@ilc.cnr.it
Simone Marchi, ILC-CNR Pisa, simone.marchi@ilc.cnr.it

Introduzione

In questo contributo si illustra il lavoro metodologico e ingegneristico in corso presso il gruppo di Literary Computing dell'Istituto di Linguistica Computazionale (ILC) del CNR di Pisa nella progettazione e nella implementazione di una piattaforma per lo studio dei testi. Tale piattaforma si basa su un'architettura minimale (microkernel) dotata di strutture dati e di funzionalità di base. Il microkernel è capace di gestire moduli, altamente disaccoppiati, attraverso cui fornire i servizi orientati al trattamento (semi-)automatico di opere letterarie, con particolare attenzione al textual scholarship e all'analisi linguistica, semantica e lessicale. Nello specifico, si mostreranno alcuni strumenti sviluppati in collaborazione con altri istituti di ricerca nel contesto di vari progetti di Digital Humanities (DH). Verranno, altresì, introdotte gli strumenti realizzate nel corso del progetto Clavius on the Web.

L'uso delle tecnologie informatiche e dei sistemi digitali per lo studio scientifico e l'analisi di testi letterari ha sortito, negli ultimi anni, un duplice effetto: da un lato, una ingente disponibilità di risorse e di dati testuali in formati standard, aperti e machine-actionable e, dall'altro, lo sviluppo di complesse procedure in grado di elaborare automaticamente tali risorse, producendo nuova conoscenza e nuove evidenze. Tuttavia, i dati e i servizi digitali sono spesso carenti in progettazione e in modellazione. In particolare - alla luce della nostra esperienza - raramente si affronta la specificazione di importanti aspetti legati al contesto applicativo di interesse, all'integrazione e all'estensione dei modelli sviluppati (sia dati, sia software). Queste caratteristiche, infatti, se prese in considerazione, facilitano sensibilmente la riusabilità e la flessibilità degli strumenti computazionali implementati. In più, la definizione di astrazioni e casi d'uso in collaborazione con gli esperti di dominio permette di conservare l'originale contesto applicativo, talvolta snaturato dall'impiego di strumenti sviluppati per altri scopi (a tal proposito si veda, tra gli altri, (Ciotti 2016) per una panoramica critica sull'uso del "topic modeling" nell'ambito del literary computing). Infatti, come sottolineato da Franz Fischer in un suo noto contributo: "There is no out-of-the-box software available for creating truly critical and truly digital editions at the same time" (Fischer 2013). Della stessa idea - aggiungendo anche una nota di urgenza - è la recente affermazione di Elena Pierazzo, la quale indica che "da molte parti si lamenta la mancanza di software e strumenti facili da usare e che possano limitare la necessità da parte degli editori di fare tutto da soli [...], ci si potrebbe chiedere come mai con quasi 30 anni di ricerca nel settore delle edizioni digitali ci siano a tutt'oggi un numero così limitato di strumenti di tale genere" (Pierazzo 2016). Allo stesso modo, Monica Berti indica come sia molto sentita l'esigenza di sviluppare un modello per rappresentare anche citazioni e "text reuses" in un ambiente digitale (Berti et al. 2014).

Lavori simili

Nel corso degli ultimi anni, alcuni importanti progetti di ricerca si sono prefissati l'obiettivo di realizzare modelli generali, servizi digitali e strumenti online integrabili in una infrastruttura a lungo

termine sia per la gestione e l'analisi di edizioni digitali sia per soddisfare le necessità degli studiosi di opere culturali.

In questa visione il progetto Open Philology - e in particolare Bridget Almas - ha, tra le proprie finalità, quella di sviluppare una piattaforma integrata, denominata Perseids, per la trascrizione collaborativa, l'editing e la traduzione di documenti storici (Almas and Beaulieu 2013).

Tra le altre iniziative interessanti che reputiamo allo stato dell'arte nel settore dell'analisi testuale citiamo: (a) Textual Community project, i cui responsabili scientifici sono Peter Robinson and Barbara Bordalejo (Robinson and Bordalejo 2016), (b) il progetto AustESE inserito all'interno del gruppo australiano eResearch (Shillingsburg 2015), (c) il progetto Tagore Online Variorum "Bichitra" di Sukanta Chaudhuri (Chaudhuri 2015), (d) il progetto Homer Multitext, i cui responsabili scientifici sono Neel Smith e Christopher Blackwell (Smith 2010), (e) il progetto Sharing Ancient Wisdoms finanziato dal network HERA (Hedges et al. 2016). E' in questo scenario che si inserisce il nostro lavoro, attraverso il quale si incoraggia un processo di progettazione e sviluppo maggiormente orientato alla definizione di tipi di dato astratti e di interfacce di programmazione condivise.

Metodo: l'approccio della piattaforma Omega

Il gruppo di Literary Computing dell'Istituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa (ILC-CNR) ha intrapreso una linea di ricerca volta alla progettazione di modelli software per lo studio scientifico del testo. La parte di indagine più teoretica è costantemente affiancata da un'attività implementativa che ha portato, negli ultimi anni, alla realizzazione di strumenti software prototipali in grado di favorire il lavoro di indagine critico-letteraria attraverso approcci statistico-computazionali (Giovannetti et al. 2016, Benotto et al. 2016, Bozzi 2013). Questa esperienza di ricerca ci ha permesso, oggi, di intraprendere un percorso di sviluppo più solido e generale per la realizzazione di un ambiente digitale, chiamato Omega, in grado di rispondere alle esigenze proprie degli studiosi di documenti storico-letterari (Del Grosso et al. 2016). A tale scopo, questo framework è progettato seguendo il paradigma orientato agli oggetti integrando tecnologie del web semantico. L'ambiente di studio qui proposto è intrinsecamente modulare grazie alla sua natura object-oriented, alla definizione di dettagliati casi d'uso e all'impiego del Domain-Driven Design (Evans 2014) che garantisce una stretta collaborazione tra i progettisti e gli esperti di dominio.

Similmente al paradigma architetturale di MINIX, sistema operativo sviluppato da Andrew Tanenbaum (Tanenbaum 2014, 63-68), Omega poggia su un'architettura a microkernel, concepita per favorire l'estensibilità e la flessibilità delle sue varie componenti. L'approccio di modellazione adottato, inoltre, impiega alcuni tra i più noti Design Pattern di progettazione e programmazione orientata agli oggetti (Factory, Visitor, Composite Component, ecc.) per l'impiego di soluzioni generiche a problemi ricorrenti nel dominio dello studio del testo (Ackerman and Gonzalez 2011).

Uno dei principali risultati dell'attività di progettazione è rappresentato dalla individuazione di Tipi di Dato Astratti (ADT) definiti formalmente attraverso interfacce di programmazione (API, Application Programming Interfaces) (Martini 2016). Gli ADT in questione costituiscono delle astrazioni che combinano la rappresentazione interna dei dati testuali con le elaborazioni definite su di essi in una singola unità di riferimento atomica (DS-ADT).

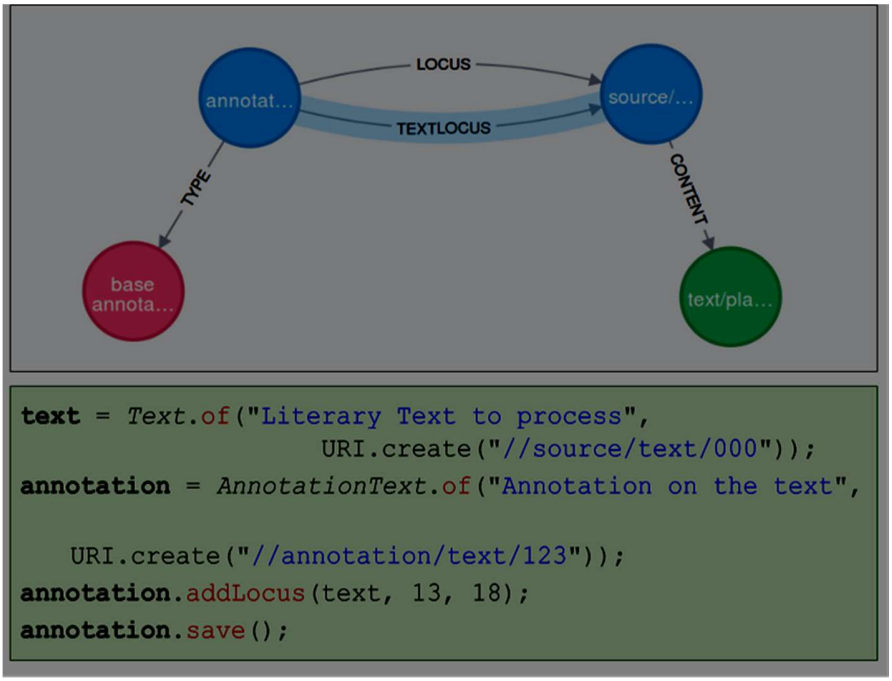


Figura 1. Esempio di ADTs e di APIs definite attraverso un approccio guidato dal dominio (DS-ADT e DS-API).

Le API, invece, sono modellate e descritte a partire dai requisiti del dominio di interesse (DS-API) e sono indipendenti da qualsiasi dettaglio tecnologico-implementativo (Del Grosso et al. 2016). Queste, inoltre, rappresentano i servizi forniti dalla piattaforma ai vari utenti con diversi livelli di granularità e scalabilità. In tal modo è stato possibile organizzare l'infrastruttura attraverso diversi "layers" di astrazione: (a) low-level (Class API); (b) middle-level (Module API); (c) high-level (Web API) (Knoernschild 2012). Per esempio, il servizio di analisi linguistica ad oggi implementato nella piattaforma è costituito da un modulo di analisi dei testi in latino (Module API) che si compone di varie unità di elaborazione come la tokenizzazione o l'individuazione di entità nominate (Class API) che possono essere invocate tramite chiamate RESTful da client remoti (Web API).

Nel frammento di codice in Figura 1 sono riportati la rappresentazione del modello dei dati e l'uso di middle-level API relative al caso d'uso per la creazione e la persistenza di una annotazione del testo. In particolare l'API permette di: a) istanziare un oggetto *text* con contenuto e relativo URI, b) istanziare una *annotation* dotata di contenuto e URI, c) associare l'annotazione al testo tramite un oggetto di tipo *Locus* e, infine, d) memorizzare l'oggetto annotazione in un database all'uopo costituito (Del Grosso et al. 2016b). Il caso d'uso (*what*) e la relativa API (*how*) qui introdotti sono stati sviluppati tenendo in considerazione i modelli standard per l'annotazione di risorse testuali maggiormente diffusi, quali, ad esempio, l'Open Annotation Collaboration (OAC)[1]. Questa attenzione agli standard e ai modelli esistenti (come ad esempio TEI per la creazione di edizioni digitali), presente nello sviluppo di tutti i vari casi d'uso che saranno definiti e implementati, consentirà di delineare scenari di possibili usi esterni con gli altri strumenti fino ad ora sviluppati.

Caso di studio: Il progetto Clavius on The Web

L'approccio metodologico di progettazione software per lo studio e il trattamento (anche automatico) del testo letterario, descritto precedentemente, è stato adottato nel progetto "Clavius on the Web",

1 L'iniziativa OAC è adesso confluito nel progetto W3C del Web Annotation Data Model https://goo.gl/ovxLG1

inaugurato nel 2012 e finanziato dal Registro.it. Il progetto vede coinvolti l'Istituto di Linguistica Computazionale "A. Zampolli" (ILC) e l'Istituto di Informatica e Telematica (IIT), entrambi del CNR di Pisa, in collaborazione con l'Archivio Storico della Pontificia Università Gregoriana (APUG) di Roma (Abrate et al. 2014).

Lo scopo di questa iniziativa è quello di preservare, valorizzare e promuovere la corrispondenza di Cristoforo Clavio (1537-1612) intercorsa tra il Gesuita ed importanti scienziati coevi (tra cui Galileo Galilei e Tycho Brahe) .

Clavius è stato un noto matematico e astronomo, conosciuto soprattutto per il suo contributo nella riforma del calendario gregoriano, per il lavoro editoriale sugli Elementi di Euclide e per il commentario all'opera *De sphaera mundi* di Giovanni Sacrobosco. I manoscritti analizzati nel progetto, contenenti la corrispondenza di Clavius (in lingua latina e in lingua italiana), sono conservati presso l'APUG.

Durante le attività del progetto, grazie allo sforzo congiunto dei partner, sono stati sviluppati strumenti Web di analisi ed elaborazione automatica del testo. Queste applicazioni permettono agli studiosi di trascrivere il contenuto testuale dei manoscritti acquisiti digitalmente, di elaborarne automaticamente i dati linguistici e, infine, di annotare lessicalmente e semanticamente il corpus gestito dal sistema (Figura 2).

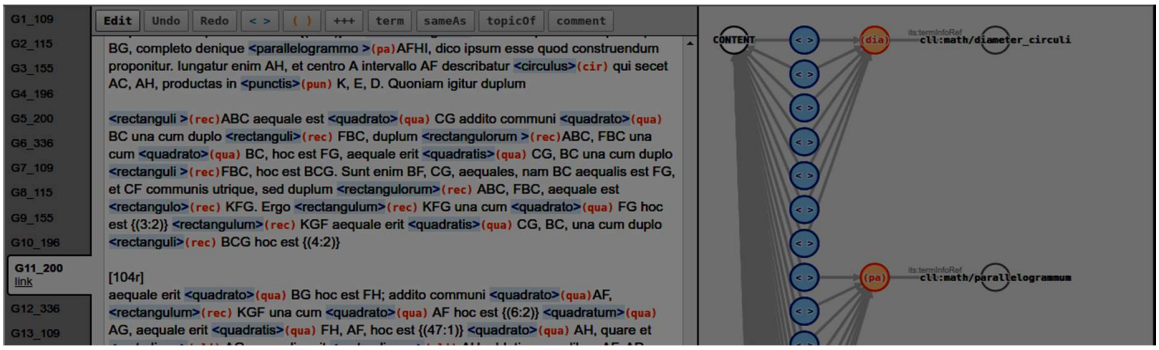


Figura 2. Esempio di trascrizione e annotazione semantica di una lettera a Cristoforo Clavio

In virtù dell'approccio modulare e flessibile utilizzato nelle fasi di progettazione dell'ambiente è stato possibile adattare le componenti software anche per una iniziativa collaterale: Clavius@School (Marchetti et al. 2015). Questa attività ha visto il coinvolgimento di studenti liceali impegnati nell'uso degli strumenti sviluppati in "Clavius on the Web", consentendo loro, inoltre, di approfondire la consapevolezza delle potenzialità e dei rischi connaturati al paradigma educativo digitale. Sintetizzando, il lavoro di ricerca e sviluppo condotto nel progetto ha portato alla realizzazione di diverse componenti che saranno integrate nell'ambiente Omega sfruttandone l'architettura a microkernel già descritta.

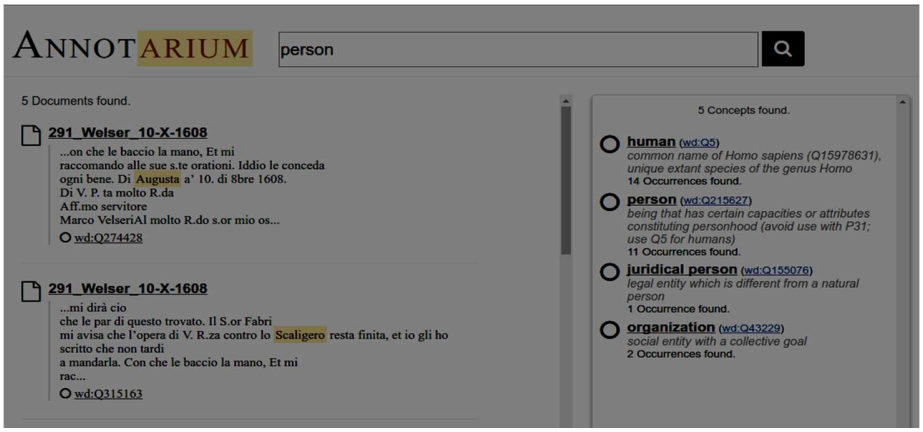


Figura 3. *Annotarium*: sistema di interrogazione full-text e ricerca semantica.

Di seguito si elencano i principali strumenti - accompagnati da alcune immagini dell'interfaccia grafica (Figura 3, Figura 4 e Figura 5) - che ne riassumono lo stato di avanzamento (Piccini et al. 2016, Valsecchi et al. 2016):

- l'archivio digitale: claviusontheweb.it/
- il Text Encoder and Annotator (TEA): wafi.iit.cnr.it/webvis/tea/
- lo strumento di ricerca Annotarium: wafi.iit.cnr.it/webvis/annotarium/
- l'annotazione linguistica: claviusontheweb.it/annotations/annotations.php?letter_id=147
- l'interfaccia di interrogazione in linguaggio controllato: http://licodemo.ilc.cnr.it:8080/clavius
- il lessico specialistico e l'ontologia: https://goo.gl/MXtFdy

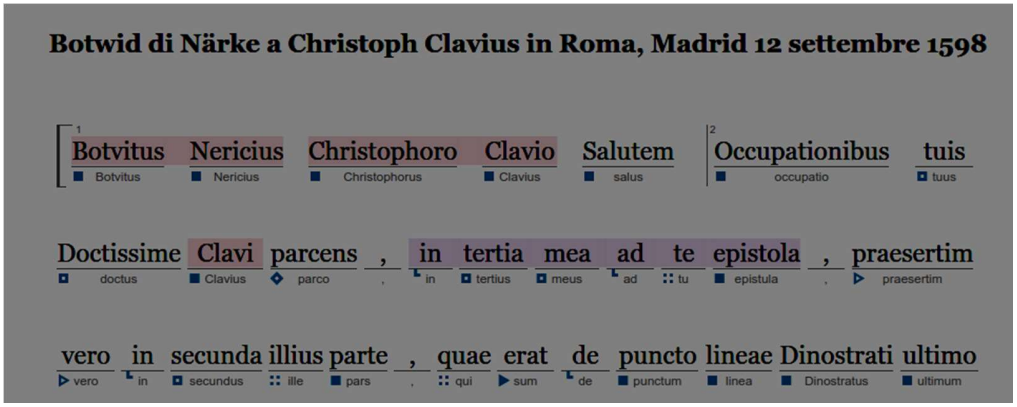


Figura 4. Visualizzazione dell'analisi semiautomatica della lingua e delle entità nominate.

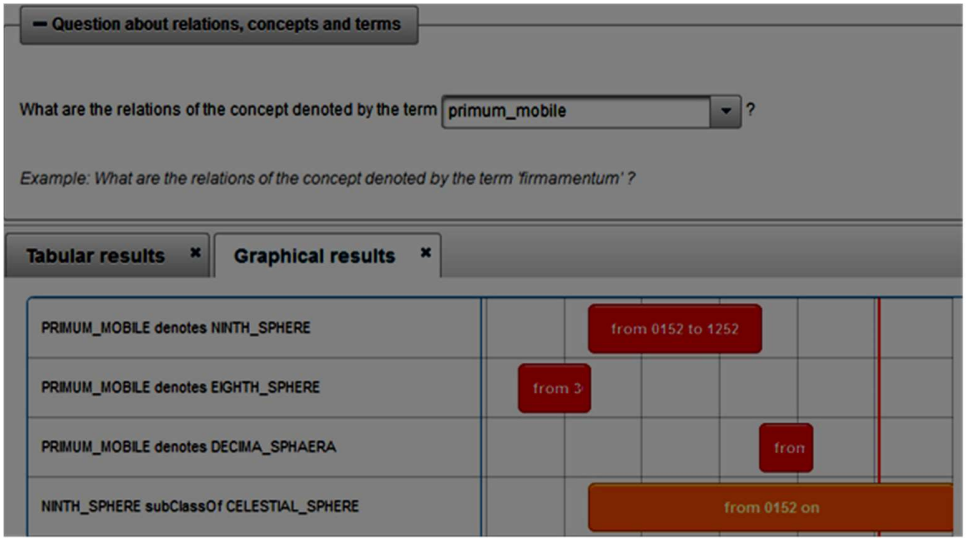


Figura 5. L'interfaccia per l'interrogazione della risorsa termino-ontologica diacronica di Clavius on the Web.

