Eurasian Latin Archive

Emmanuela Carbé, Nicola Giannelli
Università di Siena & QuestIT

ELA: FASI DEL PROGETTO, BILANCI E PROSPETTIVE

IL PROGETTO

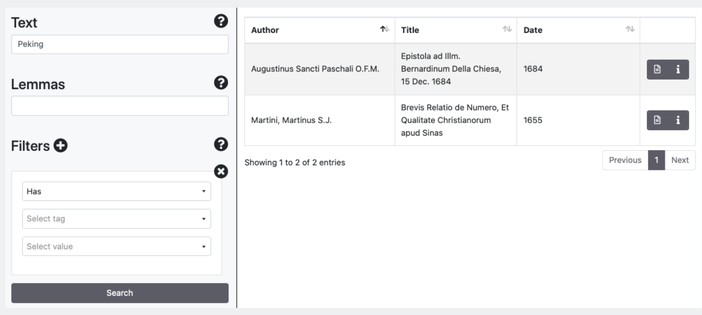
DAS-MeMo (Data mining e analisi statistica su fonti testuali storiche del periodo medievale e moderno) è la fase di start-up di Eurasian Latin Archive, un archivio digitale di documenti latini e multilingua riguardanti l'Asia Orientale dotato di strumenti per analisi linguistiche e semantiche.

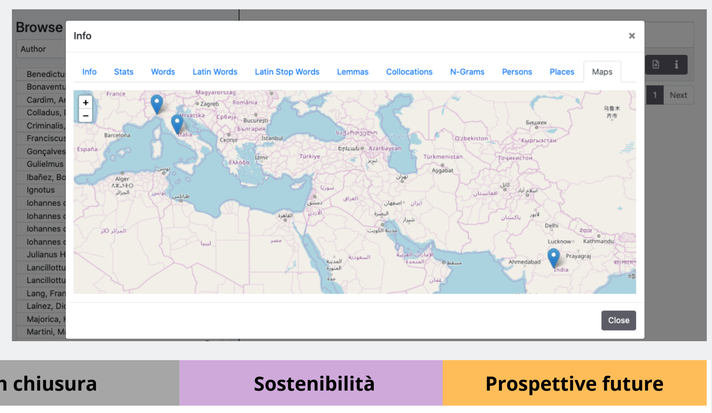
Il gruppo di lavoro, coordinato da Francesco Stella, unisce ricercatori dell'Università di Siena (dip. di Filologia e critica delle letterature antiche e moderne, dip. di Ingegneria dell'Informazione e Scienze Matematiche) e sviluppatori dell'azienda QuestIT, specializzatata in IA e NLP; Pacini Editore collabora al progetto con la pubblicazione di tre e-book. DAS-MeMo ha ricevuto da Regione Toscana il cofinanziamento di un assegno di ricerca di due anni (marzo 2018-febbraio 2020).

In chiusura del biennio ripercorriamo le fasi del progetto, anche sulla scorta della *Checklist for Digital Outputs Assessment* (Ciula 2019a, 2019b) del King's Digital Lab.

AVVIO E PIANIFICAZIONE

1. Realizzazione di un piano del progetto, basato su obiettivi, tempi e risorse; analisi SWOT; valutazione di alcuni casi di studio, in particolare ALIM – Archivio della Latinità Italiana del Medioevo (Russo 2005; Ferrarini 2017; Manos 2018)
2. Obiettivi:
   - 2.1 Creazione di un modello e di un workflow di lavoro; revisione e controllo della qualità
   - 2.2 Definizione e creazione del corpus, con relativa codifica
   - 2.3 Progettazione, analisi dei requisiti e realizzazione del prototipo ELA
   - 2.4 Indagine preliminare di una parte del corpus con primi risultati, pubblicati in e-book grazie all'editore partner del progetto
   - 2.5 Disseminazione e comunicazione dei risultati
   - 2.6 Realizzazione di un piano di sostenibilità
3. Creazione di una lista di specifiche basate su metodo MoSCoW (Must Have, Should Have, Could Have, Would Have)
4. Assegnazione di ruoli all'interno del gruppo di collaboratori e definizioni di prassi e metodi per il lavoro condiviso. ELA utilizza G Suite Education di UniSI come forma di repository non pubblico e come spazio di lavoro (ELA Team Drive)
5. Schede di monitoraggio dello stato di avanzamento dei lavori (SAL) per l'individuazione di azioni correttive in caso di criticità, relazioni intermedie inviate ai soggetti cofinanziatori





