CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

AI*IA SERIES

Vol-1509
urn:nbn:de:0074-1509-2

IT@LIA 2015
Intelligent Techniques At LIbraries and Archives 2015

Proceedings of 1st AI*IA Workshop on Intelligent Techniques At LIbraries and Archives co-located with XIV Conference of the Italian Association for Artificial Intelligence (AI*IA 2015)

Ferrara, Italy, September 22, 2015.

Edited by

Stefano Ferilli *
Nicola Ferro **

* University of Bari "Aldo Moro", Department of Computer Science, Via E. Orabona 4, 70126 Bari, Italy
** University of Padua, Department of Information Engineering, Via G. Gradenigo 6/B, 35131 Padova, Italy

Table of Contents

2015-11-09: submitted by Nicola Ferro, metadata incl. bibliographic data published under Creative Commons CC0
2015-11-12: published on CEUR-WS.org |valid HTML5|

CEUR-WS.org

Biblioteche digitali e ontologie

Maurizio Lana, Università del Piemonte Orientale

maurizio.lana@uniupo.it

Abstract. L'annotazione ontologica della conoscenza contenuta nelle opere latine disponibili nella biblioteca digitale digilibLT permette di aprire prospettive di ricerca del tutto nuove non solo per gli studiosi del settore ma anche per il mondo dell'IA in quanto l'ontologia opportunamente utilizzata esplicita in modo formalizzato la conoscenza e offre così una base affidabile a procedure di reasoning. Il modello concettuale descritto per le opere latine può essere esteso ad altre lingue e letterature.

Keywords. biblioteche digitali, digital libraries, ontologia

1 DigilibLT, una biblioteca digitale del latino tardo

La biblioteca digitale del latino tardo DigilibLT è in corso di realizzazione presso il Dipartimento di Studi Umanistici dell'Università del Piemonte Orientale sotto la direzione di Raffaella Tabacco, responsabile del progetto, che ne cura la parte latinistica e di Maurizio Lana che ne cura la parte bibliotecaria e informatica insieme con Fabio Ciotti. Per la letteratura latina arcaica e classica esiste da più di 15 anni fa una raccolta digitale delle opere nota come PHI CDROM, mentre per il latino tardo una collezione completa delle opere costituita da testi di alta e controllata qualità, ad accesso aperto, non esisteva. Cosicché un gruppo di studiosi decise nel 2008 di costruirla, con una iniziativa dal basso tipica delle biblioteche molto specializzate che nascono da specifiche necessità di studio e ricerca.

1.1 Come funziona e che cosa offre

La biblioteca offre essenzialmente quattro tipi di servizi centrati sulle opere:

- schede introduttive: ad eccezione di autori come Boezio o Ammiano Marcellino, la letteratura latina tarda è in gran parte molto poco conosciuta al di fuori dell'ambito degli esperti e dunque per ogni autore ed opera sono disponibili schede critiche di introduzione e presentazione;
- lettura: l'opera può essere letta a schermo con l'aiuto di una struttura di navigazione che corrisponde alla struttura formale dell'opera (suddivisione in libri, capitoli, paragrafi);

- ricerca di testo: è possibile effettuare ricerche su singole opere e autori, o su gruppi definiti per periodo, argomenti, genere letterario, eccetera;
- download: è possibile scaricare le opere nei formati XML/TEI, PDF, TXT, ePUB. In particolare è da sottolineare che i formati PDF, TXT, ePUB sono generati automaticamente, per mezzo di fogli XSLT a partire dall'opera annotata in XML/TEI: ciò permette di ridurre il lavoro tedioso esposto ad errori e di evitare incoerenze tra i contenuti testuali delle diverse versioni dell'opera[1].

1.2 Una biblioteca focalizzata

Un elemento caratterizzante la biblioteca è che il corpus di riferimento (il latino tardo dal I/II d secolo d.C. al 476 d.C.) costituisce una raccolta definita e chiusa, per la quale si può ragionevolmente prevedere che la biblioteca contenga, nel giro di un anno circa, tutte le opere oggi note di tale periodo. Un obiettivo possibile da raggiungere è quello di ampliare la biblioteca da latino tardo al latino arcaico e classico, cosicché essa arrivi a raccogliere tutta quanta la letteratura latina, senza barriere di epoche o tipologie di testi. Fino ad ora ciò non è mai accaduto e quindi nessuno studioso ha mai potuto indagare un tema sull'intero arco cronologico delle opere di quella letteratura, di quelle civiltà. Per quanto complessa da realizzare, si tratta di un'opportunità di grande rilevanza, una sorta di LHA[2] degli studi letterari: da un lato è certo che questa disponibilità darà luogo a nuove letture, a nuove interpretazioni di testi o di fenomeni letterari; dall'altro è altrettanto certo che occorrerà che in qualche misura cambi anche la formazione degli studiosi (se non gli studiosi tout-court).