Conclusioni

In questo contributo sono descritti i progressi effettuati nella progettazione e nello sviluppo di Omega, una piattaforma per lo studio scientifico del testo. Sono stati illustrati alcuni risultati ottenuti nell'ambito del progetto Clavius on the Web che dimostrano quanto sia cruciale l'adozione di approcci robusti e flessibili nella realizzazione di strumenti digitali, soprattutto nei casi in cui si necessiti di riadattare tali strumenti per compiti diversi.

Lo sviluppo della piattaforma Omega è disponibile su Github al seguente indirizzo: https://github.com/literarycomputinglab[2]. Al momento sono in corso le attività di definizione di ulteriori API (a partire dalla gestione di lessici e dalla rappresentazione digitale di risorse testuali) e di documentazione delle singole API già sviluppate e della piattaforma generale. Inoltre, sono in corso di studio le modalità per il rilascio del software all'interno di altre infrastrutture di ricerca, come CLARIN e DARIAH. Come anticipato in contributi precedenti, sarebbe per noi di fondamentale importanza, soprattutto in questa fase dello sviluppo, ricevere feedback dalla comunità sui modelli e gli approcci che abbiamo deciso di impiegare per la realizzazione del framework.

Riferimenti Bibliografici

Abrate, Matteo, Angelo Mario Del Grosso, Emiliano Giovannetti, Angelica Lo Duca, Damiana Luzzi, Lorenzo Mancini, Andrea Marchetti, Irene Pedretti, and Silvia Piccini. 2014. "Sharing Cultural Heritage: The Clavius on the Web Project." In *LREC2014 Proceedings*, 627–634. ELRA.

Ackerman, Lee, and Celso Gonzalez. 2011. *Patterns-Based Engineering: Successfully Delivering Solutions Via Patterns*. Addison-Wesley.

Almas, Bridget, and Marie-Claire Beaulieu. 2013. "Developing a New Integrated Editing Platform for Source Documents in Classics." *LLC* 28 (4): 493–503.

Benotto, Giulia, Emiliano Giovannetti, and Ouafae Nahli. 2016. "An Application of Distributional Semantics for the Analysis of the Holy Quran." In *CIST2016 Proceedings*. Tangier-Assilah, Morocco.: IEEE.

Berti, Monica, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova, and Gregory Ralph Crane. 2014. "The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors." *JTEI* 8. doi:10.4000/jtei.1218.

Bozzi, Andrea. 2013. "G2A: A Web Application to Study, Annotate and Scholarly Edit Ancient Texts and Their Aligned Translations." Edited by ERC Ideas 249431. *Studia Graeco-Arabica* 3: 159–171.

Chaudhuri, Sukanta. 2015. *Bichitra: The Making of an Online Tagore Variorum*. 1sted. Springer Publishing Company, Incorporated.

Ciotti, Fabio. 2016. "What's in a Topic Model. I Fondamenti Del Text Mining Negli Studi Letterari." In *DH2016 Abstracts*, 149–51. Kraków.

Del Grosso, Angelo Mario, Federico Boschetti, Emiliano Giovannetti, and Simone Marchi. 2016. "Vantaggi dell'Astrazione Attraverso l'Approccio Orientato Agli Oggetti per Il Digital Scholarly Editing." In *AIUCD2016 Abstracts*, 213–18. Venezia.

Del Grosso, Angelo Mario, Davide Albanesi, Emiliano Giovannetti, and Simone Marchi. 2016. "Defining the Core Entities of an Environment for Textual Processing in Literary Computing." In *DH2016 Abstracts*, 771–75. Kraków.

Evans, Eric. 2014. *Domain-Driven Design Reference: Definitions and Pattern Summaries*. Dog Ear Publishing.

Fischer, Franz. 2013. "All Texts Are Equal, But… Textual Plurality and the Critical Text in Digital Scholarly Editions." *Variants* 10: 77–92.

Giovannetti, Emiliano, 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*. doi:10.1093/llc/fqw054.

2 Chiunque voglia contribuire alla implementazione di Omega è invitato ad unirsi al team di sviluppo su GitHub.

Hedges, Mark, Anna Jordanous, Faith Lawrence, Charlotte Roueche, and Charlotte Tupman. 2016. "Digital Publication of Highly Interconnected Manuscripts: Sharing Ancient Wisdoms." *The Journal of Data Mining and Digital Humanities*.
Knoernschild, Kirk. 2012. *Java Application Architecture: Modularity Patterns with Examples Using OSGi*. 1sted. Robert C. Martin Series. Prentice Hall.
Marchetti, Andrea, Martin Maria Morales, and Simone Marchi. 2015. "Il Progetto Clavius on the Web Entra Nelle Scuole." *Digitalia*, 74–84.
Martini, Simone. 2016. "Types in Programming Languages, between Modelling, Abstraction, and Correctness." In *CiE2016 Proceedings*, 164–69. Parigi: Springer.
Piccini, Silvia, Matteo Abrate, Clara Bacciu, Andrea Bellandi, Emiliano Giovannetti, Lorenzo Mancini, and Andrea Marchetti. 2016. "When Traditional Ontologies Are Not Enough: Modelling and Visualizing Dynamic Ontologies in Semantic-Based Access to Texts." In *DH2016 Abstracts*, 653–55. Kraków.
Pierazzo, Elena. 2016. "Quale Futuro per Le Edizioni Digitali? Dall'haute Couture Al Prêt-À-Porter." In *AIUCD2016 Abstracts*, 51–52. Venezia.
Robinson, Peter, and Barbara Bordalejo. 2016. "Textual Communities." In *DH2016 Abstracts*, 876–77. Kraków.
Shillingsburg, Peter. 2015. "Development Principles for Virtual Archives and Editions." *Variants* 11: 9–28. doi:10.1163/9789401212113_002.
Smith, Neel. 2010. "Digital Infrastructure and the Homer Multitext Project." In *Digital Research in the Study of Classical Antiquity*, 121–37. Farnham: Ashgate.
Tanenbaum, Andrew. 2014. *Modern Operating Systems*. 4th ed. Pearson Education, Inc.
Valsecchi, Fabio, Matteo Abrate, Clara Bacciu, Silvia Piccini, and Andrea Marchetti. 2016. "Text Encoder and Annotator: An All-in-One Editor for Transcribing and Annotating Manuscripts with RDF." In *ESWC2016 Proceedings*, 9989:399–407. Lecture Notes in Computer Science. Heraklion: Springer.

Reading Habits: Marginalia in Author's Libraries (Interpretation, Encoding, and Digital Publication)

José Luis Losada Palenzuela, University of Wroclaw (Poland), jose-luis.losada@uwr.edu.pl

Among the books that once belonged to Arthur Schopenhauer, including his Spanish books, we had found handwritten annotations, reading marks and drawings (Hübscher 1968; Losada 2011). Most of the annotations are underlining, vertical lines in the margin, cross references, etc. We refer to a specific type of annotation within an author's personal library: the author's marginalia (Jackson 2001). The corpus of annotations and marks on books are in a few cases the focus of a critical edition, as part, for example, of the critical apparatus; they also are included in typical studies of author's reception, but normally they are excluded in the textual scholarship. Given the recent impulse of the cultural turn, many of them has been recollected to examine the "potential value of readers' notes for historical studies of reception and reader response" (Jackson 2001: 6). All marginalia types share, though, a common perspective: the reading process and the traces left in the text.

Very few are digitized and disseminated, due to the complexity of the editions; even for the new digital approach they present a challenge, because they are frequently linked to the context in which they appear, e.g., the materiality of the book. Only scanning the container (scans & metadata) does not facilitate their diffusion due to the lack of contextualization, although it helps to locate them. A selection of Schopenhauer's digitized books (not all of his Spanish books are digitized) is available at the Frankfurt University Library digital catalog dedicated to the Schopenhauer Archiv < http://sammlungen.ub.uni-frankfurt.de/schopenhauer >.

A few institutions have undertaken major research projects for encoding and displaying annotations or scribblings of other authors' printed works, such as, the digital project "Digitizing Walt Whitman's annotations and marginalia" — within the vast project The Walt Whitman Archive < www.whitmanarchive.org >. Also in the United States we find "Melville's Marginalia Online", which is a virtual archive of books owned and borrowed by Herman Melville. The project supplies links to digital copies of Melville's books accompanied by bibliographical descriptions and documentary notes, as well as documentation and transcriptions of marginalia < http://melvillesmarginalia.org >. In Europe, The Beckett Digital Manuscript Project < www.beckettarchive.org > has collected manuscripts and transcriptions of Samuel Beckett to facilitate the research within the methodology of genetic criticism. Marks and notes are exhaustively referenced in the digitized library, and in some cases some marked passages have been selectively transcribed. The project offers a unique opportunity to learn how Becket, who called himself "phrase-hunting" (Hulle and Nixon, 2013: XV-XVI) interacted with English, French, and Italian literature classics, showing that he was, as well as Arthur Schopenhauer, a great polyglot reader. These major academic enterprises, with a extended economical and personal support and institutional commitment, show us that the analysis (interpretation), the digital mark-up (encoding) and the scholarly digital edition (diffusion) of marginalia is a necessary task which may contribute to the classical text interpretation, to the theory of reading, as well as to the current digital edition procedures.

The analysis of Schopenhauer's marginalia reveals reading processes in several ways, such as types of interaction (writing comprehension, translation, studying), influences, personal and creative author's perspectives. In order to set a fully editorial criteria for the encoding, we have sort out Schopenhauer's typical styles of marking: formal (underlining, interlinear gloss, vertical lines, exclamation marks), intentional (appreciative comments, indexation, cross references), material (errors, variations, amendments), etc.

Although the project will offer transcription and facsimile, the goal is to transcribe wholly or partially

some of the original text in books containing annotations. The TEI modules related to the encoding of printed resources need to be used, but once the taxonomy of the different types of annotation has been set, the TEI modules normally used for the manuscript modeling are also taken into consideration and the different specifications evaluated: highlighting and quotation; additions, deletions, and omissions; substitutions; writing, decoration; spans and interpretations. Some of the elements and attributes that can be used for the transcription of marginalia are <add>, <gloss>; <note>; ; <addSpan>; <anchor/>; <gloss>; <handNotes>, @type, @place, @rendition; @rend; @hand, etc., but we have decided to use mostly the <add> element, which "contains letters, words, or phrases inserted in the source text by an author, scribe, or a previous annotator or corrector" (TEI, P5: 3.0.0) adding an attribute selection of @type and @subtype (also @hand, @place, @medium…), which allows us to cover all the taxonomy of annotations, e.g.:

```
<add type="Glosse" subtype="index"> Cross reference to a page number
<add type="Glosse" subtype="statement"> Gloss on a passage
<add type="Korrektur" subtype="union"> Amend which joins to separated words.
<add type="Korrektur" subtype="punctuation"> Modification of the original punctuation
<add type="Randstreichung" subtype="triple"> Triple lateral mark on the page margin
<add type="Unterstreichung"> Underlining.
```

We are aware of the fact that marginalia are a sort of textual manifestation prone to cause problems with elements overlapping in XML (Schmidt 2010: 343-344). The final editorial decision has reflected about other encoding criteria, e.g, the possibility of using <note @resp>, <gloss> or <hi @rend>, which however could not solve the potentially overlap issues. It may be argued that there are indeed other ways to get around it. The <add> element has itself a variant in the empty element <addSpan/> (used together with @spanTo, @xml:id and <anchor/>); <hi> could be replace with a simple ; for deletions reaching over different elements there is the <delSpan> element. But after having set the taxonomy in our corpus, we did not find the need of using them.

Our main goal of the project is to be consistent in the set of elements, so that with <add> we could describe all external marks added to a printed edition during the lecture. Furthermore the editorial criteria considered also the visualization and publication, because as Elena Pierazzo (2015: 107) puts it, “determining what a digital edition should do will determine the theoretical model that should lie at its base”.

One fundamental idea leading this editorial project comes also from the need for a complete reference model for digital editing, where analysis, modeling, transcription, encoding, visualization and publication can be simplified in order to facilitate to less experienced (digital) editors, or with less resources, the control of their editorial work, generally planed as an individual project and normally without a strong technical support.

The complexity of editorial decisions taken in digital scholarly editing from the modeling until the publication can be overwhelming. Every criteria should take into consideration the nature of the documents, capabilities of the publishing technology, costs and time, etc. (Pierazzo 2015: 107). This thought is again behind the decision of using the TEICHI module for Drupal, a digital publishing framework, which helps to overcome easily (for our needs) the barriers between the encoded text and the online publication. TEICHI is a modular tool for displaying documents encoded according to the guidelines of the Text Encoding Initiative (TEI Lite P5) as pages in a Drupal-based website < www.teichi.org >. Although it is not an out of the box publishing solution, the module allows, briefly put, to use the XSLT capabilities to process TEI/XML within the Drupal environment.

This poster presents an ongoing project < http://www.schopenhauer.uni.wroc.pl > and aims to show the utility of such digital edition, the criteria used for the encoding (XML/TEI) and the framework for the publication (TEICHI).

Bibliographic References

Hübscher, Arthur (ed.). 1966-1975. *Der handschriftliche Nachlaß*. I-V. Frankfurt am Main: Waldemar Kramer.
Hulle, Dirk van, and Mark Nixon. 2013. *Samuel Beckett's Library*. Cambridge: Cambridge University Press.
Jackson, Heather Joanna. 2001. *Marginalia: Readers writing in Books*. New Haven: Yale University Press.
Losada Palenzuela, José Luis. 2011. *Schopenhauer traductor de Gracián. Diálogo y formación*. Valladolid: Servicio de Publicaciones de la Universidad de Valladolid.
Pape, Sebastian, Christof Schöch, and Lutz Wegner. 2012. 'TEICHI and the Tools Paradox. Developing a Publishing Framework for Digital Editions'. *Journal of the Text Encoding Initiative*, 2. doi:10.4000/jtei.432.
Pierazzo, Elena. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate Publishing, Ltd.
Schmidt, Desmont. 2010. 'The Inadequacy of Embedded Markup for Cultural Heritage Texts'. *Literary and Linguistic Computing* 25 (3): 337–56. doi:10.1093/llc/fqq007.
TEI Consortium. P5: *Guidelines for Electronic Text Encoding and Interchange*. Version 3.0.0. Last updated on 29th March 2016. < www.tei-c.org/release/doc/tei-p5-doc/en/html >

Distant reading in the history of philosophy: Wittgenstein and academic success

Guido Bonino (Università di Torino), guido.bonino@unito.it
Paolo Tripodi (Università di Torino), paolo.tripodi@unito.it

This contribution is part of the larger project DR2 (Distant Reading and Data-Driven Research in the History of Philosophy, www.filosofia.unito.it/dr2 – University of Turin), which aims at getting together and coordinating a series of research activities, in which Franco Moretti's distant reading methods are applied to the history of philosophy and, more in general, to the history of ideas. This contribution provides a sample of how such methods may usefully interact with more traditional methods in the history of philosophy, resulting in a more or less deep revision of the received views.

As suggested in the title, the topic of our contribution is the place of Wittgenstein in contemporary analytic philosophy; or, perhaps more precisely, the relationship between two philosophical traditions, the analytic and the Wittgensteinian. The main aim of the present contribution is to check whether the application of a distant reading approach can add some interesting details and insights to the historical-philosophical understanding of the "decline" of the Wittgensteinian tradition in contemporary analytic philosophy (a topic that has already been studied using traditional methods of the history of philosophy, see for example Hacker 1996 and Tripodi 2009). We consider a the period 1980-2010 in the US, by analysing the corpus of more than 20,000 PhD theses in philosophy provided by Proquest (www.proquest.com). This corpus contains the metadata (such as author, title, year of publication, name of the supervisor, university, department, abstract, keywords, and so forth) of the PhD dissertations. Within this corpus, we select and cut out the metadata of the dissertations in which the name "Wittgenstein" occurs in the abstract. They are almost 450, and half of them are directly concerned with Wittgenstein's philosophy (i.e., they are entirely devoted to Wittgenstein). For each dissertation we find out and register the main subject matter and the names that co-occur with the name "Wittgenstein". Then we try to find out, with the aid of search engines, what kind of academic career (if any) the PhD candidates were able to pursue: for example, how many of them became full professors, associate professors, assistant professors, adjunct professors; how many of them got an academic job in the US, how many went abroad; how many of them worked in the highest ranked departments, in lower ranked ones, in liberal arts colleges or in community colleges (only for undergraduates). By combining such variables together and by assigning a value to each of them, we are able to obtain a sort of "Academic Success Index" (ASI), which roughly but quite reasonably measures the academic success of PhD candidates in philosophy who wrote their dissertation on Wittgenstein (or, at least, mentioned Wittgenstein in the abstract of their dissertation). We do the same operation with other philosophers, that is, with other names occurring in the abstract of the dissertations (for example, Gadamer, Spinoza), as well as with a random sample. A first interesting result is that the index of academic success of those candidates who mention Wittgenstein in the abstract of their dissertation is significantly *lower* than the index of those who mention analytic philosophers such as David Lewis, Saul Kripke, Michael Dummett and Jerry Fodor.

This interesting fact – the fact that in the last 30-35 years a PhD candidate working in the analytic philosophical field, to borrow Pierre Bourdieu's phrase, has more chances to get a good academic job than one who belongs to the Wittgensteinian field – can be explained or interpreted in many ways, inspired by different disciplines and perspectives: for example, there are sociological explanations that are more or less plausible (some professors of philosophy had and still have more academic power than others; since certain topics are more difficult, they attract better PhD students, and so forth), but there are also historical-philosophical interpretations (philosophical fashion makes it more "profitable" to work on, say, recent mainstream analytic philosophy rather than on Wittgenstein), and many other possible answers. We have a number of good reasons, however, not to accept such

explanations and interpretations as entirely correct, or at least as complete. Once again, we try to find a somewhat novel answer to our question by applying a distant reading approach. We use a visualization software (VOSviewer; www.vosviewer.com) to represent the more frequent words occurring in the almost 450 "Wittgensteinian" dissertations and in the almost 500 "analytic" ones, respectively. The impressive result is that this kind of visualization seems to provide a key to a better understanding of the difference between the indexes of academic success: looking at the "analytic" visualization chart and considering, for example, the 50 words that are more frequently used in the abstracts (but similar results would be obtained by considering the first 10 or the first 100 of the list as well), we find the prevalence of words such as "theory", "argument", "result", "consequence", problem", "solution", "account", and so forth, whereas the Wittgensteinian visualization chart presents a different configuration and a different set of frequently used words. We would like to suggest that the presence (and the absence) of this semantic pattern refers to the presence (and the absence) of a science-oriented philosophical style and metaphilosophy. Since we think that a science-oriented philosophical style should be conceived of as part of a process of academic and scientific legitimation, the main thesis of our contribution is that the index of academic success for PhD candidates in US philosophy departments in the last 40 years is quite strictly connected to the choice of a more or less science-oriented philosophical style and metaphilosophy. Such a contention, suggested by the application of distant reading methods to the history of philosophy, throws new light on the issue of the decline of the Wittgensteinian tradition in contemporary analytic philosophy.

Bibliographic References

P. Bourdieu, *Homo academicus*, Editions de Minuit, Paris, 1984
P.M.S. Hacker, *Wittgenstein's place in 20th century analytic philosophy*, Blackwell, Oxford, 1996
F. Moretti, *Distant reading*, Verso, London-New York, 2013
P. Tripodi, *Dimenticare Wittgenstein*, Il Mulino, Bologna, 2009

Coping with interoperability in cultural heritage data infrastructures: the Europeana network of Ancient Greek and Latin Epigraphy

Giuseppe Amato, CNR-ISTI, giuseppe.amato@isti.cnr.it
Andrea Mannocci, CNR-ISTI, andrea.mannocci@isti.cnr.it
Lucia Vadicamo, CNR-ISTI, lucia.vadicamo@isti.cnr.it
Franco Zoppi, CNR-ISTI, franco.zoppi@isti.cnr.it

The EAGLE Project

Ancient inscriptions are a valuable source of information about otherwise undocumented historical events and past laws and customs. However, centuries of unregulated collection by individuals and by different institutions has led to an extremely fractioned situation, where items of the same period or from the same geographical area are presently scattered across several different collections, very often in different cities or countries.

One of the main motivations of the project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy, a Best Practice Network partially funded by the European Commission) is to restore some unity of our past by collecting in a single repository information about the thousands of inscriptions now scattered across all Europe.