ESECUZIONE

Completato | In chiusura | Sostenibilità | Prospettive future

CENSIMENTO

- Censimento di oltre trecento testi candidati per entrare nell'Eurasian Latin Archive
- Gestito con OpenRefine (Hooland, Verborgh e De Vilde 2013; Williamson 2017)
- Contiene metadati descrittivi , strutturali e amministrativi delle risorse (poi in <teiHeader>): informazioni su autore/i (VIAF/Wikidata), titolo, eventuale riferimento all'edizione dell'opera e pp. (Worldcat), link a Bibliotheca Sinica 2.0 e CCT-Christian Texts Database.
- Se il documento è già stato digitalizzato: nome della digital library e link alla risorsa online che si intende utilizzare
- Il censimento funge anche da gestionale per il lavoro di digitalizzazione, trascrizione e codifica in XML/TEI dei testi: i collaboratori che prendono in carico un documento aggiungono progressivamente all'item le indicazioni del processo (nome editor, ORCID, data di inizio del lavoro, punteggio sulla qualità e attendibilità della trascrizione o OCR (da 1 a 3), qualità della codifica TEI (da 1 a 3), eventuale software utilizzato per gli OCR, data di pubblicazione in ELA, indicazione sui diritti di utilizzo del documento, eventuali modifiche dopo la prima versione (data, tipo di modifica, identificativo ORCID del responsabile della modifica).

- piano sostenibile per l'aggiornamento del censimento
- il documento è mantenuto su ELA Team Drive ed è pubblicato in forma semplice (plugin per csv WordPress) sulla piattaforma ELA

CREAZIONE ITEM

- I processi si differenziano in base alla situazione di partenza, una risorsa ad esempio può essere già stata digitalizzata da altre istituzioni e pubblicata online. Il lavoro può comprendere digitalizzazione, OCR e controllo, trascrizioni, codifica TEI, adattamento della TEI di un testo già codificato da altre istituzioni
- Testo codificato in XML/TEI P5 (Tei Consortium 2015) con un modello simile a ALIM
- Semplificazione di alcuni livelli di codifica strutturale e utilizzo di elementi di codifica semantica
- Lo scioglimento delle abbreviazioni (<abbr><expan>) migliora le prestazioni del lemmatizzatore
- <persName> con VIAF o Wikidata
- <placeName> con Pleiades o GeoNames
- <foreign>: marcatura degli inserti in lingua non latina (ISO 639), compreso il Pinyin; sono marcate come lang non ISO, sebbene non siano propriamente lingue, alcune traslitterazioni o rese di pronuncia in latino del cinese e giapponese
- realizzati alcuni script in Python per l'inserimento automatico dei tag in occorrenze già individuate: processano un file .csv arricchito con OpenRefine (VIAF, Wikidata, Pleiades, GeoNames)

- piano sostenibile per nuove digitalizzazioni e trascrizioni, accordi con archivi e istituzioni

- processo automatico di creazione del <teiHeader> in base ai dati forniti dal censimento

- miglioramento dello script di tagging automatico dei file XML con l'introduzione di euristiche più raffinate
- In ELA Tool restituzione di documenti XML/TEI arricchiti
- Implementazione NER di CLTK e messa a disposizione di dataset (<placeName>, <geogName>, <persName>, <date>)

- miglioramento del modello di codifica TEI e integrazione con modelli semantici (Ciotti et al. 2016; Ciotti 2018)

ELA TOOL

ELA Tool è un framework basato su CLTK (Burns, 2019) e NLTK (Bird et al. 2015) per il trattamento dei testi ELA in latino e multilingua. Svolge le seguenti operazioni, restituendo i risultati in formato JSON:

- Parsing <teiHeader>
- Estrazione metadati
- Parsing <text>
- Normalizzazione e cleanup
- Tokenizzazione
- [POS]
- Lemmatizzazione latino (backoff method)
- Statistiche
- Collocations
- Word N-Grams
- Estrazione coordinate geografiche da Pleiades e GeogNames

ELA Tool è disponibile su GitHub

- Evoluzione della NLP pipeline di ELA
- Integrazione di ulteriori processi sui documenti multilingua

- comparazione con altri lemmatizzatori (Mambrini e Passarotti 2019; Eger et al. 2015, 2016)