2 Geolat, o: che cosa fare di una biblioteca digitale

La grande crescita delle biblioteche digitali porta con sé una varietà di questioni, prima fra tutte la sostenibilità: la gestione ordinaria comporta attività specialistiche di alto livello e il sistema informatico che costituisce la biblioteca è esposto ad obsolescenza insieme ai contenuti stessi. Occorre quindi che la biblioteca non veda esaurita la sua funzione nel dare accesso in lettura alle opere, per quanto raffinato esso sia; ma permetta e promuova un *uso* dei testi altrimenti impossibile. E quando scriviamo uso intendiamo proprio il lavoro di ricerca che si fa con i testi e sui testi. Dal lavoro di sviluppo di digilibLT è nata l'idea di usare la biblioteca per studiare uno specifico tema trasversale: la conoscenza geografica contenuta nei testi classici. Il punto di partenza è una biblioteca digitale costituita per portare nel mondo digitale le opere che possediamo a stampa e consultiamo nel mondo fisico; e le opere posseduta dalla biblioteca riproducono le caratteristiche essenziali di un'edizione a stampa (il testo sta-

[1] Particolarmente significativo il formato TXT che è l'unico a permettere a chi lo desideri autonome elaborazioni computazionali.

[2] LHA è il Large Hadron Collider del CERN di Ginevra: l'acceleratore di particelle più grande e potente finora realizzato. Che cosa possa emergere dal suo uso nella ricerca non si sa con certezza, ma che qualcosa di importante emergerà è certo.

bilito indica lacune, integrazioni, espunzioni, e così via). Ma si può fare di più: la conoscenza geografica può essere individuata e descritta in modo altamente formalizzato, quale un'ontologia permette[3].
Annotare ontologicamente in un testo la conoscenza geografica ivi presente è operazione molto complessa in primo luogo perché non può essere automatizzata, se non parzialmente. Nella sua natura profonda l'annotazione ontologica della conoscenza geografica richiede che l'annotatore *possieda la conoscenza geografica* e quindi quando la incontra nel testo la riconosca come tale in tutte le sue valenze e implicazioni e sia in grado di esplicitarle. Peraltro, dal momento che un certo numero di nomi geografici ricorre frequentemente – basti pensare a "Roma" – è utile compiere operazioni di parsing e di NER per individuare almeno un buon numero di nomi geografici e su di essi effettuare l'annotazione senza dover leggere il testo riga per riga, parola per parola. L'altro approccio possibile è quello basato su parsing e NER, senza annotazione ontologica, che permette di procedere più velocemente ma al prezzo della perdita della profondità e ricchezza semantica.

2.1 Perché l'ontologia

Ciò che in genere si evidenzia, quando si spiega che cos'è un'ontologia, è che la conoscenza in un dominio umano viene espressa in modo formalizzato utilizzabile dalle macchine e quindi l'ontologia costituisce un 'ponte' tra gli umani e le macchine. Ma ci sono almeno tre punti di vista da cui l'ontologia è interessante come strumento di studio e di lavoro sui testi:

- esplicitare la conoscenza implicita: qualunque studente o studioso del mondo classico e della letteratura latina possiede un insieme di conoscenze che implicitamente sono coinvolte nella, e attivate e arricchite dalla lettura di un'opera; ma questo insieme di conoscenza essendo frutto di esperienza e formazione individuale non è identico per ogni lettore e pertanto scrivere un'ontologia significa portare alla luce e dichiarare questo insieme di conoscenze; che non è chiuso e dato una volta per tutte se l'ontologia costituisce – come è opportuno – un ambiente di lavoro a cui tutte le parti interessate possono collaborare
- esplicitare in modo formalizzato: questa esplicitazione di conoscenza avviene in modo altamente formalizzato per poter essere utilizzata dalle macchine, come è noto; ma l'esplicitazione formalizzata permette anche altri usi da parte degli umani che l'hanno prodotta, per esempio permette di concepire nuovi tipi di pubblicazio-

[3] Ontologia della conoscenza geografica, non della geografia perché non interessa tanto o solo la struttura dei luoghi e dei territori (una sorta di Geonames dell'antichità), ma anche e soprattutto la percezione dello spazio, in modo in cui lo spazio geografico veniva interpretato e vissuto: percorsi (di spedizioni militari, di pellegrinaggi, di commerci, …), luoghi immaginari (l'Ade, di cui però venivano individuati vari punti di ingresso sulla superficie terrestre: uno a Capo Tenaro nel Peloponneso, uno al lago d'Averno presso Cuma, e altri ancora in Sicilia, in Grecia, …), luoghi connessi con eventi storici (il fiume Rubicone), luoghi connessi con istituzioni (il colle Palatino in Roma, sede del palazzo imperiale di Augusto), e così via

ni: edizioni critiche non più solo centrate sullo stato del testo nei testimoni che lo hanno trasmesso, come da alcuni secoli si fa, ma anche sulla conoscenza contenuta nel testo: geografia, persone/personaggi, ruoli, eventi storici, tempi, ed altro ancora, sono tutti tipi di conoscenza che possono essere espressi in modo formalizzato per mezzo di ontologie e che possono dar luogo a specifiche tipologie di edizioni digitali

- esplicitare in contesto: utilizzando in modo appropriato lo standard TEI di annotazione dei testi, basato sul XML, è possibile annotare in contesto la conoscenza geografica in una modalità che unisce *inline markup* e *standoff markup*: inline si trova il rimando all'ontologia, offline si trova l'ontologia che è concepita come T-box + A-box cioè non solo una struttura concettuale ma anche l'applicazione della struttura concettuale alla descrizione delle entità geografiche, dando così luogo ad una descrizione di conoscenza, una knowledge base che man mano che cresce diventa per certi versi autonoma dai (utilizzabile autonomamente rispetto ai) testi che descrive.