The collected information (about 1,5 million digital objects at project's end, representing approximately 80% of the total amount of classified inscriptions in the Mediterranean area) are ingested into Europeana, as they represent the origins of the European culture. That information is also made available to the scholarly community and to the general public, for research and cultural dissemination, through a user-friendly portal supporting advanced query and search capabilities (Figure 1).

In addition to the traditional search options (full text search a la Google, fielded search, faceted search and filtering), the EAGLE portal supports two applications intended to make the fruition of the epigraphic material easier and more useful.

The EAGLE Mobile Application enables a user to get information about one visible epigraph by taking a picture with a mobile device, and sending it to the EAGLE portal. The application uses a visual search engine to retrieve the photographed object from the EAGLE database and provides to the user the information associated with that object.

The Story Telling Application provides tools for an expert user (say a teacher) to assemble epigraphy-based narratives providing an introduction to themes and stories linking various inscriptions together (e.g. all the public works done by an emperor). The stories are then made available at the EAGLE portal, and are intended for the fruition of the epigraphic material by less knowledgeable users or young students.

Along the same lines, in order to make the epigraphic material more interesting and usable also by non-epigraphists, EAGLE, in collaboration with the Italian chapter of the Wikimedia Foundation, is leading an effort for the enrichment of the epigraphic images and text with additional information and translations into modern languages. This additional material, residing on Wikimedia, is periodically harvested and included in the information associated with each epigraph.

During the whole project life frame the maintainability and sustainability issues have been constantly considered from both the technical and the scientific point of view. This led to the foundation of IDEA (The International Digital Epigraphy Association, http://www.eagle-network.eu/founded-idea-the-international-digital-epigraphy-association/) whose aim is the promotion of the use of advanced methodologies in the research, study, enhancement, and publication

of “written monuments”, beginning with those of antiquity, in order to increase knowledge of them at multiple levels of expertise, from that of specialists to that of the occasional tourist. Furthermore, scope of the association is to expand and enlarge the results of EAGLE providing a sustainability model to ensure the long-term maintenance of the project results and to continue to cope with its original aims. The presentation of that activity is however outside the scope of this poster.

This poster gives some insights of the overall infrastructure. The two following sections describe respectively the core of the Aggregation Infrastructure and some key characteristics of the Image Retrieval System and the Mobile Application.

Details on the characteristics and use of the two applications and the other resources can be found at:

http://www.eagle-network.eu/resources/

Figure 1 - Searching in EAGLE.

The EAGLE Aggregation Infrastructure

EAGLE aggregates content provided from 15 different archives from all over Europe. While most of them are providing records based on EpiDoc (an application profile of TEI, today the de-facto standard for describing inscription), some archives are supplying records in “personalized” formats. EAGLE aggregates data also from two other different sources: Mediawiki pages, containing translations of inscriptions, and “Trismegistos records”, containing information about inscriptions that appear in more than one collection.

The need for expressing queries against such heterogeneous material has led to the definition of a data model being able of relating separate concepts and objects in a seamless way, thus allowing both the scholarly research and the general public to achieve results which could hardly be obtained with the existing EpiDoc archives.

The EAGLE data model (Casarosa 2014) consists of an abstract root entity (the Main Object) from which four sub-entities can be instantiated: (i) Artefact (capturing the physical nature of an epigraphy); (ii) Inscription (capturing the textual and semantic nature of a text region possibly present on an artefact); (iii) Visual representation (capturing the information related to the “visual nature” of a generic artefact); (iv) Documental manifestation (capturing the description of an inscription’s text in its original language and its possible translations in modern languages). All the information to be aggregated in EAGLE will find its place into one or multiple instances of such sub-entities.

The EAGLE Aggregation Infrastructure is built on top of the D-NET software, developed by CNR-ISTI in the course of its participation in a number of European projects. D-NET is an open source solution specifically devised for the construction and operation of customized infrastructures for data aggregation, which provides a service-oriented framework where data infrastructures can be built in a LEGO-like approach, by selecting and properly combining the required services (Manghi 2014). For EAGLE, D-NET has been extended with image processing services to support the Mobile Application (Figure 2).

In D-NET, data processing is specified by defining workflows (i.e. a graph of elementary steps, with optional fork and join nodes) and meta-workflows (i.e. a sequence of workflows). A (meta-) workflow can be easily configured, scheduled and started through a D-NET tool with a graphical user interface, while the implementation of the elementary steps is done by writing programs actually executing the needed processing.

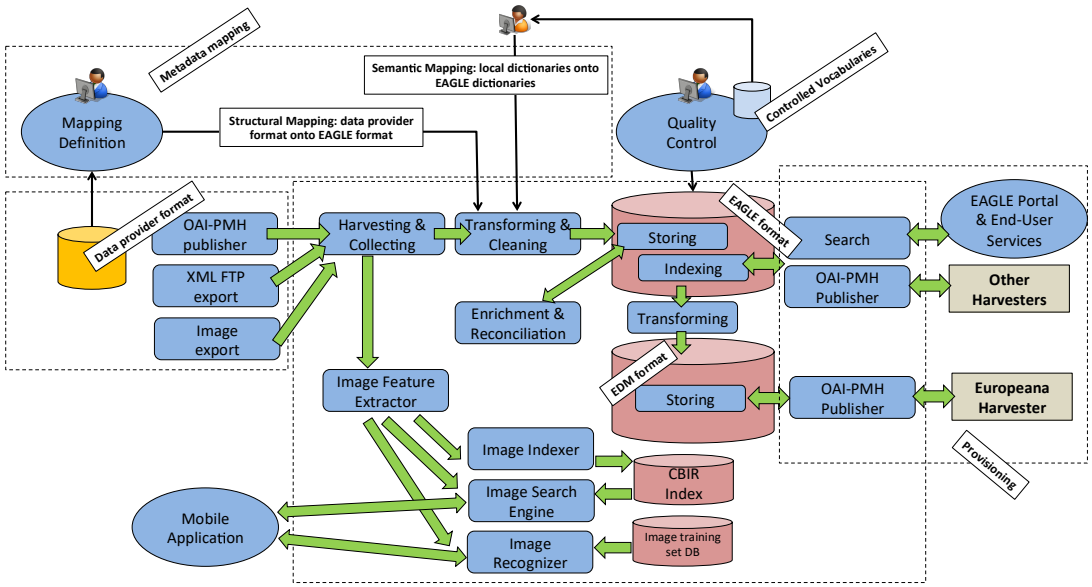


Figure 2 - The EAGLE Aggregation Infrastructure.

The EAGLE Image Retrieval System and the Mobile Application

The EAGLE Image Retrieval System allows users (like tourists or epigraphists) to retrieve information about an inscription by simply taking a photo, e.g. by using the EAGLE Mobile Application (Figure 3), or by uploading a query image on the EAGLE Web Portal (Figure 4). This represents a profitable and user-friendly alternative to the traditional way of retrieving information from an epigraphic database, which is mainly based on submitting text queries, for example, related to the place where an item has been found or where it is currently located.

The system offers two modes to search for an image provided as input query. In the first mode, called *Similarity Search*, the result will be a list of images contained in the EAGLE database, ranked in order of visual similarity to the query image. In the second mode, called *Recognition Mode*, the result of the query will be the information associated with the recognized inscription (whenever the object depicted in the query image is present in the EAGLE database). In the recognition mode, it is possible for an epigraph to appear in any position of the query image (Figure 5), also as part of a more general picture (e.g. a photo of an archeological site, or a scanned image of a page of a book).

The image search and recognition services are based on the use of visual features, i.e. numerical representation of the visual content of the image, for comparing different images, judging their similarity, and identifying common content. The image features are inserted into a Content-Based Image Retrieval (CBIR) index that allows image search to be executed very efficiently even in presence of huge quantity of images. Examples of image features are the *local features* (e.g., SIFT and SURF), the *quantization* and/or *aggregation* of local features (e.g., BoW, VLAD, and Fisher

Vectors), and the emerging *deep features* (e.g., CNN features).

During the EAGLE project, several state-of-the-art image features had been compared in order to find the most prominent approaches to visually retrieve and recognize ancient inscriptions. An extensive experimental evaluation was conducted on 17,155 photos related to 14,560 inscriptions of the Epigraphic Database Roma (EDR) that were made available by Sapienza University of Rome, within the EAGLE project. The results on EDR, presented in (Amato 2014, Amato 2016), show that the BoW (Sivic and Zisserman, 2003) and the VLAD (Jégou 2010) approaches are outperformed by both Fisher Vectors (Perronnin and Dance 2007) and Convolutional Neural Network (CNN) features (Donahue 2013) for visual retrieving ancient inscriptions. More interestingly, the combination of Fisher Vectors and CNN features into a single image representation achieved a very high effectiveness: the query inscriptions were correctly recognized in more than 90% of the cases.

Typically, the visual descriptors extracted from images have to be inserted into a CBIR index to efficiently execute the retrieval and recognition process. The EAGLE image indexer uses the functionality of the *Melampo CBIR System*. Melampo stands for *Multimedia enhancement for Lucene to advanced metric pivOting*. It is an open source CBIR library developed by CNR-ISTI, which allows efficient searching of images by encoding image features into strings of text suitable to be indexed and searched by a standard full-text search engine. In this way, the mature technology of the text search engines is exploited.

As trade-off between efficiency and effectiveness, in the EAGLE Mobile Application, the deep CNNs features have been selected and used as image features for the similarity search mode, while the VLAD has been used for the recognition functionality. Currently, more than 1.1 million epigraphs and inscriptions are visually recognizable.

Figure 3 - The EAGLE Mobile Application, which is available for download on Google Play Store, allows users to get information on a visible inscription by simply taking a picture from a mobile device.

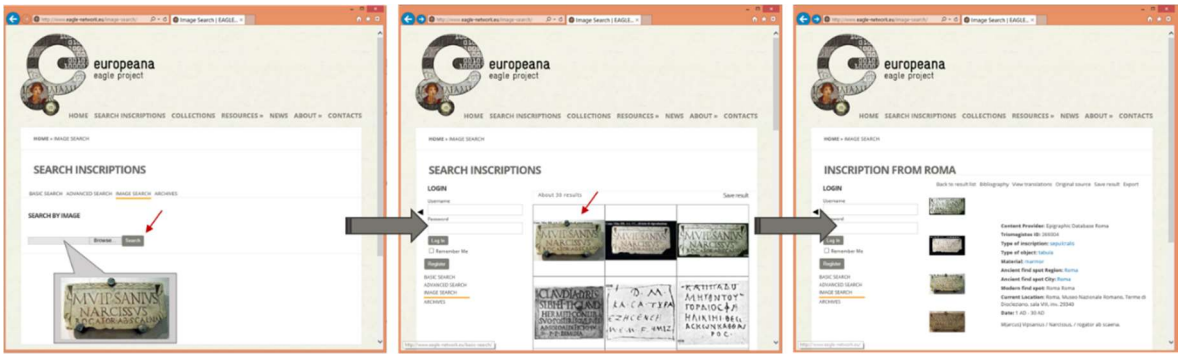


Figure 4 - Example of image search functionality in the EAGLE Web Portal (http://www.eagle-network.eu/image-search/). Given a query image, the system retrieves the most visually similar inscriptions from all EAGLE images.

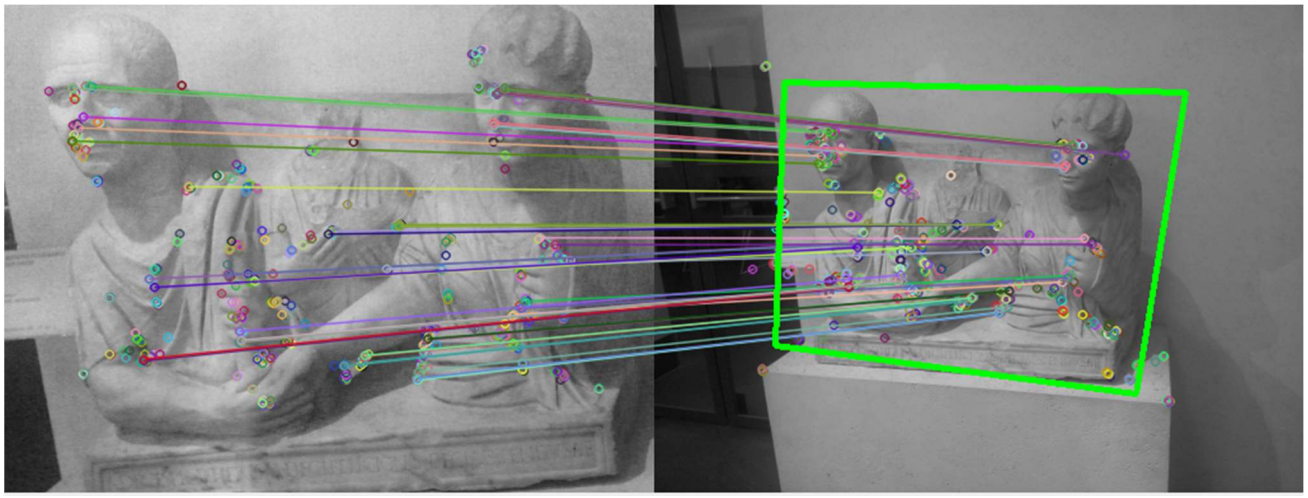


Figure 5 - Example of object recognition.

Riferimenti Bibliografici

Amato, G., Falchi, F., Rabitti, F., Vadicamo, L. 2014. "*Epigraphs Visual Recognition - A comparison of state-of-the-art object recognition approaches*". EAGLE International Conference on Information Technologies for Epigraphy and Digital Cultural Heritage in the Ancient World, Paris, September 29-30 and October 1 2014.

Amato, G., Falchi, F., Vadicamo, L. 2016. "*Visual Recognition of Ancient Inscriptions Using Convolutional Neural Network and Fisher Vector*". Journal on Computing and Cultural Heritage, Volume 9, Issue 4, Article 21 (December 2016), 24 pages. DOI: https://doi.org/10.1145/2964911

Casarosa, V., Manghi, P., Mannocci, A., Rivero Ruiz, E., Zoppi, F. 2014. "*A Conceptual Model for Inscriptions: Harmonizing Digital Epigraphy Data Sources*". EAGLE International Conference on Information Technologies for Epigraphy and Digital Cultural Heritage in the Ancient World, Paris, September 29-30 and October 1 2014.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., TDarrell, T. 2013. "*DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition*". CoRR abs/1310.1531.

Jégou, H., Douze, M., Schmid, C., Pérez, P. 2010. "*Aggregating local descriptors into a compact image representation*". In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, 3304-3311.

Manghi, P., Artini, M., Atzori C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D., Pagano, P. 2014. "*The D-NET Software Toolkit: A Framework for the Realization, Maintenance, and Operation of Aggregative Infrastructures".* In Emerald Insight, DOI http://dx.doi.org/10.1108/PROG-08-2013-0045.

F. Perronnin, F., Dance, C. 2007. "*Fisher kernels on visual vocabularies for image categorization*". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07). 1–8. DOI:http://dx.doi.org/10.1109/CVPR.2007.383266

Sivic, J., Zisserman, A. 2003. "*Video google: A text retrieval approach to object matching in videos*". In Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV'03), Vol. 2. IEEE Computer Society, 1470–1477.DOI:http://dx.doi.org/10.1109/ICCV.2003.1238663

The EAGLE project. http://www.eagle-network.eu/

Europeana. http://www.europeana.eu/

D-NET Software Toolkit. http://www.d-net.research-infrastructures.eu/

Melampo library. https://github.com/claudiogennaro/Melampo

EADH DAY

LIGHTNING TALK

DiMPO - a DARIAH infrastructure survey on digital practices and needs of European scholarship

Claire Clivaz, Vital-DH projects@Vital-IT, Swiss Institute of Bioinformatics (CH)
claire.clivaz@sib.swiss
Costis Dallas, Digital Curation Unit, IMIS-Athena Research Centre (GR)
& Faculty of Information, University of Toronto (CA) c.dallas@dcu.gr
Nephelie Chatzidiakou, Digital Curation Unit, IMIS-Athena Research Centre (GR)
n.chatzidiakou@dcu.gr
Elena Gonzalez-Blanco, Universidad Nacional de Educación a Distancia (SP)
egonzalezblanco@flog.uned.es
Jurij Hadalin, Institute of Contemporary History (SLO) jurij.hadalin@inz.si
Beat Immenhauser, Schweizerische Akademie der Geistes- und Sozialwissenschaften,
Haus der Akademien (CH) beat.immenhauser@sagw.ch
Ingrida Kelpšienė, Vilnius University (LT) ingrida.vosyliute@kf.vu.lt
Maciej Maryl, Institute of Literary Research, Polish Academy of Sciences (PL)
maciej.maryl@ibl.waw.pl
Gerlinde Schneider, Karl-Franzens-Universität Graz (A) gerlinde.schneider@uni-graz.at
Walter Scholger, Karl-Franzens-Universität Graz (A) walter.scholger@uni-graz.at
Toma Tasovac, Belgrade Center for Digital Humanities (SRB) ttasovac@humanistika.org

Abstract

In 2015, the Digital Methods and Practices Observatory (DiMPO) Working Group of DARIAH-EU conducted a European survey on scholarly digital practices and needs, which was translated into ten languages and gathered 2,177 responses from humanities researchers residing in more than 20 European countries. The full results will be presented in early 2017 at the DiMPO website, http://observatory.dariah.eu. The summary of the main results is included in a highlights report, translated into the diverse languages of the team (French, German, Greek, Polish, Serbian, Spanish; translations into other languages are expected). The survey, the first of its kind in Europe, is a perfect case of multiculturalism and multilingualism, as well as transcultural and transnational collaboration and communication, in full alignment with the 2017 topic of the EADH day.

Our presentation will outline the data-gathering process and main findings of the survey, with the aim of encouraging debate on the current state of digital practice in the humanities across Europe. The survey questionnaire consists of twenty-one questions designed to be relevant to researchers from different European countries and humanities disciplines. The main focus has been on of specific research activities, methods and tools used by the researchers. The definition of the questionnaire drew from the findings of earlier qualitative research in the context of user requirements within Preparing DARIAH (Benardou et al. 2010a; 2013), the European Holocaust Research Infrastructure (EHRI) (Angelis et al. 2012; Benardou et al. 2012; 2013), and the Europeana Research initiative (Benardou et al. 2014), building upon broader scholarship on scholarly information behaviour (e.g., Stone 1982; Bates et al. 1985; Unsworth 2000; Borgman 2007; Palmer 2009). Its structure drew from the Scholarly Research Activity Model, an activity-theoretical formal framework on scholarly activity (Benardou et al. 2010; 2013) which culminated more recently in the development of the NeDiMAH Methods Ontology (Hughes et al. 2016; Pertsas and Constantopoulos 2016).

After filtering and normalizing the dataset, the results were statistically analyzed using descriptive statistics, although simple tests of two-way association were also performed to assess the relationship

of particular responses to the respondents' country of residence, discipline, academic status and other relevant factors. In addition to the consolidated European results, six detailed national profiles have been produced, namely for Austria, Greece, Lithuania, Poland, Serbia and Switzerland. The findings suggest that the use of digital resources, methods, services and tools is widespread among European humanities researchers, and is present across the whole scholarly work lifecycle, from data collection to publication and dissemination. Results add to our understanding of how users of digital resources, methods, services and tools conduct their research, and what they perceive as important for their work. This is salient in order to ensure appropriate priorities for digital infrastructures, as well as activities and strategies for digital inception which will shape future initiatives regarding the diverse communities of researchers in the humanities.

In the next edition of the survey - scheduled for 2017 - we envisage the incorporation of questions specific to certain regions or countries so as to address the diversity of different cultural contexts. We would like also get a better idea of the familiarity of the respondents with the Digital Humanities: are they advanced DHers, or beginners, or do they define themselves not at all as DHers? In addition, we would like to explore further which digital methods are used by respondents in their research, what involves the use, creation and curation of digital resources, and what is the context of digital engagement in humanities inquiry. Such information will allow us to better situate the role and impact of DiMPO vis-a-vis general reports on the Humanities, such as the *Humanities World Report* (Holm et al. 2015). Data sustainability and consistency as we conduct the survey in the future are of central importance for our DiMPO working group.