BACKEND

- Backend Java EE
- Database MySql
- ElasticSearch
- Documentazione

- Estrazione e mappatura della stuttura del documento TEI e della lemmatizzazione di ELA Tool
- Permette ricerche full text (sintassi Lucene), per lemma, per tag, con filtri
- Gestione utenti e ruoli (admin, editor, reviewer, user non autenticato)
- Gestione documento (id, upload, modifica, cancellazione, versioni)
- Esportazione documento (PDF/XML/TXT)
- Gestione versioni di ELA Tool: refresh dei testi quando è disponibile una nuova versione del tool

- Il backend è stato progettato pensando a ulteriori implementazioni, in particolare alla realizzazione di un'interfaccia collaborativa per creare e modificare i documenti con un editor di testo online

- Integrazione di OxGarage

INTERFACCE

1. Interfaccia Generic User

- CMS WordPress
  - Presentazione progetto
  - Guida per l'utente
  - Censimento
  - Notizie
- Motore di ricerca (Javascript):
  - Search (ricerche full text con sintassi Lucene, per lemma, con filtri)
  - 'Browse by' (tag)
  - Lista semplice dei testi
- Risultati di ELA Tool (Javascript)
- Mappe (Leaflet)
- Visualizzatore Tei Boilerplate

2. Interfaccia Content Creator

- Autenticazione
- Ricerca full text
- Caricamento/eliminazione item
- Refresh ELA Tool

- Piano di manutenzione dell'Interfaccia Generic Costumer, in particolare degli aggiornamenti di WP e PHP
- L'interfaccia della Digital Libary è completamente svincolata da WP per limitare gli interventi di manutenzione evolutiva.
- Piano degli aggiornamenti redazionali del sito (sezione news e pagine di progetto)

- User Experience: miglioramento interfaccia responsive e supporto per i dispositivi mobili

- Realizzazione di un visualizzatore dei testi ad hoc con l'uso di xslt

ALTRI OUTPUT

- Sito informativo del progetto di start-up (dasmemo.unisi.it)
- Realizzazione di tre e-book pubblicati in Creative Commons da Pacini (*Hagiographica Coreana I-III,* Pacini 2007-2017 ; P. Intorcetta, *Sapientia Sinica,* 1662; testi di Alessandro Valignano), scaricabili dalla piattaforma ELA e dal sito dell'editore
- Convegno *Global Latin* (31 gennaio-1 febbraio 2019), con la prima riunione del gruppo di lavoro internazionale ELA
- Presentazione del progetto nel in occasione di conferenze, seminari e summer school (QQML 2019 - Firenze, First LiLa Workshop - Milano, DH 2019 - Utrecht, University of Nitra, LiSeH 2019- Graz, V4Py - Prague 2019, Master Infotext UniSi).
- Paper / poster
- Attività di divulgazione durante le giornate di orientamento UniSi e Bright 2018-2019

- Pubblicazione e-book Pacini

- Il sito di DAS-MeMo è sui server di UniSI: non ha bisogno di manutenzione particolare perché è costruito con un generatore di siti statici (Jekyll). Una copia di backup è conservata su ELA Team Drive

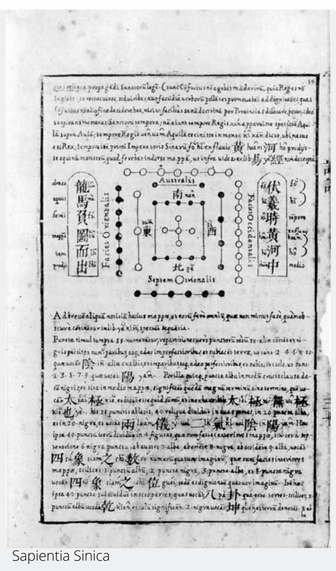
- Prosecuzione delle attività di presentazione e di divulgazione

- Partecipazione a bandi per finanziare il consolidamento e l'evoluzione di ELA

- Trasferimento della piattaforma sulla macchina virtuale ELA di UniSI e pubblicazione
- Pubblicazione dell'Eurasian Latin Archive Handbook e del data model

- Piano di manutenzione della macchina e report di eventuali criticità a cura del Centro di calcolo UniSI
- Piano di manutenzione della piattaforma a cura del gruppo ELA e pianificazione di backup

- Implementazione degli strumenti di analisi linguistica e semantica
- Creazione di strumenti per l'analisi comparata tra i testi con processi in real-time