Quest'ultimo aspetto presenta molteplici elementi di complessità perché il testo dell'opera quando viene digitalizzato e annotato diventa ibrido: è il testo dell'autore originario ma innestato con contenuti dell'annotatore che diventa una sorta di coautore contemporaneo, cioè il testo e il commento si fondono e per il fatto che si tratta di codice, essi permettono di generare prodotti digitali del tutto nuovi: si parla spesso (almeno tra coloro che si interessano di questi temi!), di edizioni (critiche) digitali che possono andare dalla riproposizione delle forme tradizionali dell'edizione critica a stampa, costituita da testo stabilito e apparato critico, fino a forme che ne sfruttano in vari modi la caratteristica di codice sorgente da interpretare in modo appropriato[4].
Cercando di vedere le questioni dal punto di vista dell'intelligenza artificiale, ciò che si prefigura in questo modo è non tanto il fatto di estrarre dal testo delle opere la conoscenza in essi contenuta (operazione che appare almeno dal punto di vista degli studiosi del settore molto esposta a difetti e limiti connessi con il fatto che si tratta di un tipo di conoscenza dai confini e contenuti molto sfumati) quanto piuttosto la possibilità di operare su insiemi di conoscenza espressi come mai prima in modo formalizzato nei quali quindi si possono sviluppare procedure di *reasoning* che hanno il vantaggio di partire da premesse formali, e premesse formali certe perché una per una validate da un esperto umano del dominio.

2.2 Di che cosa stiamo parlando

Le prospettive ontologiche di annotazione di conoscenza geografica contenuta in testi latini tardi (e in prospettiva anche testi latini arcaici e classici) sono l'oggetto dell'attività del gruppo di ricerca Geolat, costituito presso il Dipartimento di Studi Umanistici dell'Università del Piemonte Orientale, finanziato dalla Fondazione Com-

[4] Su questo tema si possono vedere, senza pretesa di dare una lista esauriente di fonti, [1], [2] e [5].

pagnia di San Paolo, e diretto da chi scrive[5]. Nel corso del progetto, con un complesso lavoro partito dalla lettura e analisi di una serie di testi latini e da uno stato della questione delle ontologie geografiche (e di interesse geografico in senso lato) esistenti, è stata prodotta l'ontologia geografica GO! (Geolat Ontology ma anche Geographical Ontology)[6], strutturata in 4 moduli (4 ontologie specifiche): TOP che contiene concetti generali, PHY che descrive la geografia fisica, HUM che descrive gli aspetti antropico-sociali, FAR che descrive ciò che è specificamente connesso con il mondo antico.

Il contenuto del progetto è modulare e replicabile in quanto il quadro di riferimento concettuale si può estendere all'interno del latino (dal tardo all'arcaico e classico) o ampliare ad altre lingue e letterature.

2.3 Bibliografia

1. D. Buzzetti, *Digital Editions and Text Processing*, in M. Deegan and K. Sutherland (eds.), "Text Editing, Print, and the Digital World", Aldershot, Ashgate, 2009, pp. 45-62
2. D. Buzzetti, J. McGann, *Critical Editing in a Digital Horizon*, in J. Unsworth, K. O'Brien O'Keeffe, L. Burnard (eds.) "Electronic Textual Editing", The Modern Language Association of America, 2006, pp. 51-71
3. F. Ciotti, F. Tomasi, M. Lana, D. Magro, S. Peroni, F. Vitali, *Dialogue and linking between TEI and other semantic models*, "TEI Conference and Members Meeting 2013", Universitalia, 2013, pp. 145-158
4. M. Lana, F. Tomasi, F. Ciotti, *TEI, ontologies, linked open data: Geolat and beyond*, "Journal of the Text Encoding Initiative", 8, 2015
5. P. Monella, *Why are there no comprehensively digital scholarly editions of classical texts?*, "IV Meeting of digital philology", Verona September 15, 2012, http://www1.unipa.it/paolo.monella/lincei/files/why/why_paper. pdf
6. F. Tomasi, F. Ciotti, M. Lana, D. Magro, S. Peroni, F. Vitali, *Annotating texts with ontologies, from geography to persons and events*, Digital Humanities 2014, 2014, pp. 494-496

[5] Fanno parte del gruppo di ricerca D. Magro, F. Ciotti, R. Afferni, G. Vanotti, C. Meini, M. Benzi, e gli assegnisti T. Tambassi e A. Borgna.

[6] Scritta da C. Corcione, P. De Caro e S. Naro, con la collaborazione di D. Magro, T. Tambassi e M. Lana. Non appena concluse le ultime messe a punto sarà pubblicata online con un permanent URL.