Ultimately, the analysis of digital practices may provide original evidence, information and insight to strengthen our understanding of how humanists work, and of the nature of the humanities proper. Stanford University defines the humanities as:

> "the study of how people process and document the human experience. Since humans have been able, we have used philosophy, literature, religion, art, music, history and language to understand and record our world. These modes of expression have become some of the subjects that traditionally fall under the humanities umbrella. Knowledge of these records of human experience gives us the opportunity to feel a sense of connection to those who have come before us, as well as to our contemporaries" ("What are the Humanities ?").

Understanding the needs and actual work practices of humanists, the main purpose of the DiMPO European survey, is a *sine qua non* condition to ensure that the fundamental purpose of the arts and humanities continues to be served in the digital era. Thus, the findings of the survey will seek to strengthen the link between an empirical inquiry on scholarly digital practices (Palmer et al. 2009; Benardou et al. 2013), and the general concerns in the evolution of the humanities, such as presented in diverse national and international reports (Holm et al. 2015; SAGW 2016), or in scholarly essays (Benardou et al. 2010a and 2010b; Bod 2013; Hughes et al. 2016; Unsworth 2000). The 2015 survey is the first step towards developing a fully-fledged online observatory on the use of digital resources, methods and tools in Europe, a fact reflected in the name of the DiMPO website, http://observatory.dariah.eu, and representing the final focus and expected outcome of this DARIAH working group.

Bibliographical References

Bates, Marcia J., D. N Wilde, and S. Siegfried. 1995. "Research Practices of Humanities Scholars in an Online Environment: The Getty Online Searching Project Report No. 3." *Library and Information Science Research* 17 (1): 5–40.

Benardou, Agiatis, Panos Constantopoulos, and Costis Dallas. 2013. "An Approach to Analyzing Working Practices of Research Communities in the Humanities." *International Journal of Humanities and Arts Computing* 7 (1–2): 105–27. doi:10.3366/ijhac.2013.0084.

Benardou, Agiatis, Panos Constantopoulos, Costis Dallas, and Dimitris Gavrilis. 2010a. "A Conceptual Model for Scholarly Research Activity." In *iConference 2010: The Fifth Annual iConference,* edited by John Unsworth, Howard Rosenbaum, and Karen E. Fisher, 26–32. Urbana-Champaign, IL: University of Illinois. Accessed January 8, 2017. http://nora.lis.uiuc.edu/images/iConferences/2010papers_Allen-Ortiz.pdf.
———. 2010b. "Understanding the Information Requirements of Arts and Humanities Scholarship: Implications for Digital Curation." *International Journal of Digital Curation* 5 (1): 18–33.
Benardou, Agiatis, Costis Dallas, and Alastair Dunning. 2014. "From Europeana Cloud to Europeana Research: The Challenges of a Community-Driven Platform Exploiting Europeana Content." In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 5th International Conference, EuroMed 2014, Limassol, Cyprus, November 3-8, 2014, Proceedings*, edited by Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen, and Ewald Quak, 802–10. Lecture Notes in Computer Science 8740. Cham; Heidelberg; New York; Dordrecht; London: Springer International Publishing. Accessed January 8, 2017. http://link.springer.com/chapter/10.1007/978-3-319-13695-0_82.
Bod, Rens. 2013. *A New History of the Humanities. The Search for Principles and Patterns from Antiquity to the Present*. Oxford: Oxford University Press.
Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet.* Cambridge, MA; London: MIT Press.
DiMPO. "DARIAH-EU Digital Methods and Practices Observatory – DiMPO". Accessed January 6, 2017. http://observatory.dariah.eu
Holm, Poul, Arne Jarrick and Dominic Scott. 2015. *Humanities World Report 2015.* New York: Palgrave MacMillan. Accessed January 6, 2017. http://link.springer.com/book/10.1057%2F9781137500281
Hughes, Lorna, Panos Constantopoulos and Costis Dallas. 2016. "Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines", in *A New Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens and John Unsworth, Malden, Oxford and Chichester: John Wiley & Sons, 2016. doi:10.1111/b.9781118680643.2016.00013.x
Palmer, Carole L., Laurie C. Teffeau, and Carri M. Pirmann. 2009. "Scholarly Information Practices in the Online Environment." Dublin, Ohio: OCLC. Accessed January 8, 2017. http://0-www.oclc.org.millennium.mohave.edu/programs/publications/reports/2009-02.pdf.
Pertsas, Vayianos, and Panos Constantopoulos. 2016. "Scholarly Ontology: Modelling Scholarly Practices." *International Journal on Digital Libraries,* May 2016, 1–18. doi:10.1007/s00799-016-0169-3.
Schweizerischen Akademie der Geistes- und Sozialwissenschaften SAGW. 2016. *It's the humanities, stupid!* Bern: SAGW publication. Accessed January 6, 2017. https://abouthumanities.sagw.ch/
Stanford Humanities Center. "What are the Humanities?". Accessed January 6, 2017. http://shc.stanford.edu/what-are-the-humanities
Stone, Sue. 1982. "Humanities Scholars: Information Needs and Uses." *Journal of Documentation* 38 (4): 292–313.
Unsworth, John. 2000. "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" In *Humanities Computing: Formal Methods, Experimental Practice Symposium,* King's College, London. http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html.

Tracing the patterns of change between Jane Austen's *Pride and Prejudice* and a simplified version of the novel: what are the rules of text simplification for foreign language learners?

Emily Franzini, Georg-August-Universität Göttingen, efranzini@etrap.eu

Introduction

Authentic text and graded reader

One of the objectives of second language (L2) learning is to be able to read and understand a variety of texts, from novels to newspaper articles, written in the language of interest. These texts written with a native audience in mind are commonly referred to as authentic texts or "real life texts, not written for pedagogic purposes" [Wallace, 1992]. Authentic texts, however, can present too many obstacles for L2 learners with too low a level of knowledge. The complex language structures and advanced vocabulary of these 'real' texts can have the unwanted effect of demotivating the reader [Richard, 2001]. The gap between the learner's limited L2 knowledge and the fluency of authentic texts creates an ideal space for graded readers. Graded readers are "simplified books written at varying levels of difficulty for second language learners" [Waring, 2012]. Through graded readers, original classic works can be adapted to match the learner's level of knowledge, thus providing an ideal tool to tackle 'real' themes, narratives and dialogues.

From authentic text to graded reader

One such graded reader is a newly adapted version of Jane Austen's *Pride and Prejudice* (edition of 1813) that the author of this paper wrote [Franzini, 2016] as part of a collection for learners of English as a foreign language (EFL).

For authors, the process of adaptation of a text for a learning audience is complex. In order to simplify the text the author will necessarily have to make grammatical changes and lexical substitutions following vocabulary lists, shorten the text by cutting out entire paragraphs and events, and in some cases eliminate entire chapters and characters. Together with these changes, which can be defined as 'structural' because they are dictated by hard requirements of length and standardised level of difficulty, the author will also make a series of judgment calls at a style, sentence and word level. These changes, which are here defined as 'cognitive', include processes that are more intangible and that are a consequence of a native author's 'feeling' of how best to convey the original text. These include elaborating, clarifying, providing context and motivation for unfamiliar information and non-explicit connections [Beck et al., 1991].

Research Objective

The objective of this study is to computationally analyse the manual process behind the simplification of a historical authentic text aimed at producing a graded reader. More specifically, it aims to classify and understand the structural and cognitive processes of change that a human author, more or less consciously, is able to perform manually. Do the applied changes follow strict rules? Can they be classified as forming a pattern? And if so, can they be reproduced computationally?

Related Research

Researchers have long been addressing the issue of text simplification for a variety of purposes. A similar study to this was made by Petersen who compared authentic newspaper articles with abridged versions [Petersen and Ostendorf, 1991]. Other studies have been made, for example, to create a reading aid for people with disabilities [Canning, 2000][Allen, 2009].

Data

This study considers two sets of data. The first is a file containing the entire original novel (ON) *Pride and Prejudice*. The second set of data is a file of the graded reader (GR) published by Liberty. The GR has been compressed from the 61 chapters of the ON to 10 chapters. When comparing word tokens, the GR is in size 12.6% of the ON [Table 1]. The language was simplified to match the upper intermediate level B2[1]. To guide the choice of vocabulary, the author chose to follow the Lexitronics Syllabus.[2]

	Line count	Word tokens	Word types	Average sentence length
Original Novel	5,974	143,386	6,823	24.00
Graded Reader	1,115	18,086	1,813	16.22
% GR size in respect to ON	18.6%	12.6%	26.5%	67.5%

Table 1: Quantitative comparison between data sets

Methodology

Readability

As a first step towards analysing the differences and similarities between an authentic text and a graded reader, it was decided to evaluate if what is published as a graded reader can computationally be considered a simplified version of the original. The method chosen to make this investigation was to conduct two different readability tests, namely the ARI (Automatic Readability Index) test and the Dale-Chall Index test on the data. Both tests were designed to gauge the comprehension difficulty of a text by providing a numeric value, which corresponds to a particular school level of a native speaker of the language tested.

The results show [Table 2] that both tests yield similar scores and satisfy the hypothesis that this particular GR can be computationally proven to be, in terms of understandability, a simplification of the ON.

	ARI	Dale-Chall
Original Novel	14-15 year olds	14-16 year olds
Graded Reader	11-12 year olds	11-13 year olds

Table 2: Age level of text understandability

1 European CEFR - Common Framework of References for Languages. Language Policy of the Council of Europe: http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

2 Lexitronics Syllabus: https://tvo.wikispaces.com/file/view/20386024-Common-English-Lexical-Framework.pdf

Difference Analysis

In order to analyse the process of adaptation, a difference analysis was conducted by considering both those elements that changed from the ON to the GR, and those that, by contrast, remained the same. The analysis is structured into chapters, sentences and words, so as to proceed in order from the largest unit of text to the smallest.

When adapting a text, whether it is for a graded reader, a play or a film, the rationale behind the selection of certain parts over others is normally content-based. The author selected the most dynamic parts of the novel, which included dialogues, moments of suspense, movements of the characters and revelations. The selection of some scenes of the plot over others is purely a 'cognitive' choice of the author. As long as the main thread of the story and its main characters are preserved, the choice of scenes is entirely subjective. However, by using text reuse detection software on both texts it was possible to visualise where the majority of reuses occur. These concentrate in particular around the beginning and the end of the novel (dark green in Fig. 1).

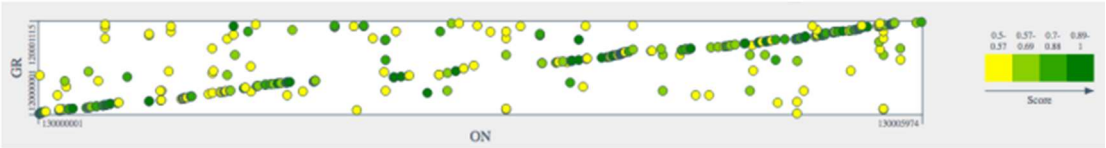


Figure 1: Visualisation of the reuses between the ON and the GR

'Structural' changes made at a sentence level present patterns that can be more systematically identified. For example, by comparing sentence length, it was noted that on average the ON contains longer sentences (24 words) than the GR (16.22 words) [Fig. 2]. Though this might seem like an obvious result, it appears less so when one thinks that, in order to simplify a concept for a language learner, it is often necessary to use additional words to elaborate or clarify it.

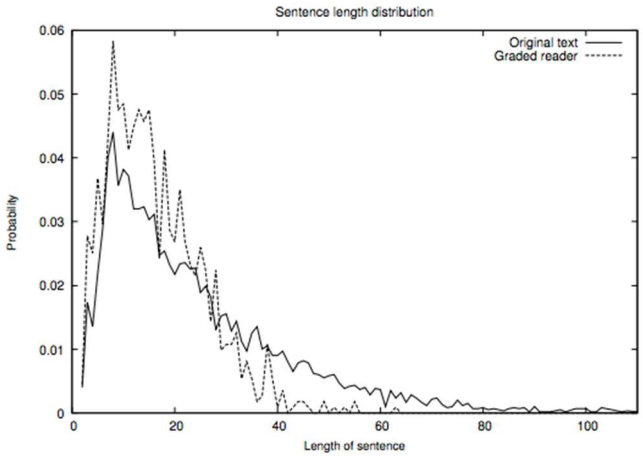


Figure 2: Sentence length distribution

In order to conduct a difference analysis on the smallest unit of text - the word - we looked at all the words that appear frequently in the ON, but that never appear in the GR, in order to understand what kinds of words the author found necessary to drop.

Word	Frequency	Word	Frequency
upon	75	table	31
least	65	astonishment	30
acquaintance	63	fancy	30
either	59	attempt	29
whose	59	dine	29
dare	53	beg	28
regard	53	depend	28
determine	47	highly	28
scarcely	45	satisfaction	28
ladyship	42	acknowledge	27
former	38	credit	27
put	36	thus	27
amiable	35	disposition	26
deal	34	exceedingly	26
design	32	praise	26
satisfy	32	pray	26
society	32	wholly	26

Table 3: Number of words that appear only in the ON

Table [3] shows that 14 out of the 34 words listed (ca. 35%) are too advanced for level B2. Some of the other words, though accessible to B2 learners, were replaced with easier synonyms. We also conducted an analysis on parts of speech and how they differ in the two data sets [Table 4].

PoS	More frequent in ON	Similar frquency	More frequent in GR
JJS adjective, superlative	X		
JJR adjective, comparative	X		
PDT predeterminer	X		
RBS adverb, superlative	X		
WDT WH-determiner	X		
FW foreign word	X		
: colon	X		
WP$ WH-pronoun, possessive	X		
NNPS noun, proper, plural	X		
SYM symbol	X		
RP particle		X	
RB adverb		X	
VB verb, base form		X	
TO 'to' as preposition		X	
JJ adjective or numeral, ordinal		X	
NNS noun, proper, singular		X	
CC conjunction, coordinating		X	
PRP$ prounoun, possessive		X	
NN noun, common, singular		X	
MD modal auxiliary		X	
IN preposition or conjuction, subordinating		X	
DT determiner		X	
VBN verb, past participle		X	
VBG verb, present participle		X	
POS genitive marker		X	
RBR adverb, comparative		X	
EX existential 'there'		X	
UH interjection			X
NNP noun, proper, plural			X
WRB WH-adverb			X
VBD verb, past tense			X
VBP verb, present tense, not 3rd person singular			X
VBZ verb, present tense, 3rd person singular			X
WP WH-pronoun			X
CD numeral, cardinal			X
PRP prounoun, personal			X

Table 4: Parts of speech frequency in the ON vs. in the GR

Conclusions and further research

This study is a first step into the realm of text simplification regarding graded readers for L2 learners. By conducting a difference analysis between the two texts, it was observed that at plot level the

selection of scenes has no impact on the difficulty of a text. The text reuse detection software used,[3] however, identified which parts of the plot have been preserved and which have been eliminated for the sake of a consistent, yet shorter, story line. It was observed that the beginning and the end of the novel were the parts that were adapted most faithfully.

The identification of reuse over the whole novel was also a step towards pinpointing where sentences were reused *verbatim* and where they were not. Where the sentences have undergone heavy changes, we can observe to what extent they were modified, how and why. At a sentence level, we noted that reducing the length of the sentences is a successful simplification strategy. A further study would have to be conducted to best understand how sentences were split or reduced, and consequently how the syntax of a sentence was affected by its shortening.

At a word level, the simplification of the text appeared to be dictated by the elimination and replacement of difficult vocabulary and certain parts of speech, such as comparative and superlative adjectives. The word length does not appear to be an indicator of difficulty. While it was observed that both the readability tests were based on sentence length as a parameter, only the ARI test, however, considers word length as another parameter. A test on the word-length distribution of the ON versus the GR shows that, in this case, the word length bears no importance in assessing the difficulty of a text. Further research would have to be conducted in order to learn if it is easier for an L2 learner to remember a word not because of its length, but because of its repeated presence in a text. The insights gained from this study will be useful in future work on automating the simplification process.

References

[Allen, 2009] Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. System, 37(4):585–599.

[Beck et al., 1991] Beck, I. L., McKeown, M. G., Sinatra, G. M., and Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. Reading Research Quarterly, Vol. 26(No. 3):251–276.

[Canning, 2000] Canning, Y. (2000). Cohesive regeneration of syntactically simplified newspaper text. Proc. ROMAND, pages 3–16. 14

[Franzini, 2016] Franzini, E. (2016). Adapted Edition of Jane Austen's *Pride and Prejudice*. Liberty Publishing.

[Petersen and Ostendorf, 1991] Petersen, S. E. and Ostendorf, M. (1991). Text simplification for lanugage learners: A corpus analysis. Speech and Lanuguage Technology in Education (SLaTE2007).

[Richard, 2001] Richard, J. C. (2001). Curriculum development in language teaching Cambridge c.u.p.

[Wallace, 1992] Wallace, C. (1992). Reading oxford, o.u.p.

[Waring, 2012] Wallace, R. (2012). Writing graded readers.

3 TRACER text reuse detection machine: http://www.etrap.eu/research/tracer/

Latin Text Reuse Detection at Scale
Orosius' *Histories:* A Digital Intertextual Investigation into the First Christian History of Rome

Greta Franzini, University of Göttingen, gfranzini@etrap.eu
Marco Büchler, University of Göttingen, mbuechler@etrap.eu

Introduction

This ongoing research aims at performing semi-automatic analysis and comparison of Paulus Orosius' (385-420 AD) most celebrated work, the *Historiarum adversum Paganos Libri VII,* against its sources. The *Histories*, as this work is known in English, were commissioned to Orosius by his mentor Saint Augustine as complementary to his own *De civitate Dei contra Paganos* and constitute the first history (752 BC-417 AD) to have been written from a Christian perspective. To do so, Orosius drew from and reused earlier and contemporary authors, including the pagans Caesar, Vergil, Suetonius, Livy, Lucan and Tacitus, thus providing a rich narrative fraught with intertextual references to poetry and prose alike.

Related Work

Text reuse in the *Histories* has already been surveyed in the *Corpus Scriptorum Ecclesiasticorum Latinorum [CSEL]* (vol. 5, 1882) and in the *Patrologia Latina [PL]* (vol. 31, cols. 0663-1174B, 1846). There, the editors list the reuses together with detailed information about the source passages. However, no information is given regarding the style of Orosius' reuses. Furthermore, one can only *trust* that the CSEL and PL indices are complete. Looser forms of reuse, such as allusions or echoes, may have eluded the editors.

> "It would be burdensome to list all of the Vergilian echoes [...]" (Coffin 1936, 237)

What Coffin describes as "burdensome" can be accomplished with machine assistance. To the best of our knowledge, the present research is the first attempt to computationally corroborate known text reuse in the *Histories* and to use its rich intertextuality to refine algorithms for historical text reuse detection.

Challenges and Research Questions

Orosius' reuse style is extremely diverse, ranging from two words to longer sentences, and from verbatim quotations to paraphrase or reuses in inverted word order. This diversity challenges automatic text reuse detection as no single algorithm can extract all of the different reuse styles.

The Latin corpus under investigation is also challenging due to its size and diachronicity. While presently testing our detection methodology on a sample number of Orosius' sources, corresponding to roughly 1.3 million words of Latin, the corpus will grow to include all of his sources. Such a large corpus forces one to experiment with different detection tasks and settings in order to tease out as many reuses as possible. Covering a 500-year period of Latin language, the texts contain differences in vocabulary and different spelling conventions, requiring non-invasive but considerable data pre-processing work to help produce usable machine-results.

The research questions underpinning this research are: how does Orosius adapt his sources? Can we categorise his text reuse styles and what is the optimal precision-recall retrieval ratio on this large historical corpus? How does automatic detection at scale affect performance?

The Corpus

All of the public-domain works for this study were downloaded from *The Latin Library*[1]. Unlike analogous resources, *The Latin Library* provides clean and plain texts (.txt), the format required by the text reuse detection machine used in this study, TRACER.

Table 1 below outlines the authors and works under investigation in chronological order. To give an idea of the size of the texts, the 'Tokens' column provides a total word-count for each work; the 'Types' column provides the total number of unique words; and the 'Token-Type Ratio' shows how often a type occurs in the text (e.g. a TTR of 3 indicates that for every type in a text there are three tokens. Generally, the higher the ratio the less linguistic variance in a text). This table reveals the language and challenges we should expect when detecting reuse. For instance, Caesar, Lucan and Tacitus share similar text lengths but Caesar has a higher TTR; this tells us that Caesar's text has less linguistic variety than Lucan and Tacitus. Conversely, if we look at Suetonius in comparison to Lucan and Tacitus, we notice a larger text but a similar TTR. This indicates a high linguistic variance in Suetonius' text, and one that can prove challenging for text reuse detection.

Author [date]	Latin Style	Work (type)	Tokens	Types	Token-Type Ratio (TTR)
Julius Caesar [100-44BC]	Classical	De Bello Gallico (prose)	51,723	11,100	4.65
Vergil [70-19 BC]	Classical	Aeneid (epic poem)	63,715	16,799	3.79
Vergil [70-19 BC]	Classical	Georgics (epic poem)	14,175	6,974	2.03
Livy [59 BC-17 AD]	Classical	Ab urbe condita (prose)	507,120	50,774	9.98
Lucan [39-65 AD]	Classical	De Bello Civili sive Pharsalia (epic poem)	51,033	14,780	3.45
Tacitus [56-117 AD]	Classical	Historiae (prose)	51,417	15,347	3.35
Suetonius [69-ca.130 AD]	Classical	De Vitis Caesarum (biography)	71,040	21,565	3.29
Florus [74-ca. 130AD]	Classical	Epitome de T. Livio Bellorum Omnium Annorum DCC Libri Duo (prose)	26,750	9,181	2.91
*Justin [3rd century]	Late	Historiarum Philippicarum T. Pompeii Trogi Libri XLIV in Epitomen Redacti (prose)	61,256	15,134	4.04
Eutropius [n.d.-ca. 399AD]	Late	Breviarium ab Urbe Condita (prose)	18,873	5,575	3.38
St. Augustine [354-430AD]	Late (Ecclesiastical)	De civitate Dei contra Paganos (prose)	274,720	35,430	7.75
Orosius [385-420 AD]	Late (Ecclesiastical)	Historia adversum Paganos (prose)	74,929	19,748	3.79
					Total tokens (words to be processed): 1,266,751

Table 1. Overview of analysed texts. Excluded texts will be included in a second phase of the project. Justin is still being process so we exclude him from the discussion for now.

1 At: http://www.thelatinlibrary.com/ (Accessed: 7 January 2017).

Methodology

Our workflow makes use of five "tools": the *TreeTagger* Latin parameter file[2] and *LemLat 3*[3] for Part-of-Speech (PoS) tagging and lemmatisation; the *BabelNet 3.7*[4] and *Latin WordNet*[5] Latin lemma lists and synonym sets to support the detection of paraphrase and the extraction of paradigmatic relations; and TRACER, our text reuse detection machine.[6]

First, the data is acquired, cleaned through custom scripts and normalised. Next, the texts are tagged for PoS and lemmatised using first *TreeTagger*, which disambiguates *tokens*, and then *LemLat 3*, which disambiguates *types*. We use both tools to ensure the best possible tagging and lemmatisation output. Word forms that *TreeTagger* and *LemLat 3* do not recognise are called *unknowns*. These can be caused by residual dirt in the text (e.g. missing white-space, symbols, etc.) or by missing entries in the tools' embedded dictionaries. We manually filter unknowns into two lists, *dirt* and *missing forms*, and correct all those caused by dirty text by identifying and rectifying the problem in the corpus. The tagging and cleaning of the corpus is performed iteratively until the only unknowns are those caused by *missing forms* (e.g. named entities), which we store separately for the potential improvement of *TreeTagger* and *LemLat 3*.[7] At the time of writing, the corpus is being processed and cleaned for a third time.

In order to detect both verbatim and looser forms of text reuse, TRACER requires as input: 1) the corpus, 2) the PoS/lemma information extracted from the corpus, and 3) the Latin WordNet.

TRACER is a powerful suite of some 700 algorithms packaged to detect text reuse in different texts and languages. TRACER offers complete control over the algorithmic process, giving the user the choice between being guided by the software or to intervene by adjusting search parameters. In this way, results are produced through a critical evaluation of the detection.

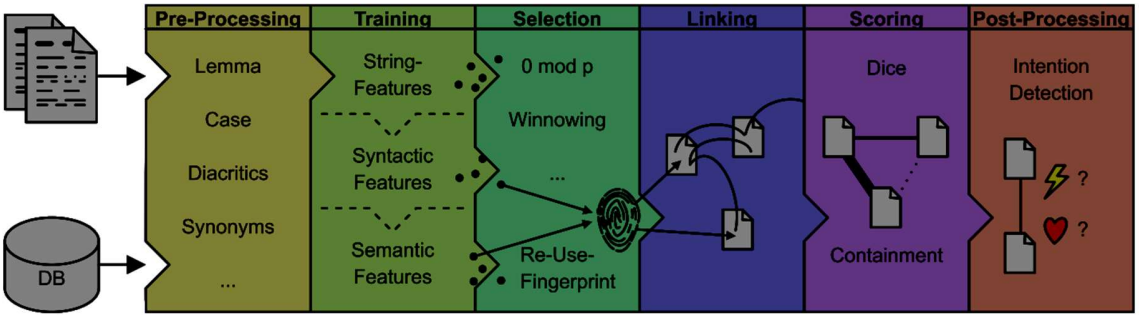


Figure 1: TRACER splits every detection task into six steps (from left to right).

The text reuse diversity in Orosius' *Histories* calls for different TRACER detection settings and parameters. For every detection task we keep a record of the parameters used and the results produced. The computed results are manually compared against the known reuses documented in the aforementioned editions to check for

2 At: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Accessed: 7 January 2017).

3 At: http://www.lemlat3.eu/ (Accessed: 7 January 2017).

4 At: http://babelnet.org/ (Accessed: 7 January 2017). We are also in possession of the Latin synonym-set provided by *BabelNet* but we have yet to test it.

5 That is, the Latin contained in the *Ancient Greek WordNet*, available at: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-56 (Accessed: 7 January 2017).

6 At: http://www.etrap.eu/research/tracer/ (Accessed: 25 October 2016).

7 The processing of the corpus with *LemLat 3* is performed by Marco Passarotti and Paolo Ruffolo of the Università Cattolica in Milan.

matches and new discoveries yielded by TRACER, if any. The manual comparison is facilitated by an XML-encoded copy of the corpus that we are currently creating, in which we annotate text reuses documented by the CSEL and PL editions.

Preliminary Results

Here we present the results of an initial TRACER run that was performed on the first version of the downloaded corpus as a proof-of-concept[8].

The texts were segmentised by sentence. The average sentence length measured across the entire corpus is 31 words per sentence. A first text reuse detection experiment at the sentence-level failed due to the presence of very short reuses. For this reason, the segmentation was changed to a moving window of ten words, restricting TRACER's detection from entire sentences to smaller units. In the *Selection* step (see Figure 1), we experimented with different pruning parameters, which produced few but precise results. We eventually opted for PoS selection, which considered nouns, verbs and adjectives as relevant reuse features, in order to obtain a higher recall (i.e. more reuse candidates) over precision.

In the *Scoring* step, we used the resemblance score, which measures the ratio of overlapping features with the overall unique set of features of two alignment candidates. The results of this first run of TRACER: one notices that almost 50% of all scored alignment pairs of two text passages have a four-word overlap (e.g. nouns, verbs, etc.), and that 93.9% of all candidates have overlaps of 3, 4 or 5 words, indicating a fragmentary reuse style rather than block-copying.

Again, this result is based on an analysis of the first version of the corpus and does not claim to extract all reuses from the *Histories*. It was performed as a proof-of-concept and will be refined through cleaner versions of the corpus and additional TRACER detection tasks with different settings. We expect to draw meaningful conclusions from a comparison and merge of the results of different detection tasks.

Research Value

From a computational standpoint, this research establishes a text reuse detection workflow for Latin that integrates *TreeTagger*, *LemLat 3* and the *Latin WordNet* with TRACER, a workflow that is able to perform detection *at scale*. From a humanities perspective, this research explores the ways in which Orosius adapts his sources; the computed results are compared against the reuse identified in the aforementioned editions of the *Histories* as a means of bridging the gap between close and distant reading, and of potentially revealing previously unknown reuse.

This project also serves as a case study for the testing of linguistic resources for Latin and, through collaborations, works towards the establishment of a Gold Standard for Latin lemmatisation, one that accounts for the evolution of the Latin language.

Deliverables and Next Steps

Once complete, we plan to publish the corpus in both plain text and XML formats, as well as an index of the text reuses manually and computationally extracted, organised by reuse style.

Works Cited

H. C. Coffin, “Vergil and Orosius”, *The Classical Journal*, 31(4) (1936): 235-241

8 That is, the version of the corpus that did not include the improvements made following the first *TreeTagger* and *LemLat 3* analysis.

Analyzing poetry databases to develop a metadata application profile. Why each language uses a different way of modelling?

Patricia Garrido, LINHD-UNED and UCM, pgarrido@linhd.uned.es

Introduction

This lightning talk is a description of a work-in-progress which explains my collaboration in the POSTDATA[1] project where I am working as a student in practices, contributing to the Project with my knowledge in philology, and learning how to use DH tools and methodologies to analyze traditional philological problems.

POSTDATA project and its process

My contribution to this project belongs to the first of its work packages: "semantic web and ontology development", which deals with the development of a metadata application profile for poetry. It is reverse engineering process, as we analyze the logical models of different databases and create particular conceptual models in order to create a final and common conceptual model to all the existing ones. For the accomplishment of this work, a classification of the different databases has been made taking into account the language in which poetry is written. At the moment, I am working in specific repertoires and databases devoted to Latin poetry from different provenances and universities: *Pedecerto*, the *Corpus Rhytmorum Musicum Analecta Hymnica Digitalia and Analecta carminum medii aevi*, and comparing them with other repertoires of German, English, French, Spanish and Portuguese poetry.

First, it is necessary to analyze the logical model of each database in order to understand the concepts that are represented by each table making a description of the different terms that were chosen by the designers. An example of this procedure can be well explained using *Pedecerto*[2] as a case study, a digital instrument for the analysis of Latin verses. It is a repertoire which is composed by two different databases, sending user information from both of them. For example, the word "sistema" appears in the model without any contextualization and it becomes difficult to interpret it. For that reason, it is necessary to go back to the website and look for disambiguation. In the case of "sistema", the conclusion is that this term describes "the type of behavior in the metric system" (If there is a "D" it is a dactylic system; if "E" it is an elegiac couplet; if "N" we have hexameter and pentameter meters mixed with other kind of meter...)

A similar phenomenon happens to the *Corpus Rhytmorum Musicum*[3], which is a musical and textual philological database of the earliest Medieval Latin Songs. This one is more related to music and manuscripts, so I find terms such as "NRMano" and exploring the website as I have explained below, I can describe the term as the "number of hands which have written a determinate manuscript".

It is necessary to build an abstract model in which the terms used for describing general concepts, such as "manuscript", "poem" or "literary work" have identical or very similar meaning across the different databases.

There is a second phase in this process, which consists of the analysis and grouping of the controlled vocabularies from each different literary tradition, which are collected by the search tools

1 The POSTDATA project: http://postdata.linhd.es/

2 The Pedecerto repertoire, supported by the University of Udine, has its own website: http://www.pedecerto.eu/

3 The The Corpus Rhytmorum Musicum is supported by the University of Siena in Arezo and its website is: http://www.corimu.unisi.it/

of the repertories. The study of controlled vocabularies can be focused from different perspectives, but we first classify the term looking later for groups and hyperonyms. The execution of this task is very positive for the review of the previous one, since in the logical entity we find terms that refer to controlled vocabularies and must not appear in the conceptual model. As many databases do not show a regular work on controlled vocabularies, it is sometimes not easy to identify and extract their terms and keywords. In this sense, *ReMetCa* Project is a repertoire of special relevance, as it has developed a great effort to study controlled vocabularies using external tools, as Tematres.

So, this Lightning Talk will describe all these methods to compare and analyze poetry databases, but also will reflect on the idiosyncrasy of classifying poetry and the differences of conceptualization among the different languages, literatures and traditions and its representation in the digital world.

References

González-Blanco García Elena and Rodríguez Gómez, José Luis, "ReMetCa, an integration proposal of MySQL and TEI-Verse"." Issue 8 del Journal of the Text Encoding Initiative (2015)
González-Blanco García, Elena, del Rio Riande, Gimena, and Martínez Cantón, "Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires", LREC 2016 Proceedings (2016)
González-Blanco García, Elena, "Un nuevo camino hacia las Humanidades Digitales: El Laboratorio de Innovación en Humanidades Digitales de la UNED (LINHD)", Signa, Revista de la Asociación Española de Semiótica, 25 (2016): 79-93.

Repertoires and projects

Corimu: http://www.corimu.unisi.it/
POSTDATA: http://postdata.linhd.es/
Pedecerto: http://www.pedecerto.eu/
ReMetCa: http://www.remetca.uned.es/index.php?lang=es
Analecta Hymnica Digitalia and Analecta carmine medii aevii: http://webserver.erwin-rauner.de/crophius/Analecta_conspectus.htm

EVILINHD, a Virtual Research Environment open and collaborative for DH Scholars

Elena González-Blanco, LINHD-UNED, egonzalezblanco@flog.uned.es

Virtual Research Environments (VREs) have become central objects for digital humanist community, as they help global, interdisciplinary and networked research taking of profit of the changes in "data production, curation and (re-)use, by new scientific methods, by changes in technology supply" (Voss and Procter, 2009: 174-190). DH Centers, labs or less formal structures such as associations benefit from many kinds of VREs, as they facilitate researchers and users a place to develop, store, share and preserve their work, making it more visible. The focus and implementation of each of these VREs is different, as Carusi and Reimer (2010) show in their comparative analysis, but there are some common guidelines, philosophy and standards that are generally shared (as an example, see the Centernet map and guidelines of TGIR Huma-Num 2015).

This lighting talk presents the structure and design o

f the VRE of LINHD, the Digital Innovation Lab at UNED (http://linhd.uned.es), and the first Digital Humanities Center in Spain. EVILINHD focuses on the possibilities of a collaborative environment for (profane or advanced) DH Scholars.

The platform developed offers a bilingual English-Spanish interface that allows users register, create new projects and join the existing ones. Projects are shared by teams and created and published from the beginning to the final publication on the web without exiting the platform. Three types of projects may be created: 1) digital scholarly TEI-based editions using eXistDB, TEIscribe and TEIPublisher, 2) digital libraries using Omeka, and 3) simple and beautiful websites using Wordpress. There is also a customized option which allows to create projects combining all of these ingredients or part of them.

Once projects are finished, the environment offers the possibility of publication in LINDAT repository, the Clarin.eu infrastructure for deposit data and projects, as LINHD is part of the Spanish Clarin-K Centre (Bel, González-Blanco and Iruskieta, forthcoming). To publish projects into the repository, additional metadata are requested following the TADIRAH DH classification created by DARIAH. Once projects are published in LINDAT, they get permanent identifiers provided by Handle and they are harvested by the Clarin.eu Virtual Language Observatory.

The environment combines open-access free software tools well-known and widespread in the DH communities, and also some proprietary developments, like the TEIScribe visual XML cloud editor, developed at LINHD. All of them are integrated in a single log on environment based on Ubuntu and covered with an architecture of web standard technologies (such as PHP, SQL, Python and eXistDB).

Bibliographical references

Bel N., González-Blanco García, E., and Iruskieta M. 2016 (forthcoming). CLARIN Centro-K-español. *Procesamiento del Lenguaje Natural* (forthcoming for *Revista de la SEPLN*).

Candela, L. Virtual Research Environments. *GRDI2020.* http://www.grdi2020.eu/Repository/FileScaricati/eb0e8fea-c496-45b7-a0c5-831b90fe0045.pdf (accessed 28-10-2015).

Carusi, A. & T. Reimer. *Virtual Research Environment Collaborative Landscape Study. A JISC funded project (January 2010).* Oxford e-Research Centre, University of Oxford and Centre for e-Research, King's College London https://www.jisc.ac.uk/rd/projects/virtual-research-environments (accessed 28-10-2015).

Tgir Huma-Num, *Le guide des bonnes pratiques numérique.* 2011. http://www.huma-num.fr/ressources/guide-des-bonnes-pratiques-numeriques (version of 13-1-2015). (accessed 28-10-2015).
Voss, A & Procter, R. 2009. Virtual research environments in scholarly work and communications. *Library Hi Tech*. Vol. 27 Iss: 2, pp.174-190.

One-member non-bursary text reuse project in a minor language – is it manageable?

Ernesta Kazakėnaitė, Vilnius University

The idea of lightning talk is to share the experience of carrying out a one-member project with no financial support except a PhD scholarship and no possibilities to use any DH tools for Text Analysis, because the specific type of materials under study are 16th- and 17th- century writings of a minor morphologically rich language.

The official title of the PhD project that will be presented is *The First Latvian Translation of a Lutheran Bible and its Printed Excerpts from 16th and 17th Century: Interconnected Links and Influences*. This is a kind of text reuse project. Its main goal is to show that the translator of the first Latvian Bible (which was published in 1685) used not only originals, the Vulgata or Lutheran Bible, as was common practice for translations at that time, but also earlier Latvian printed texts (as pericopes) where possible. For now I have made my own corpora to compare all sources word-by-word. The reason why I am not able to use Text Analysis tools is that there are no common XML-TEI documents available and it was unmanageable for one person in such a short time to convert all the necessary texts. Moreover, I could not use modern programs to detect plagiarism which might aid finding textual similarities, because all the relevant texts are in their own writing systems, which are not comparable with one another. For example, one lexeme can be written in more than ten different ways: *wueſſe notal – wueſſenotal – wueſſenotalle – wueſſe notalle – wiſſunotaļļ – wiſſunotaļ – wißnotaļļ – wiſſinotaļ – wiſnohtał – wiſs notaļ – wiſſnotaļ* ‘decidedly’ etc. That is why to identify text reuse I came up with the simple idea of comparing these texts in a *Microsoft Word Doc* file. In presentation I will demonstrate methodology and the analysis criteria.

WeME: A Linked Data Metadata Editor for Archives

Angelica Lo Duca, IIT-CNR, angelica.loduca@iit.cnr.it

Over the last years, a great effort has been done in the field of Cultural Heritage to digitize documents and collections in different formats, such as PDF, plain texts and images. All these data are often stored either in libraries or big repositories in the form of books. Furthermore, the process of digitization requires the addition of metadata, which describe information about documents. This process is often tedious because it consists in adding well-known information about a document (such as the author's name and date of birth) to the collection, manually. In general this manual effort produces three main disadvantages: a) the probability of introducing errors increases, b) the whole process is slowed down because it is not automatic, c) inserted information is isolated, i.e. not connected to the rest of the Web.

In this presentation we illustrate the Web Metadata Editor (WeME)[1], a Web application which provides users with an easy interface to add metadata to documents belonging to a collection. WeME helps archivists to enrich their catalogues with metadata extracted from two kinds of Web sources: Linked Data and traditional Web sources. WeME mitigates the three described disadvantages produced by manual effort, by extracting well-known metadata from some Linked Data nodes (e.g. DBpedia, GeoNames) and other traditional Web sources (VIAF, Treccani and Google Books). In details, WeME exploits semantic and traditional Web to extract information through the construction of SPARQL and RESTful APIs queries to the Web sources totally transparent to the user, who must specify only the name of the resource to be searched. The advantages derived from WeME are essentially two: firstly WeME eases the task of adding metadata to documents; secondly, WeME establishes new relations both among documents within the same catalogue and with documents belonging to the Web sources. The current version of WeME does not support any refining tools, but we are going to add them as future work.

In details, in order to add information about a document, a user can insert it manually, or exploit the search option provided by WeME, which triggers the search over the Linked Data nodes and the other Web sources. If the search is successful, WeME populates the properties about the document automatically, such as author's birth and death dates, the places of birth and death and a short biography. The user can decide whether or not to accept the retrieved information.

1 https://github.com/alod83/metadata_editor

The Reception of Italian Literature in Nineteenth-Century England. A Computational Approach

Simone Rebora, Georg-August-Universität Göttingen, simone.rebora81@gmail.com

While extensive research has been focused on the reception of major authors (e.g. Dante Alighieri), not enough attention has yet been dedicated to the general reception of Italian literature abroad. This paper presents and discusses a project design that aims at filling this void, profiting of the extensive repositories available online and combining multiple computational techniques in a processing pipeline.

Limiting the analysis to secondary literature in nineteenth-century England, the corpus will be composed using the texts freely available through digital platforms and libraries such as *The Internet Archive*, *HathiTrust*, and *Europeana*. Preliminary analysis shows an inconstant quality of the optical character recognition (OCR), thus advising for a reprocessing of the scanned images. In terms of efficiency, while 100% accuracy won't be reachable, a comparison of the results provided by different tools (e.g. the free software *OCRopus*, *Ocrad*, and *Tesseract*) will allow a refinement of the overall quality.

In the second part of the project, a process will be developed for the individuation of passages dedicated to Italian authors. Two approaches are possible: (1) through named-entity recognition (NER) and (2) through topic segmentation. After having compiled an extensive list of authors' names, approach (1) will provide quicker results, matching the named entities with the authors and separating the related passages through punctuation marks. Among the free software ready for use, see *Stanford NER*, *ANNIE* and *OpenNLP*. Approach (2) is more refined, but more difficult to realize. Topic segmentation is a methodology that still lacks its "gold standard" and has been generally developed in fields other than humanities, such as medicine. However, it offers the possibility of splitting the passages with better accuracy (see software such as *TextTiling*, *C99*, and *TopicTiling*). Once again, the two approaches will be combined and compared.

In the third part of the project, the extracted passages will be analyzed using sentiment analysis tools. While these algorithms are still rarely applied to literary texts (the dominant fields of application are marketing and sociology), they may provide reliable results for the texts selected here, that are generally informative. Among the free software, see *Stanford Sentiment Analysis*, *SentiStrength*, and *NLTK*.

The final goal will be the production of graphs comparable to those of *Google Ngram Viewer*, but better refined—because able to quantify the amount of text dedicated to an author in a specific time frame, providing also an intuitive visualization of the positive/negative reception. The process will be initially tested on the corpus of 23 texts already analyzed by the submitter during his doctoral research (focused on Italian literary historiography in English language), thus choosing the best combination of methods by comparing the results. Among the possible new corpora, see the journals (e.g. *Foreign Quarterly Review*, published between 1827 and 1846) and the travel books. However, the test set may be ideally expanded to the whole corpus of digitized (secondary) literature, and the focus shifted to different subjects and countries, thus providing an effective tool for the study of literary historiography from a European—or even Global—perspective.

Genderless? Digital Cultures in a gendered perspective

Alessia Scacchi, University of Rome "Sapienza", alessia.scacchi@uniroma1.it

The scientist Evelyn Fox Keller, in an interview with Elisabetta Donini in 1991, argued that "science and technology provide us with the tools to transform the world, to deconstruct the nature in the most radical way", so it is time to de-build a formal model, extremely effective, that wants technology created and developed by a universal male. It is possible meaning at the genre as a filter, the boundary between humans and knowledge, theory and experience, between observation's power and body.

Therefore gender and technology are a phrase that hides the commitment of many scholars, scientists and philosophers who have exercised their intellectual power to define, delimit, move relatively sexual identity, at least identifying problems related to it. In this sense definition of "cyberscience" by Keller doesn't adhere to information theory, cybernetics, systems' analysis, information technology, because, unlike US physics, this study is still an historical humanities object.

Studying biographies of scientists womens can be noted that constants are a lot: these women seem to share a very important male figure that supports or is the shadow in which they work - from Hypatia who collaborated with his father Theon. Perhaps, this is the reason why they were erased from official history, even under pseudonyms with which they were forced to publish. They have shown interest in science, with extensive production of manuals, translation and teaching activities; were forward-looking, patient, capable of producing results thanks to collaborative work. This is not in the case of these women because this was their limit of investigation and mathematical technique. According to Marina Mizzau:

> nell'affannosa elaborazione di strumenti di misurazione, si dimentica che il problema non è solo come, ma soprattutto che cosa misuriamo: il metodo diventa il letto di procuste dell'oggetto, il dito sostituisce la luna.

So, you could design a computer without thinking that technology allows immediate access to the asexual composition or transmittal of contents? It would be the case of a team of scientists, working on nuclear fusion, would deliberately shun purposes' thoughts and all the consequences on the existence of human race.

Computer science, conceived as a living organism, is the technical representation of harmony as in indeterminate, as in determined reality, can be the "verb" that unifies and involves complexity. As Weil thinks, a rather mystical approach, that nevertheless explains peculiarities of scientific women's research in this digital era.

Today it is necessary to rethink digital cultures in the light of the progress made by gender theories. The metamorphosis between cyborgs and nomads, mentioned by Braidotti, now it's going on, but we need to resist disembodiment, postmodernist's cybermonsters, because we are into a multifaceted and changing reality, a technological world.

> I corpi-macchina nomadi sono potenti figurazioni del non unitario soggetto-in-divenire che io considero l'alternativa più rilevante alla crisi del soggetto umanista. Essa ha come base e porta a risoluzione il doloroso processo storico di emancipazione delle teorie della

soggettività dal concetto di individualismo. I corpi-macchina nomadi suggellano inoltre una nuova alleanza tra pensiero concettuale e creatività, ragione e immaginazione.

Changing subjects, therefore, goes beyond the boundaries between subjectivity and individualism and redefines new paths for a better cohabitation between gender and science, especially if we rethink on digital cultures.

Bibliography

Moschini, Laura. 2013. *Il rapporto tra etica scienza e tecnologia: ricerca in ottica di genere*. Roma: Aracne

Demaria, Cristina and Violi Patrizia. 2008. *Tecnologie di genere: Teoria, usi e pratiche di donne nella rete*, Bologna: Bononia University Press

Pugliese, Annarita Celeste and De Ruggieri, Francesca. 2006. *Futura: genere e tecnologia*. Roma: Meltemi

Sesti, Sara. *Donne di scienza: un percorso da tracciare*, in Badaloni, Silvana and Perini, Lorenza. 2005. *Donne e scienza: Il genere in scienza e ingegneria: Testimonianze, ricerche, idee e proposte*, Padova: CLEUP

Turkle, Sherry. 2005. *La vita sullo schermo: Nuove identità e relazioni sociali nell'epoca di internet*. Milano: Apogeo

Tugnoli Pàttaro, Sandra. 2003. *A proposito delle donne nella scienza*. Bologna: CLUEB

Braidotti, Rosi. 2002. *In metamorfosi. Verso una teoria materialista del divenire*. Milano: Feltrinelli

Bozzo, Massimo. 1996. *La grande storia del computer*: *Dall'abaco all'intelligenza artificiale*. Bari: Edizioni Dedalo

Donini, Elisabetta. 1991. *Conversazioni con Evelyn Fox Keller: Una scienziata anomala*. Milano: Elèuthera

Weil, Simone. 1988. *Quaderni. Volume terzo*. Milano: Adelphi

Rothschild, Joan. 1986. *Donne, tecnologia, scienza: un percorso al femminile attraverso mito, storia, antropologia*. Torino: Rosenberg & Sellier

Mizzau, Marina. 1979. *Eco e Narciso: Parole e silenzi nel conflitto uomo-donna*. Torino: Bollati Boringhieri.

A Trilingual Greek-Latin-Arabic Manuscript of the New Testament: A Fascinating Object as Test Case for New DH Practices

Sara Schulthess, Vital-DH projects@Vital-IT, Swiss Institute of Bioinformatics,
sara.schulthess@sib.swiss
Claire Clivaz, Vital-DH projects@Vital-IT, Swiss Institute of Bioinformatics,
claire.clivaz@sib.swiss
Anastasia Chasapi, Vital-DH projects@Vital-IT, Swiss Institute of Bioinformatics,
anastasia.chasapi@sib.swiss
Martial Sankar, Vital-DH projects@Vital-IT, Swiss Institute of Bioinformatics,
martial.sankar@sib.swiss
Ioannis Xenarios, Vital-DH projects@Vital-IT, Swiss Institute of Bioinformatics,
ioannis.xenarios@sib.swiss

Introduction

The aim of the project *HumaReC* (2016-2018)[1] is to inquiry how Humanities research is reshaped by the research and publication rhythm in the digital age and to test a new model of continuous data publishing for Humanities. The edition and study of a New Testament manuscript, Marciana Gr. Z. 11 (379), will be the test case for the development of these new practices.

A research platform for continuous data publishing

HumaReC is a digital project developed on an online research platform, with a manuscript viewer at its core, including a digital edition of the text. We have implemented a blog where regular postings will be used as the means to research results in a continuous manner and encourage discussions with the public. However, the writing of a long, well-structured text still constitutes an important aspect of the that can be continuously updated. We will investigate this new editorial format, termed "web book"2, in collaboration with the scientific publishing company Brill.

The project is inscribed in the spirit of the OA2020.org initiative, in collaboration with the diverse partners such as the Biblioteca Nazionale Marciana, computer scientists and an international board of scientific experts. scientific production in Humanities. Since a paper monograph is not adapted in our case, we will develop a format that offers the possibility to link our research outcomes to the data available on the website and *HumaReC* will run over two years (October 2016-October 2018); the features of the research platform as well as the publication of the results will increase continually and social medias will be used to inform the public of new releases.

The trilingual manuscript Marciana Gr. Z. 11 (379) as object

Marciana Gr. Z. 11 (379), the object chosen for this digital inquiry, is particularly adapted for this new model of research, because of the various challenges it presents, on many levels. Marciana Gr. Z. 11 (379) is the only trilingual Greek, Latin and Arabic manuscript of the New Testament to our knowledge.

1 http://p3.snf.ch/project-169869, last accessed 06/01/2017.

It was most likely made in Sicily during the 12th century and is a product of the 'Norman-Arab-Byzantine' culture.

First of all, the multilingual aspect of the manuscript makes it worthy of treating as a digital edition, since it can offer many features that cannot be present in a printed edition. Among them are the visualization possibilities, as the manuscript is structured in three columns, each for one of the three languages. The manuscript viewer, which is developed based on the open source visualization tool EVT,[3] allows to display the edited texts according to the reader interest. Additionally, the viewer links the transcribed texts to the manuscript images.

We will also plan to use this opportunity of working with a trilingual manuscript in order to experiment with the Handwritten Text Recognition (HTR) tool of the platform *Transkribus*, an EU-funded project.[4] It will especially be interesting to try HTR on the Arabic text, *HumaReC* being the first project working with *Transkribus* with Arabic.

Finally, a digital research, open and interactive, makes sense for such a multicultural object that connects to several controversial issues in contemporary research. We can mention here the situation of the Arabic biblical manuscripts which were neglected by the Western research for contentious reasons, the apologetic use of images of New Testament manuscripts by religious groups on the Internet and the question of the influence of the Arab world in medieval Europe, that is still a debated topic among scholars.

Bibliography

Clivaz, Claire, Sara Schulthess and Martial Sankar. 'Editing New Testament Arabic Manuscripts on a TEI-base: fostering close reading in Digital Humanities', accepted for publication in *Journal of Data Mining & Digital Humanities*, ed. M. Büchler et L. Mellerin (2016). https://hal.archives-ouvertes.fr/hal-01280627.

Clivaz, Claire. 'Common Era 2.0. Reading Digital Culture from Antiquity and Modernity'. In *Reading Tomorrow. From Ancient Manuscripts to the Digital Era / Lire Demain. Des Manuscrits Antiques à L'ère Digitale*, edited by Claire Clivaz, Jérôme Meizoz, François Vallotton, and Joseph Verheyden, Ebook., 23–60. Lausanne: PPUR, 2012.

Fitzpatrick, Kathleen. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. New York: NYU Press, 2009. http://mcpress.media-commons.org/plannedobsolescence/one/community-based-filtering/.

Martin, Jean-Marie. *Italies Normandes*. Paris: Hachette, 1994.

Nef, Annliese. 'L'histoire des "mozarabes" de Sicile. Bilan provisoire et nouveaux matériaux'. In *¿ Existe una identidad mozárabe ? Historia, lengua y cultura de los cristianos de al-Andakus (siglos IX-XII)*, edited by Cyrille Aillet, Mayte Penelas, and Philippe Roisse, 255–86. Madrid: Casa de Velásquez, 2008.

Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods,* Aldershot: Ashgate, 2015 (paperbook version; HAL version, 2014: http://hal.univ-grenoble-alpes.fr/hal-01182162.)

Rosselli Del Turco, Roberto, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti, and Jacopo Pugliese. 'Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions'. *Journal of the Text Encoding Initiative*, no. 8 (2014). http://jtei.revues.org/1077.

Schulthess, Sara. 'Les manuscrits arabes des lettres de Paul. La reprise d'un champ de recherche négligé'. PhD dissertation, Université de Lausanne/Radboud Universiteit Nijmegen, 2016. http://hdl.handle.net/2066/159141.

3 https://visualizationtechnology.wordpress.com, last accessed 06/01/2017.

4 https://transkribus.eu/, last accessed 06/01/2017.

Finding Characters. An Evaluation of Named Entity Recognition Tools for Dutch Literary Fiction

Roel Smeets, Radboud University Nijmegen, r.smeets@let.ru.nl

Character relations

When human readers interpret novels they are influenced by relations between characters. These relations are not neutral, but value-laden: e.g. the way in which we connect Clarrisa with Richard is of major importance for our interpretation of the gender relations in Mrs Dalloway (1925). In literary studies, character relations have therefore lain at the foundation of a variety of critical studies on literature (e.g. Minnaard 2010, Song 2015). A basic premise in such criticism is that ideological biases are exposed in the (hierarchical) relations between representations of certain groups (i.e. gender, ethnicity, social class).

Social Network Analysis

My PhD project departs from the hypothesis that a computational approach to character relations can reveal power relations and hierarchical structures between characters in literary texts in a more data-driven and empirically informed way. In order to test this hypothesis, I will experiment with different forms of social network analysis of characters in a large corpus of recent Dutch literary novels. The first step that has to be taken is to define the nodes which constitute the social network of a novel. For that purpose, some sort of character detection has to be done in which a practical combination of Named Entity Recognition and Resolution, pronominal resolution and coreference resolution has to be operationalized.

Named Entity Recognition

In this talk I will focus on one specific aspect of character detection in literary fiction: Named Entity Recognition. Named Entity Recognition tools are regularly used in all kinds of analytical contexts, but not so often for the analysis of literary fiction. I will report on an evaluative experiment I pursued on the accuracy of existing Named Entity Recognition tools for the Dutch language. Problems surrounding the application of Named Entity Recognition on Dutch novels will be addressed by giving an overview of the precision, recall and f-scores for a series of selected tools. Furthermore, critical recommendations will be made as to how to operationalize Named Entity Recognition tools for the detection of nodes that are constitutive of social networks of character in literary texts.

Literature

Minnaard, Liesbeth. 2010. ‘The Spectacle of an Intercultural Love Affair: Exoticism in Van Deyssel's Blank en geel’. In: *Journal of Dutch Literature* (1:1).
Song, Angeline M.G. 2015. *A Postcolonial Woman’s Encounter With Moses and Miriam.* New York: Palgrave Macmillan US.

CHALLENGES

Digital humanities and our users

Pierluigi Feliciati, Università di Macerata, pierluigi.feliciati@unimc.it

My proposal is to chair a moderated brainstorming / focus group session (30') on users' needs, behaviors and satisfaction against digital humanities web resources, based on the discussion of the following questions:

- Do we take in consideration that, building a public service, the quality of use is a key-topic to be seriously considered?
- Do we adopt methods and tools such as user profiles, scenarios, personas, cards?
- Do we adopt any method to ensure a good level of usability, in the design phase (usercentred model) or in the https://translate.google.it/?hl=it&tab=wT#it/en/rigoree evaluation phase (protocols for discount evaluation)?
- Do we annotate sistematically the users' experience of our project results, when we have the occasion to present, discuss or test them?
- Do we know what are user studies and how thay could be organised to provide a better interaction between users and the resources environments we build?

The brainstorming / focus group is intended to be an useful occasion to focus even briefly the attention of DH scholars on the crucial topic of quality of our projects with a public result on the web. i.e. efficacy, efficiency and satisfaction for users. The impact of projects should be considered not only evaluating its scientific degree of exactness, novelty and originality, but opening the evaluation to final users, adopting the proper methods.

DH infrastructures, a need, a challenge or an impossible?

Elena González-Blanco, LINHD-UNED, egonzalezblanco@flog.uned.es

The growing need of shared collaborative and web-based projects has increased the need of using cloud infrastructure to develop and support DH research. However, the access to these infrastructures is not easy for three reasons: 1) economic issues, 2) academic structure and 3) not enough knowledge of the possibilities available.

Concepts that are widely spread in the industry, such as IaaS (Infrastructure as a Service), PaaS (Platform as a Service) or SaaS (Software as a Service) are starting to reach some of the biggest DH infrastructures. Examples like EGI for managing cloud research space and virtual hosting or EUDAT for data storing are IaaS, Textgrid for working digital editions and Zooinverse for creating collaborative digital projects are PaaS and web-based tools, such as Voyant Tools or IXA Pipes might be considered as SaaS.

However, how are these resources used by DH scholars and groups? The existence of big coordinated infrastructures, such as DARIAH and CLARIN at European level plays an important role for helping researchers to know and enjoy these platforms and tools, but reality is still far from been homogeneous and differs a lot between the different countries.

The challenge proposed is: how could we approach DHers and communities of researchers to discover, use and disseminate these tools?

Working with Digital Autoptic Processes is Easier than You Think!

Lovorka Lucic, Archeological Museum in Zagreb
Federico Ponchio, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR
Angelo Mario del Grosso, Istituto di Linguistica Compiutazionale, CNR
Ivan Radman-Livaja, Archeological Museum in Zagreb
Brigitte Sabattini, Aix-Marseille Université
Bruce Robertson, Mount Allison University
Marion Lamé, entre Camille Jullian, MMSH, CNRS

In 2013, the Archaeological Museum of Zagreb (AMZ) started the Tesserarum Sisciae Sylloge (TSS), a digital and online corpus of some 1200 lead tags, labels used by dyers and fullers destined to be attached to clothing during the first three centuries A.D. One side of these tags carries personal names, the other side carries an inscription mentioning the merchandise or the services to be provided, as well as a price and more often than not an indication of quantity or weight. In the intervening three years, we have sharing our progress in conferences and we have exploring *dispositive analysis* modelling to represent an inscription (Lamé, 2015). We also tested our digital tools through several long distance teaching sessions with students in Canada, France and Senegal. Several questions came out while teaching and harvesting the results of such teaching with digital autoptic processes tools. Thanks to those students and teachers we had an extraordinary field of experimentation (Lamé et al. 2017). We would be proud to present publicly the TSS website for the very first time at the EADH-Day, remembering that the association funded this project at its early stage.

If producing the TSS 1.0 was challenging, it also opens new questions. The immateriality of the digitized cultural heritage object raises issues about the cultural and social relationship between TSS users all around the world, on one hand, and, on the other hand, some cultural heritage that does not always belong to user's own culture. TSS users can never perceive the objects with their five senses, but they are in a far different situation than using just a book. TSS users seem to be disconnected from the "valeur d'ancienneté" of Cultural Heritage (Riegl 1984, Sabattini 2006). We would like to share this new challenge with you now, considering a hybrid approach, combining digital and analog tools, as well as real objects and communication between people as a way to partially overcome this limitation.

Would you be able to decipher such a lead tag if you could hold it in your own hands or would Digital Autoptic Processes, books and collaboration with people all around the world help you in such a task? So, join one of our three team between the Monday the 17th of January and Wednesday the 25th of January and participate from wherever you are in the world!

Team 1 of Lovorka Lučić - **Epigraphy, lovin' it!**
Team 2 of Federico Ponchio - **Epigraphy, go for it!**
Team 3 of Angelo Mario Del Grosso - **Epigraphy, I like it!**

- Some archaeological objects from the Archaeological Museum in Zagreb to practice scholarly work on both digital and real objects: we will bring some Roman lead tags and coins.

- An access to TSS workflow (Digital Autoptic Processes and tools) and its digital catalogue of lead tags.
- One exemplar of both big volumes (two kilograms each!) of the luxurious paper edition of the very same lead tags, with traditional drawings and the entire study produced by its curator (Radman-Livaja, 2014).
- For 30 minutes (or so) the audience will have the opportunity to use the framework of digital scholarly editing tools of the TSS, especially the Digital Autoptic Processes and to experiment the complementarity of archaeological objects, books and SDE. The audience will produce their own transcription and discuss it with other distant users, of all type (library users, students, teachers) based in Canada, Senegal and France.

To allow such interaction and discussion, the TSS is based on some highly organised digital frameworks allowing Digital Autoptic Processes by users and textual editing that deals with philological aspects. The text of these lead tags is a complex system that is modelled by adopting the object oriented paradigm: a collection of interconnected, but loosely coupled entities (ADT) having properties (data representation), behaviour (API) and identity (object). At the end of the challenge, to reward participants, we will briefly illustrate this technical background and the structure of the TSS (see technical bibliography) and how it produces TEI-XML files for interoperability. The TSS website, which is still in beta version for testing can be found at this address: www.amz.hr/tss.

Join This First TSS Challenge Online!

Like this fist TSS challenge and support your team on the TSS "Tesserarum" Facebook page: https://www.facebook.com/tesserarum

Team 1 of Lovorka Lučić - **Epigraphy, lovin' it!**

Team 2 of Federico Ponchio - **Epigraphy, go for it!**

Team 3 of Angelo Mario Del Grosso - **Epigraphy, I like it!**

Stay tuned and follow the event on the Facebook event page and Twitter:

Event Facebook: http://tinyurl.com/jogf3wa

Twitter: https://twitter.com/tesserarum

To participate from whenever you are in the world, having access to TSS's DAP between Monday the 15th and Wednesday the 25th, register to this challenge on Eventbrite - http://tinyurl.com/hustnbd

As far as the small archaeological objects are concerned, adequate gloves will be provided by the Museum. Several authors and collaborators of this challenge have prepared their DSE in advance and are present also online on D-Day, to discuss it with you.

Riferimenti Bibliografici

Ackerman, Lee, and Celso Gonzalez. 2011. *Patterns-Based Engineering: Successfully Delivering Solutions Via Patterns*. Addison-Wesley.

Almas, Bridget, and Marie-Claire Beaulieu. 2013. "Developing a New Integrated Editing Platform for Source Documents in Classics." *LLC* 28 (4): 493–503.

Del Grosso, Angelo Mario, Federico Boschetti, Emiliano Giovannetti, and Simone Marchi. 2016. "Vantaggi dell'Astrazione attraverso l'Approccio Orientato agli Oggetti per il Digital Scholarly Editing," in *AIUCD* 2016: Venezia, pp. 213-218.

Del Grosso, Angelo Mario, Davide Albanesi, Emiliano Giovannetti, and Simone Marchi. 2016. "Defining the Core Entities of an Environment for Textual Processing in Literary Computing."

In *DH2016 Conference.* 771–775. Kraków: Jagiellonian University and Pedagogical University.
Del Grosso, Angelo Mario, Simone Marchi, Francesca Murano, and Luca Pesini. 2013. "A Collaborative Tool for Philological Research: Experiments on Ferdinand de Saussure's Manuscripts." In *2nd AIUCD Conference.* 163–175. Padova: CLEUP.
Driscoll, Matthew James, and Elena Pierazzo, eds. 2016. *Digital Scholarly Editing: Theories and Practices*. Vol. 4. Digital Humanities Series. Open Book Publishers.
Evans, Eric. 2014. *Domain-Driven Design Reference: Definitions and Pattern Summaries*. Dog Ear Publishing.
Fischer, Franz. 2013. "All texts are equal, but … Textual Plurality and the Critical Text in Digital Scholarly Editions," *Variants* 10: 77–92.
Gibbs, Fred , Trevor Owens, 2012. "Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs", *DHQ* 6 (2).
Knoernschild, Kirk. 2012. *Java Application Architecture: Modularity Patterns with Examples Using OSGi*. Robert C. Martin Series. Prentice Hall.
Lamé, Marion et al. 2017 (forthcoming). "Teaching (Digital) Epigraphy". In *Digital and Traditional Epigraphy in Context. Proceedings of the EAGLE 2016 International Conference*, S. Orlandi, P. Liuzzo, F. Mambrini, and R. Santucci (eds). Sapienza University Press, Roma.
Lamé, Marion. 2015. "Primary Sources of Information, Digitization Processes and Dispositive Analysis", in F. Tomasi – R. Rosselli Del Turco, – A. M. Tammaro (ed.) *Proceedings of the Third AIUCD Annual Conference on Humanities and their Methods in the Digital Ecosystem*. ACM, article 18. http://dl.acm.org/citation.cfm?id=2802612.
Martini, Simone. 2016. "Types in Programming Languages, between Modelling, Abstraction, and Correctness," Computability in *Europe, CiE 2016*. Pursuit of the Universa. Springer, 9709, LNCS.
Palma Gianpaolo Baldassari Monica, Favilla Maria Chiara, Scopigno Roberto. "Storytelling of a Coin Collection by Means of RTI Images: the Case of the Simoneschi Collection in Palazzo Blu", in R. Cherry, N. Proctor (ed.), *Museums and the Web 2013*, 2014.
Radman-Livaja, Ivan. 2014. *Plombs de Siscia*, vol. 9 of Musei Archaeologici Zagrabiensis Catalogi et Monographiae , Zagreb.
Riegl Alois, 1984 (1903). *Der modern Denkmalkulttus*, Vienne, 1903. traduction française par D. Wieczorek, Le culte moderne des monuments. Son essence et sa genèse, Le Seuil.
Robinson, Peter, and Barbara Bordalejo. 2016. "Textual Communities." In *DH2016 Conference*, 876–877. Kraków: Jagiellonian University and Pedagogical University.
Robinson, Peter. 2013. "Toward a Theory of Digital Editions", Variants (10). pp. 105-132.
Sabattini, Brigitte. 2006. "Documenter le présent pour assurer l'avenir in Documentation for conservation and development new heritage strategy for the future", *Actes du XI Forum Unesco Université et Patrimoine* (Florence 11-15 september). https://www.academia.edu/15853068, consulté le 13/11/2016.
Sahle, Patrick. 2013. *Digitale Editionsformen*, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels (3).
Schmidt, Desmond. 2014. "Towards an Interoperable Digital Scholarly Edition." *JTEI Journal*, no. 7.
Schreibman, Susan, Ray Siemens, John Unsworth. 2016. *A New Companion to Digital Humanities*, 2nd Edition. Wiley-Blackwell.
Serres, Michel. 1985. *Les cinq sens*, Hachette Littératures.
Shillingsburg, Peter. 2015. "Development Principles for Virtual Archives and Editions." Variants 11: 9–28.

Tabor, Sharon. W 2007. *Narrowing the Distance: Implementing a Hybrid Learning Model. Quarterly Review of Distance Education* (IAP) (Spring) 8 (1): 48–49.
Terras, Melissa, 2015. "Opening Access to collections: the making and using of open digitised cultural content", in G.E. Gorman, J. Rowley (ed.), *Open Access: Redrawing the Landscape of Scholarly Communication.* Online Information Review, special issue, 733–752.
Vaughan, Norman D. 2010. Blended Learning. In Cleveland-Innes, MF; Garrison, DR. *An Introduction to Distance Education: Understanding Teaching and Learning in a New Era*. Taylor & Francis. p. 165.
Zevi, Bruno. 1959. *Apprendre à voir l'architecture*, Editions de Minuit, Paris.

AIUCD 2017 Conference, 3rd EADH Day, DiXiT Workshop *"The Educational impact of DSE"*
Rome, 23-28 January 2017

WORKSHOP DIXIT

With the patronage of:

It has moving parts! Interactive visualisations in digital publications

James Cummings, Martin J. Hadley, Howard Noble

Abstract:

Digital scholarly editions have only infrequently included interactive data visualizations of the information that's possible to extract from the rich encoding that often serves as their basis. Often editors feel they do not have any 'data' but only 'text', but they are wrong. Abbreviations are marked and expanded but rarely do the editions provide overall statistical analysis of their use. The variations between witnesses are used to construct traditional critical apparatus and sometimes (too often manually) generate a 'stemma codicum' but rarely is the raw data behind this used to create interactive visualizations that readers are able to explore at will. However, moves are increasingly being made in academic publishing towards the embedding of interactive digital visualizations within online academic publications for greater outreach potential. In looking at a number of projects at the University of Oxford this paper will investigate the inclusion of data visualization in digital editions, journal articles, and other forms of online publishing.

For example, the recent Live Data project at the University of Oxford is investigating, creating and publishing interactive data visualizations for academic research projects. This work is enabling researchers to more easily use data visualization in public engagement activities, acts as evidence in research impact statements, and attracts the interest of funders. Simultaneously, in discussions with Oxford University Press, we've helped develop robust methodologies for the stable inclusion of such visualizations inside online publications (and are investigating possibilities for 'offline' digital publications like PDFs). In both cases investigations into suitable data holding repositories (including institutional repositories and third-party services like Figshare or Zenodo) and their roles with respect to live data and the long-term preservation of the data. How is it possible to present data from an edition in a journal article in a stable manner? What if the underlying data is constantly changing (as in crowdsourced contributions)?

Returning to digital scholarly editions (rather than academic outputs based on related research) the paper will look at the creation and generation of adjunct materials to digital editions and question why it is not standard for these to be extracted from editions. Interactive data-rich visualizations based on scholarly digital editions are still fairly rare, but increasingly more of them include such visualizations as timelines, maps, and charts able to be modified by the reader in response to criteria particular to that edition. By investigating the visualizations created for a number of projects, it is suggested how similar interactive data visualizations might benefit digital scholarly editions and their readers.

Vergil and Homer: a Digital Annotation of Knauer's Intertextual Catalogue

James Gawley*, Elizabeth Hunter*, Tessa Little*, Caitlin Diddams*

Student annotators from the University at Buffalo have produced a digital supplement to the connections between Vergil's Aeneid and the Iliad and Odyssey listed in G. N. Knauer's Die Aeneis Und Homer (1964). Student annotators generated additional content beyond what was published in Knauer's work, including a systematic description of the similarities between the passages, and a numeric ranking of the likelihood that each intertext represents a deliberate allusion. The description of similarities will allow future work to determine the most significant language features that identify allusion. A comparison between the numeric rankings and Knauer's system of notation shows that student annotators are capable of reliably distinguishing allusion from more general forms of intertext.

In the first stage of this project, participants developed an annotation scheme. We agreed upon a set of tags to describe the formal similarities between passages, and criteria that would allow annotators to consistently rank allusions on a scale from 1 to 5. Each annotator was assigned thirty intertexts from Knauer's index on a weekly basis. At the end of every week, difficulties in applying the annotation scheme were discussed and necessary changes to the annotation rules were implemented. During this stage, 10% of all intertexts were assigned to multiple annotators, and discrepancies in assessment were resolved. Participants worked primarily from a spreadsheet listing the loci of assigned parallels, and did not base their classification of intertexts on the system of symbols used by Knauer to categorize intertexts.

The second stage of this project began once 400 parallels had been assessed to the satisfaction of all annotators. At this point, the symbols used to tag parallels in Knauer's text were added to the spreadsheet. Comparison of Knauer's symbols to the rankings of graduate student annotators reveals interesting patterns. Most significantly, there is a strong correlation between the annotators' confidence in the intent of Vergil to make a deliberate allusion and the presence of certain symbols in Knauer's notation. It is equally significant that this correlation is not absolute: Knauer occasionally omits the appropriate symbols, or uses them in contradictory ways. At this stage, annotators examined all cases of discrepancy between Knauer's symbols and their own rankings. In some cases this led to a modification of our rankings and a revision of our annotation rules. In other cases, our annotators remain confident in their disagreement with Knauer. Our systematic description of formal similarities shows which features led the annotators to disagree with Knauer.

These findings show that students with a working knowledge of Greek and Latin can be rapidly trained to evaluate intertexts and distinguish cases of deliberate allusion. Digital supplements like the one produced in this project allow students to make a significant scholarly contribution before they have achieved the proficiency of a scholar like Knauer. Educators can use our project as a model for producing digital editions which enhance the versatility of traditional scholarship.

References

Knauer, Georg N. Die Aeneis Und Homer: Studien Zur Poetischen Technik Vergils, Mit Listen Der Homerzitate in Der Aeneis. Göttingen: Vandenhoeck & Ruprecht, 1964.

* University at Buffalo, State University of New York

Wiki Critical Editions: a sustainable philology

Milena Giuffrida, Università degli studi di Catania, milenagiuffrida@gmail.com
Simone Nieddu, Sapienza, Università di Roma, nieddu.sim@gmail.com

Introduction

Wiki Critical Editions are open source and collaborative platforms which host scholars' critical edition and combine scholarly edition's scientific rigor with a flexible and sustainable support, well known in its tools, the Wiki page. Two case studies: Wiki Leopardi (*Canti*'s print editions) and Wiki Gadda (*Eros e Priapo*'s manuscript of the first draft) will show their peculiarities and their advantages, both for scholars and students, as useful research and effective teaching tools.

Advantages

Access to WikiEditions is strictly regulated. Only the research team can actively modify wiki pages, whilst guest users' access is limited to consultation.

Each member works on a module in order to empower the single scholar. Furthermore, every action on each individual site is subjected to cross-check by strictly selected and highly specialized collaborators – thus allowing WikiEditions to stand out among other collaborative edition models which they are inspired to. In fact, European collaborative editions usually implement an open access system which allows users with any ability level and cultural background to authenticate themselves and transcript or correct transcriptions made by other users (let us consider, for example, the Trascribe Desk on Bentham Project -http://www.ucl.ac.uk/Bentham-Project-, in which volunteers can freely transcribe the manuscripts of Jeremy Bentham; or Transkribus - https://transkribus.eu/Transkribus/ -, a University of Innsbruck platform that enables volunteers to contribute to the transcription of documents made available by humanities scholars or archives).

An excessive free access can undermine the validity and rigor of the edition, whereas the restricted access to WikiEditions ensures scientific edition and accuracy in transcription.

Sustainability of WikiEdition compensate for its simple and basic interface. In fact, to realize WikiEditions substantial funding are not required, and continuous cooperation between philologist and computer scientist. Even though they don't know computer languages, users can insert contents of all kinds thanks to the data management method provided by the software. Moreover, platform tools can be easily enhanced by users when needed with new instruments, such as formatting keys.

Wiki Critical Editions

WikiGadda (http://www.filologiadautore.it/wikiGadda/index.php?title=Pagina_principale) is a wiki platform devoted to Carlo Emilio Gadda's works. Based on a Wiki Media adaptation, since 2010 WikiGadda has been linked to the portal filologiadautore.it. The most important and functional section of WikiGadda is dedicated to Gadda's pamphlet *Eros e Priapo*. In this section we can read the critical edition, based on the original manuscript (1944-46), censored in the Sixties, and discovered only in 2010.

The aim of WikiGadda is to make immediately clear the authorial intervention and the nature of variants. Wiki apparatus is free from abbreviations and represents different phases trough platform pages: each variant is clickable and leads to a new page where you can read the new lesson of the text.

The same method has been used for Wiki Leopardi (http://wikileopardi.altervista.org/wiki_leopardi/index.php?title=Wiki_Leopardi), a web platform capable to display the different variants of Giacomo Leopardi's *Canti*, based on Franco Gavazzeni's critical edition (Accademia della Crusca, 2006); a successful collaboration between graduate and undergraduate students of Sapienza University, supervised by Paola Italia.

References

Bentham Project. 2017. University College London. http://www.ucl.ac.uk/Bentham-Project.

Bryant, John. 2002. *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. Ann Arbor: University of Michigan Press.

Bryant, John. 2006. Introduction to *Herman Melville's Typee. A Fluid-Text edition*, http://rotunda.upress.virginia.edu/melville/intro-editing.xqy

Filologia d'autore. 2017. www.filologiadautore.it.

Italia, Paola. 2013. *Editing Novecento*. Roma: Salerno.

Italia, Paola, and Pinotti, Giorgio. 2008. *"Edizioni coatte d'autore: il caso di* Eros e Priapo *(con l'originario primo capitolo, 1944-46)"*. Ecdotica 5 (2008): 7-102.

Leopardi, Giacomo. 1998. *Canti*, by Gavazzeni, Franco, Milano: Rizzoli.

Leopardi, Giacomo. 2009. *Canti*, by Gavazzeni, Franco, and Italia, Paola. Firenze: Accademia della Crusca.

Shillingsburg, Peter. 2006. *From Gutenberg to Google. Electronic representation of literary texts*. Cambridge: Cambridge University Press.

TranScriptorium. 2017. http://transcriptorium.eu/.

Transkribus. 2017. University of Innsbruck. https://transkribus.eu/Transkribus/.

WikiGadda. 2017. http://www.filologiadautore.it/wikiGadda/index.php?title=Pagina_principale.

WikiLeopardi. 2017. http://wikileopardi.altervista.org/wiki_leopardi/index.php?title=Wiki_Leopardi.

PhiloEditor: from Digital Critical Editions to Digital Critical Analytical Editions

Teresa Gargano, Università "La Sapienza" di Roma, teresagargano_91@hotmail.it
Francesca Marocco, Università "La Sapienza" di Roma, francesca.marocco@libero.it
Ersilia Russo, Università "La Sapienza" di Roma, ersilia.russo1@gmail.com

Introduction: what is *PhiloEditor* about

PhiloEditor is a web application that automatically detects the variants of two or multiple drafts of a text, providing a diachronic and stratigraphic display that allows both to study different editions as a Digital Critical Edition and to interact as a digital scholarly infrastructure, useful in scientific and didactic perspective, and generating a brand new kind of edition: *Digital Critical Analytical Edition*.

The project is the result of a Rome University "La Sapienza" and University of Bologna "Alma Mater Studiorum" team research, led by Paola Italia, Claudia Bonsi, Fabio Vitali and Angelo Di Iorio, with the collaboration of Francesca Tomasi, which has been presented already in AIUCD Conference 2014 (Di Iorio et al. 2014) and at the International Conference *ECD/DCE Edizioni a confronto – comparing editions*, University of Rome "La Sapienza", March 27th, 2015 (Bonsi and Italia 2016).

Figure 1. *PhiloEditor 2.0* home page.

Created in 2014 by Fabio Vitali, the software uses *Versioning* by *diffing* technology, lent by the legal domain, to automatically locate all the variants between two texts. In addition to the automatic functions, there is a manual one which allows users to evaluate and categorize phenomena by applying typographical and different colored markers according to the kind of variation. Furthermore, the system allows to classify the same variant in different ways (overlapping markup). The statistics option gives the possibility to summarize and to link data resulting from marking operations, organizing them into pie charts and histograms. As a consequence, both the qualitative and quantitative way of approaching texts leads to simplify the exegetical reflection, combining in the Digital Critical Analytical Edition two different perspectives: a *philological* and an *hermeneutic* one.

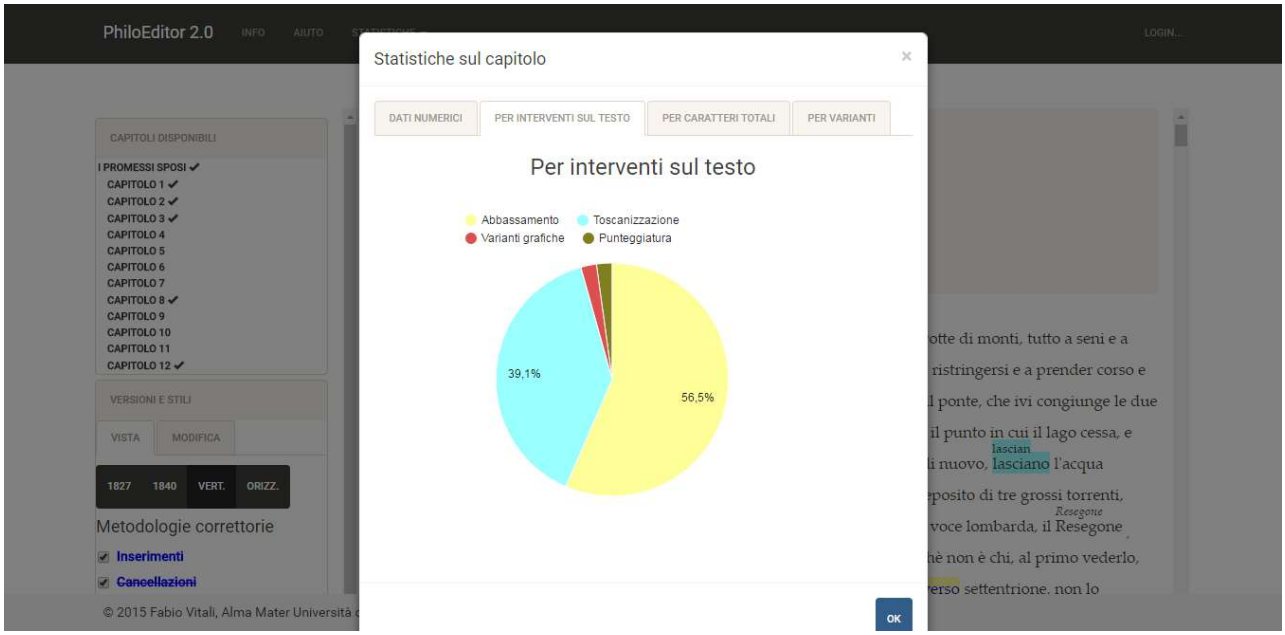


Figure 2. Statistics window on *PhiloEditor 2.0*.

The first version of *PhiloEditor*: *PhiloEditor 2.0*

With the first version of *PhiloEditor* - *PhiloEditor* 2.0 -, the two printed editions of Alessandro Manzoni's *I Promessi Sposi* (Manzoni 2002[1] and Manzoni 2002[2]) have been compared. Manzoni's novel was a perfect case study because of its several editorial variants, that can be considered under different perspectives thanks to the markers adapted on the characteristics of the text diachronic variation. *PhiloEditor 2.0* markers are classified into Methodological and Linguistic Corrections. Methodological Corrections include non-linguistic variations: deletions (~~bold strikethrough~~), dislocation of parts (normal, pink), corrections to avoid repetitions (normal, red), systemic (normal, cyan) and phraseological (underlined, blue) corrections. On the other side, Linguistic Corrections refer to changes caused by literary reasons: linguistic reduction (yellow background), tuscanization (cyan background), graphical (red background) and punctuation mark (beige background) variations. Thus, the application displays visually and intuitively the long editorial work that engaged the author for almost two decades and gives three-dimensionality back to his writing.

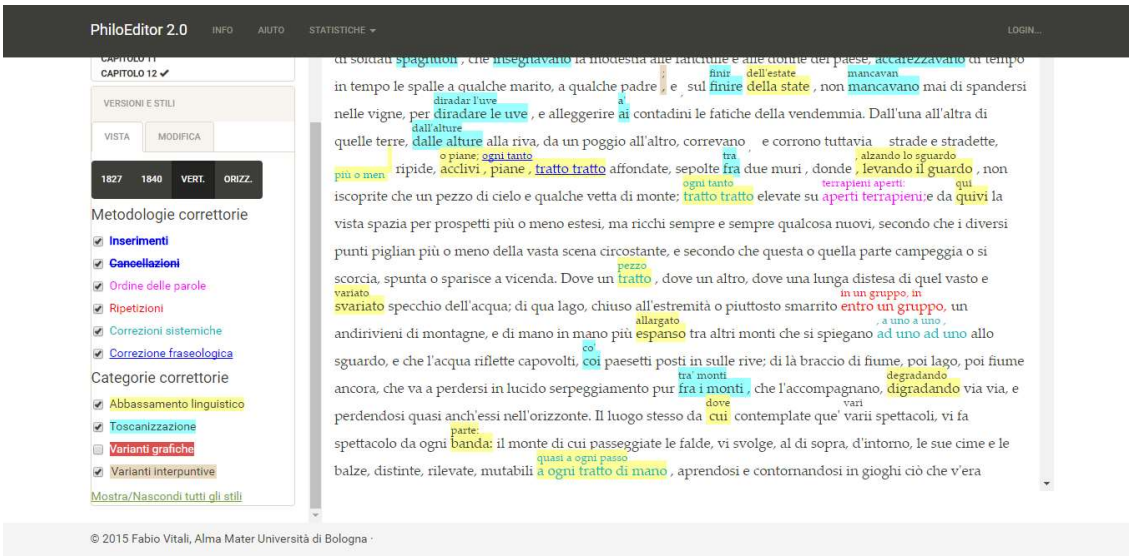


Figure 3. Markup of the first chapter variants.

The second version of *PhiloEditor*: *PhiloEditor 3.0*

The latest version of *PhiloEditor* - *PhiloEditor* 3.0 – (Donati 2015/2016) has been developed into a virtual space capable to manage more texts and different authors. The added value of *PhiloEditor 3.0* consists in its potentialities not only for specialized purposes, but also in teaching educational contexts.

The review and the book version of Carlo Collodi's *Le avventure di Pinocchio* have been chosen to test this new tool (Collodi 1881-1883 and Collodi 1983). Beside a philological-specialized analysis, the markers examine the narratology features, key topics, dialogic structures, characters: new functions which will improve and encourage the use of the platform at every level, both for undergraduate and graduate students.

Figure 4. *PhiloEditor 3.0* home page. In this version, both *I Promessi Sposi* and *Le avventure di Pinocchio* are available.

Conclusions

PhiloEditor can be considered as a new and innovative educational instrument. The layered representation of the variants and their classifications allow all users to perceive the text diachronic evolution immediately in all its peculiarities, encouraging a more active and dynamic approach to Humanities. These new possibilities could make the distance between reader and text shorter and give the chance to put the student's critical faculties to the test. The application also allows to experience a collaborative and participating way of working, that ensures the improvement of the information anytime and anywhere. As the interface is very easy to use, students become able to understand the history of texts, to use philological and exegetics tools, and to develop a critical approach to literary texts.

Primary bibliography

Collodi, C. 1881-1883. *Le avventure di Pinocchio* in "Giornale per bambini".
Collodi, C. 1983. *Le avventure di Pinocchio*, ed. O. Castellani Pollidori, Pescia: Fondazione Nazionale Carlo Collodi.
Manzoni, A. 2002[1]. *I Promessi Sposi* (1827), ed. S. S. Nigro, Milano: Mondadori.
Manzoni, A. 2002[2]. *I Promessi sposi* (1840), ed. S. S. Nigro, Milano: Mondadori.

Secondary bibliography

Bonsi C., Di Iorio A., Italia P., Vitali F., *"*Manzoni's Electronic Interpretations*"*, *Semicerchio* LIII (2/2015): 91-99.
Di Iorio A., Italia P., Vitali F., "Variants and Versioning between Textual Bibliography and Computer Science" (paper presented at AIUCD '14, Bologna, September 18-19, 2014).
Donati, G. "*Philoeditor 3.0: un web editor per la ricerca filologica"* (diss., University of Bologna Alma Mater, 2015/2016).
Bonsi C., Italia P. 2016. *Edizioni a confronto. Comparing Editions*, Roma: Sapienza Università Editrice.
Italia P., Tomasi F., 2014. "Filologia digitale. Fra teoria, metodologia e tecnica", *Ecdotica* XI (2014): 112-130.

Think Big: Digital Scholarly Editions as / and Infrastructure

Anna-Maria Sichani, Huygens ING - University of Ioannina

Digital scholarly editions are originally designed and developed as scholarly outputs of specific research questions, eg. the critical reconstruction and presentation of historical documents or the genetic study of the writing process of a literary work. It starts from a specific research need and then tries to answer as fully as possible this very need having in mind a well-defined audience.
Such a workflow usually results to digital editions that are limited in their purposes, functionalities, uses and, thus, impact and viability in the long-term.

By adopting a more infrastructure-based approach in designing and developing digital scholarly editions we initially accept that the digital edition will not be the only and/or the final output of our undertaking. It is useful, thus, in the initial phase of the design, not only to document the milestones and the deliverables of the digital editing project but also to imagine the potential - often unexpected - (re)uses of the digital edition and its components, the future frameworks of their application as well as the diverse audiences and their expectations. We need, thus, to decide in making technological and operational choices (metadata design, standards, platform-independent processing workflows, robust documentation, open access and open source policies, etc.) that will enable the creative reuse and expansion of the digital edition's components. A number of processing procedures and visualisation techniques further allows the creation of different information layers, various datasets and modular outputs from the digital edition for various purposes and different audiences or stakeholders, eg specialised scholars, educators, pupils and students, general users, publishers, librarians, etc. Such a development model could contribute valuably to a discussion concerning the financial, technological, and security aspects of maintenance and sustainability of digital editions.

Such an infrastructure-based approach in digital scholarly editing is usually undertaken and therefore fits exemplary as part of a broad national and/or institutional initiative. Early examples of this model can be found in enhanced collections of digital textual material (eg. Oxford Text Archive, Cambridge Digital Library), in virtual research environments with digital editions or textual resources for scholarly use (eg. TextGrid, eLaborate) as well as in more experimental proposals for modular design such as the model of minimal and maximal digital edition (Vanhoutte 2012).

The development of digital editions in such a framework becomes an open laboratory to imagine and design outputs that will be evolving, open to interaction and extension over time, and highly customizable; to support and enable the creative reuse that transcends scholarly fields and disciplinary boundaries; to enhance the integration of aspects of digital editing in different communities of practice and social groups; to improve and foster learning and collaboration in unexpected ways; and, finally, to enrich the ongoing and diverse impact and value of digital editions.

This presentation will propose and discuss an infrastructure-based model for the development of digital scholarly editions, by pointing out the challenges and the benefits of such an approach. Furthermore, my aim is to add an interactive part in my presentation by discussing a number of real-world examples and hypothetical case-studies and further asking the audience to assist in designing them within such an infrastructural framework.

Digital Edition of the Complete Works of Leo Tolstoy

Daniil Skorinkin, Higher School of Economics, dskorinkin@hse.ru

Introduction

This paper presents a project aiming to create a complete digital edition of Leo Tolstoy's works with rich structural, semantic, and metadata markup. The project is twofold: its first stage was a massive crowdsourcing effort to digitize Tolstoy's 90-volume comprehensive print edition. That effort, known as 'All of Tolstoy in One Click', received considerable media attention (Bury 2013, McGrane 2013) and attracted more than three thousand volunteers from all over the world. Now that the first goal of 'primary' digitization had been achieved, an obvious next step was to provide the digitized texts with TEI-conformant markup. This work is in progress at the moment. Below we describe both stages of the project (the completed and the ongoing) with a special focus on their social and educational impact.

Source description

More than 46 000 pages of text that collectively contain 14,5 mln words earned Tolstoy a place among the most productive writers of all times. The preparation of the 90-volume print edition started in 1928 (Tolstoy's 100-th anniversary) and took three decades, with last volume published in 1958. The edition is rather diverse: apart from finished works of fiction (prose, poetry, drama), essays and schoolbooks, it contains numerous drafts, letters, volumes of personal diaries, which Tolstoy kept diligently throughout his life, certain number of facsimile manuscripts and drawings, and all sorts of editorial comments. A separate volume is dedicated entirely to alphabetic and chronological indexes. Each volume had 5000 copies, and none of them were ever reprinted, so by the second decade of the 21st century the whole edition was turning into a bibliographic rarity.

OCR and primary digitization (aka 'All of Tolstoy in one click')

The 'All of Tolstoy in one click' project was a joint effort by the Leo Tolstoy State Museum and ABBYY, a Russian software company specializing in optical character recognition (OCR). The initial scanning of the print edition was performed by the Russian state library back in 2006. These images were recognized with help of ABBYY FineReader, proofread several times by volunteers, edited by professional editors and converted into e-books (now available at tolstoy.ru).

Proofreading was the most labour-intensive part of the whole project. Each volunteer was issued a special license for FineReader and a package of 20 unrecognized pages in PDF. Volunteers were supposed to recognize the PDF files using FineReader, correct the automatically identified areas on the pages if necessary (FineReader distinguishes between text, pictures, tables and so on) and then proofread the results of OCR. If the result was not uploaded back within 48 hours after the assignment, these 20 pages were returned in the initial assignments stack. The exchange was organized through a dedicated website readingtolstoy.ru, which now hosts a map with volunteers' locations, press materials about the projects and other related information.

When the organizers announced the call for volunteers, they did not have very optimistic expectations and prepared to carry a fair share of the workload by themselves. The reality, however, proved their pessimism completely wrong. Within two hours after the launch of the crowdsourcing website (readingtolstoy.ru) more than two hundred people signed up and started working already, taking care of

the first 5 volumes. In the end, the entire body of 46820 pages was recognized and proofread within 14 days (8,5 volumes per day) by 3249 volunteers from 49 countries. The most active volunteers processed up to two thousand pages. The leaders were awarded tours to Tolstoy's family estate in Yasnaya Polyana, many other hardworking participants received free e-book readers and OCR software. When interviewed, many of the volunteers noted they could not stop working on the project because they were fascinated by Tolstoy's text and experienced a surge of enthusiasm. Thanks to their hard work the organizers were able to prepare the entire electronic edition (91 original volumes plus 579 separate works extracted from these volumes) in all contemporary e-book formats in just 1,5 years.

TEI markup (aka 'Tolstoy.Digital')

The diversity and scope of the 90-volume edition that we described above obviously call for various digital editorial practices (established or emerging), especially those associated with the TEI standards. To implement these, the second part of the project was launched under the codename Tolstoy.Digital. It is run jointly by the Leo Tolstoy State Museum and the National Research University 'Higher School of Economics'. Though the main managers of the project are university professors and museum researchers, most of the actual research, planning, development and implementation is being done by students specializing in such fields as computational linguistics/NLP, digital humanities and (digital) literary studies. Some work is done in the form of student group projects for which credits are awarded, while other tasks are carried out by individual students as their personal course projects.

On one hand, a lot of effort is being put into re-encoding of the pre-existent metadata and editorial information in the digital environment. One particular example is the footnotes (more than 80 000 of them). Among them editorial and Tolstoy's own comments, explanations and translations, plus all sorts of 'critical edition' style notes. The latter represent diverse editorial 'secondary evidence', e.g. 'here Tolstoy wrote word A first, but then replaced it with an unclear word which is probably word B' or 'this phrase was crossed out with a dry pen, most likely by Tolstoy's wife' or 'original page contained this addition on the margin'. As the size of the material suggests automation, currently our efforts are focused on automatic (or at least machine-aided) classification of notes and their subsequent conversion into TEI tags.

On the other hand, we are trying to augment the markup with new kinds of information that become available as text processing technologies advance. For instance, we have been experimenting a lot with reliable extraction of characters and identification of dialogue between them (with attribution of each speech utterance to its fictional speaker). This data later allows research on differences in the verbal behavior of different characters, which seems to have been a part o Tolstoy's technique. Another area of active research is semantic role labeling within Tolstoy's text (see Bonch-Osmolovskaya and Skorinkin, 2016).

The third major area of our work concerns letters (Bonch-Osmolovskaya and Kolbasov, 2015), which make up one third of the complete works. We have already extracted the metadata (addressee, date, place etc.) from the print edition in TEI format, and are building an extensive search environment/web interface upon it at the moment. Its current version is available at http://digital.tolstoy.ru/tolstoy_search/.

Acknowledgements

This work was supported by grant 15-06-99523 from the Russian Foundation for Basic Research.

References

Bonch-Osmolovskaya A., Kolbasov M. 2015. Tolstoy digital: Mining biographical data in literary heritage editions, in: 1st Conference on Biographical Data in a Digital World 2015, BD 2015; Amsterdam.

Bonch-Osmolovskaya, A.; Skorinkin, D. 2016. Text mining War and Peace: Automatic extraction of character traits from literary pieces. In: Digital Scholarship in the Humanities. Oxford: Oxford University Press.

Bury, L. 2013. Thousands volunteer for Leo Tolstoy digitization. In: The Guardian. https://www.theguardian.com/books/2013/oct/16/all-leo-tolstoy-one-click-project-digitisation

McGrane, S. 2013. Crowdsourcing Tolstoy. In: The New Yorker. http://www.newyorker.com/books/page-turner/crowdsourcing-tolstoy

A spoonful of sugar: encoding and publishing in the classroom

Elena Spadini

This paper pursues the use of text encoding and digital publication in teaching textual criticism.

A number of concepts and rules of textual criticism can be put into practice during a course thanks to the use of digital resources and tools. In dealing with original materials (text sources), the students or participants have to learn the importance of, among others: identify and analyse the document's structure; select relevant features for their research question, establish transcription criteria and conventions, understand the content and identify entities within the text.

These concepts and rules can be addresses through exercises in text encoding. This paper suggests that, in addition to text encoding, an appropriate and not too technical demanding solution for digital publication of the encoded texts will further foster the understanding of these key points.

More and more training courses are now available on how to encode texts, following the Guidelines of the TEI Proposal 5. At the end of these courses, the students or participants have produced a number of documents with markup. While the separation between the encoded text and how it will be rendered is fundamental to descriptive markup, focusing only on the encoding may result in more difficulties for the students to grasp the key concepts of markup and specific practices suggested by the Guidelines. Rendering the encoded texts will thus not only stimulate the participants' enthusiasm, but also foster their overall understanding about markup and its various applications.

The visualization *per se* of TEI data can be accomplished through the TEI transformation framework in oXygen, or through dedicated "lightweight solutions" as TEI Boilerplate and CETEIcean. Another option has recently been released as a common effort from the TEI and the eXist-DB communities, the TEI-Publisher Tool Box. It is based on the TEI-Simple Processing Model, integrated into the native XML database eXist. The Tool Box includes an App Generator, that will automatically create a web application, where to upload the encoded texts and customize the rendition through the ODD if needed. If compared with other publishing framework, the TEI-Publisher offers extra functionalities, due to the fact that it is built upon a database. Two searches options, for instance, are available in the web application automatically generated, one for the text and one for the metadata.

The use of TEI-Publisher in an academic course is underway at the Laboratorio Monaci, a workshop for undergraduate and graduate students held at Sapienza University of Rome, whose goals are the study, promotion and edition of the materials of the Archive of Ernesto Monaci (1844-1918). During a section of this workshop, students are introduced to the Text Encoding Initiative and to how to apply its Guidelines. As soon as the letters are aptly encoded, students are able to upload them into the web application generated through the TEI-Publisher, to browse and search them. When combining the work on the XML editor with exercises on the web application, it may be easier to understand the above-mentioned concepts and procedures: significant text structures are visualized, as well as the relevant features that had been encoded; discrepancy in transcription criteria can be detected, alongside misinterpretation of the references within the text, that may lead to unsatisfactory results of a query.

To conclude, also a number of downside aspects of the use of digital tools, and in particular of publication framework, in the educational context will be discussed.

A Digital Research Platform for Giacomo Leopardi's Zibaldone in the literature classroom

Dr. Silvia Stoyanova, University of Macerata, sms116@caa.columbia.edu

I will present my recent experiences of introducing a digital research platform for Giacomo Leopardi's Zibaldone to Master students in Italian Literature at the University of Macerata and of working with a volunteer student to enhance the platform's editorial apparatus. I would like to discuss the students' feedback on adopting the digital platform to conduct research on the Zibaldone and the project's potential for creating a community of contributors and editors among university students. In conclusion, I will offer some methodological suggestions for the implementation of user experience in the construction of digital scholarly editions.

The Zibaldone project (http://digitalzibaldone.net) addresses the semantic organization of Leopardi's research notes collection which the author indexed thematically and linked with cross-references at the paragraph level with the intention to organize their fragmented discourse into scholarly narratives. The project's premise is that the affordances of the digital medium could articulate this mediation by aligning the fragments into semantic networks and providing scholars with a platform for annotating them further and sharing research results. Opening the platform to a community of editors, which collective knowledge building privileges process over end result (Siemens et al. 2012), is particularly pertinent to the Zibaldone's processual textuality and distributed authorial agency.

Teaching a course on employing the digital platform to one of its targeted user audiences allowed me to probe the collaborative potential of the project while giving participants the opportunity to receive a hands-on introduction to the methods and tools adopted for creating the platform, such as document analysis, encoding in TEI, semantic network visualization, etc. I was able to gather user feedback on the platform's existing and perceived affordances by asking students to conduct thematic research on the Zibaldone both with the print edition and with the digital tool. At the end of the course, students were asked to fill out a questionnaire on the interface design, the functionalities of the platform and their interest in contributing to the project, as well as to share their methodological experience of working with the platform in a short paper. The course therefore tested several scholarly and pedagogical uses of the digital platform which I would like to discuss, namely: the comparison of studying the text with the digital tool and with the print edition; the method of learning how to use a digital edition by learning about its editorial history and doing hands-on exercises exemplifying its key editorial procedures; the level of engagement of students with no prior experience of digital technologies or the use of digital editions; the level of the students' engagement as potential co-editors of the platform; the level of usefulness of their feedback on the platform's future development.

Bibliographic References

Siemens, Raymond et al. "Toward Modeling the Social Edition", Digital Scholarship in the Humanities 27 (4) (2012): 445-461.