Sapientia Sinica

CHIUSURA DEL PROGETTO

- Rispetto alla MoSCoW Analysis tutti i "Must Have" e la maggior parte dei "Should Have" sono stati portati a compimento
- Il gruppo di lavoro ha acquisito nuove competenze ed esperienze spendibili in futuri progetti
- Nella fase di consolidamento di ELA andranno prese in considerazione alcune difficoltà riscontrate durante il progetto, che hanno incrementato i cosiddetti "colli di bottiglia". In termini di risultati, le criticità maggiori riguardano la quantità di testi pubblicati (non sufficienti per rilievi statistici) e la carenza di strumenti per le analisi semantiche
- Sono state apportate delle modifiche al progetto di pubblicazione di e-book, dovute a ritardi sull'indagine preliminare complessiva dei risultati di analisi del corpus

LEZIONI APPRESE

- è fondamentale una corretta e continua documentazione del progetto
- un team estremamente interdisciplinare necessita di una maggiore attenzione nella comunicazione tra membri
- ogni incertezza su scelte progettuali (es. modifiche al modello TEI, cambio di strategia per il raggiungimento di un obiettivo) comporta ritardi: vale la pena spendere più tempo nella progettazione, a costo di ridimensionare una parte del progetto. Non è inoltre possibile prevedere tutto fin dall'inizio, non si può sapere ad esempio come evolveranno nel corso di pochi anni le tecnologie che possono essere utilizzate: occorre valutare rischi e benefici di un eventuale cambiamento di rotta dovuto a cause esterne
- attenzione alla sostenibilità: porsi fin da subito il problema della longevità degli output e della prosecuzione del lavoro. Può essere utile cercare di prevedere i futuri interventi di manutenzione evolutiva e pianificare la preservazione a medio e lungo termine;
- la condivisione del progetto, compresi i problemi in itinere, è un'opportunità: il dialogo con studiosi di altre università, italiane e internazionali, è stato fondamentale e spesso ha aiutato a risolvere velocemente alcune criticità
- è necessario trovare un equilibrio tra le effettive disponibilità del progetto (team, finanziamento, tempo a disposizione) e le idee: progettare pragmaticamente, predisporre nell'architettura le basi per implementazioni future

CONTATTI

Centro Studi Comparati I Deug-Su
DFCLAM - Università di Siena
centrostudicomparati@libero.it
www.centroideugsu.unisi.it
www.dasmemo.unisi.it

QuestIT
info@quest-it.com
https://www.quest-it.com

Download poster e riferimenti bibliografici

Riferimenti bibliografici

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

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

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

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

Arianna Ciulia. 2019a. *What Makes Good Honey? KDL Checklist for Digital Outputs Assessment in the REF. Thoughs and relfections from the Lab.* Aug. 7, 2019. URL: https://www.kdl.kcl.ac.uk/blog/checklist-digitaloutputs-ref/

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

Steffen Eger, Rüdiger Gleim and Alexander Mehler. 2016. Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. *Proceedings of the 10th International Conference on Language Resources and Evaluation*. European Language Resources Association:1507-1513. URL: https://www.aclweb.org/anthology/L16-1239

Steffen Eger, Tim Vor der Brück and Alexander Mehler. 2015. Lexicon-assisted Tagging and Lemmatization in Latin: A Comparison of Six Taggers and Two Lemmatization Methods. *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Association for Computational Linguistics:105–113. URL: https://www.aclweb.org/anthology/W15-3716

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

Seth van Hooland, Ruben Verborgh and Max De Vilde. 2013. Cleaning Data with Open Refine. *The Programming Historian*, 2. URL: https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine

Francesco Mambrini and Marco Passarotti. 2019. Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin. *Proceedings of the 13th Linguistic Annotation Workshop (LAW XII),* Association for Computational Linguistics, Florence, 2019, ed. by A. Friedrich and D. Zeyrek:71-80. URL: https://sigann.github.io/LAW-XIII-2019/pdf/W19-4009.pdf

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

Marco Passarotti et al. 2019. Lila: Linking Latin – A Knowledge Base of Linguistic Resources at NLP Tools. *Proceedings of the Poster Session of the 2nd Conference on Language, Data and Knowledge (LKD-PS 2019)*, Leipzig, May 2019, ed. by T. Declerck and J.P. McCrae:20-23. URL: http://ceur-ws.org/Vol-2402/paper2.pdf

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

Tei Consortium. 2015. P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.6.0. Last updated on 16th July 2019. URI: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Evan Peter Williamson. 2017. Fetching and Parsing Data from the Web with Open Refine. *The Programming Historian*, 6. URL: https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine