

Idee, persone, *realia*: un ambiente digitale per la Via della Seta

Emmanuela Carbé¹, Andrea Balbo², Chiara Ombretta Tommasi³, Francesco Stella⁴, Mario G.C.A. Cimino⁵,
Federico A. Galatolo⁶, Chiara Aiola⁷

¹Università di Siena, Italia – emmanuela.carbe@unisi.it

²Università di Torino, Italia – andrea.balbo@unito.it

³Università di Pisa, Italia – chiara.tommasi@unipi.it

⁴Università di Siena, Italia – francesco.stella@unisi.it

⁵Università di Pisa, Italia – mario.cimino@unipi.it

⁶Università di Pisa, Italia – federico.galatolo@ing.unipi.it

⁷Net7, Italia – aiola@netseven.it

ABSTRACT

Il contributo delinea la fase di progettazione di un ambiente digitale collaborativo multilingua dedicato a SERICA (Sino-European Religious Intersections in Central Asia. Interactive Texts and Intelligent Networks), un progetto che riunisce studiosi di diverse discipline per la realizzazione di una piattaforma integrata che raccolga documentazione testuale sulle connessioni tra mondo europeo e mondo orientale in un percorso diacronico molto ampio (dal II sec. a.C. all'età moderna). A partire da esperienze precedenti, e in particolare dal progetto ELA - Eurasian Latin Archive, è stata prodotta un'analisi dei requisiti e un'architettura della piattaforma per la realizzazione di un primo prototipo.

PAROLE CHIAVE

Digital Library, Multilingual Texts, NLP, Artificial Intelligence.

INTERVENTO

1. INTRODUZIONE

SERICA (Sino-European Religious Intersections in Central Asia. Interactive Texts and Intelligent Networks) è un progetto FISIR a cura dell'Università di Pisa e di Torino, in collaborazione con studiosi dell'Università di Ravenna, di Napoli Orientale e di Siena, e con alcune aziende, tra cui Net7. Il progetto, avviato a giugno 2021, si concentra sulle rotte dell'Asia Centrale tra Cina e Europa, con un'estensione diacronica dei documenti presi in esame che parte dal mondo ellenistico e tardoantico e arriva al XIX secolo, adottando dunque una prospettiva di ricerca globale e transdisciplinare in grado di mettere in luce le interconnessioni tra epoche diverse. Per questo ampio attraversamento sono stati individuati tre assi portanti, sintetizzabili nei concetti di 1. idee (scambi tra Oriente e Occidente in ambito filosofico, religioso, antropologico), 2. persone (storiografi, esploratori, missionari che contribuirono alla mediazione tra i due mondi), e 3. *realia* (documentazione su elementi materiali delle civiltà osservate, opere d'arte e oggetti d'uso quotidiano, strumenti scientifici e capi di abbigliamento).

L'obiettivo finale del progetto SERICA è la creazione di un ambiente digitale collaborativo e multilingua, basato anche su tecniche di NLP e di Intelligenza Artificiale, in grado di offrire agli utenti una piattaforma di Digital Library con strumenti di elaborazione, analisi avanzata e annotazione degli oggetti di ricerca, affiancati da percorsi di approfondimento e da strumenti specifici come mappe interattive.

2. BACKGROUND

Il gruppo di SERICA è partito da diversi case studies, prima di tutto dall'esperienza pregressa di ELA – Eurasian Latin Archive (ela.unisi.it, Stella 2021), un progetto biennale dell'Università di Siena cofinanziato da Regione Toscana tra il 2018 e il 2020, e realizzato con la collaborazione dell'azienda QuestIT. Il progetto ELA ha censito un gruppo di testi del XIII-XVIII secolo, digitalizzando un nucleo di documenti in lingua latina, contenenti tuttavia anche porzioni di testo in altre lingue ([6]). ELA ha realizzato un modello di codifica in XML TEI ponendo una particolare attenzione ad alcuni aspetti semantici dei testi, a partire dai nomi di luogo e nomi di persona, trattati da uno script (realizzato in Python) che in parte automatizza il processo di marcatura utilizzando le occorrenze già individuate in precedenza. La Digital Library di ELA dialoga con un framework messo a disposizione su GitHub (ELA Tools: [4]) e basato su CLTK ([3]) e NLTK ([2]): i testi codificati in TEI vengono dunque acquisiti da ELA Tool, in cui si attivano i processi di parsing, normalizzazione, tokenizzazione, individuazione delle parti del discorso, lemmatizzazione, estrazione di collocazioni, n-grammi, e di

coordinate geografiche dei luoghi citati nel testo attraverso i database di Pleiades (pleiades.stoa.org) e GeoNames (geonames.org). I dati così raccolti, insieme ad alcune elaborazioni statistiche, vengono restituiti in formato JSON alla Digital Library, che integra i risultati acquisiti nell'interfaccia utente ([5]).

Oltre a Eurasian Latin Archive, il progetto SERICA è partito da altri strumenti già esistenti, come la biblioteca digitale Sinica 2.0 dell'Università di Vienna, l'archivio ARSI di Roma (Archivum Romanum Societatis Iesu, in parte digitalizzato e disponibile in Internet Archive), il Chinese Christian Texts Database dell'Università di Lovanio e il progetto China Historical GIS (CHGIS) dell'Università di Harvard.

3. ANALISI DEI REQUISITI E REALIZZAZIONE DEL PROTOTIPO

Per il progetto di ambiente digitale è stato costituito un gruppo di lavoro specifico che ha collaborato alla stesura dell'analisi dei requisiti e alla realizzazione dei modelli di dati e dell'architettura della piattaforma. L'ambizione del progetto è quella di creare un ambiente modulare, in grado di unire armonicamente tecnologie, framework e modelli (IIIF per la gestione e l'annotazione delle immagini, codifica del testo in XML TEI, annotazione semantica in RDF, trattamento del linguaggio tramite strumenti NLP e di Intelligenza Artificiale) per rispondere alle esigenze dei gruppi di ricerca e alle maggiori criticità intrinseche al progetto: la prima di queste è l'estrema eterogeneità del *corpus*, costituito da lingue diverse di epoche diverse (testi in greco antico, latino, cinese, lingue persiane, sanscrito, italiano, francese, spagnolo, tedesco), lingue che dovranno essere riconosciute e processate nell'ambito degli strumenti NLP messi a disposizione dalla Digital Library.

Sono stati definiti, all'interno dei requisiti, i ruoli degli utenti (amministratore, editore, autore-revisore, lettore, visitatore) ed è stata data una particolare attenzione alla natura dei metadati da gestire.

Allo scopo di creare uno strumento flessibile e scalabile, il cuore dell'ambiente digitale è stato suddiviso in quattro moduli: 1. SERICA DL (Digital Library e relativo gestionale), che prevede più livelli di rappresentazione degli oggetti digitali: nel caso di documenti testuali la Digital Library include le immagini digitalizzate (se disponibili), il testo codificato, le possibili annotazioni semantiche, l'esito delle elaborazioni NLP; 2. SERICA MAP (Atlante per la navigazione interattiva) gestisce la realizzazione delle mappe, distinte per epoca e per tema e collegate a pagine informative, immagini, riferimenti ai testi della Digital Library, riferimenti ad altri progetti e basi di dati; 3. SERICA NLP (strumenti di analisi linguistica e semantica basati su tecniche di Natural Language Processing) è il modulo che attiva il livello NLP della Digital Library, estrae e gestisce i dati linguistici e i concetti chiave dei testi, con particolare attenzione alla risoluzione dei problemi legati alle diverse lingue presenti nei testi; 4. SERICA AI (strumenti di analisi e di guida alla navigazione basati sull'Intelligenza Artificiale) è il modulo per realizzare la ricerca di testi per similarità tramite algoritmi di apprendimento, e per creare una guida interattiva visuale in base ad algoritmi di raccomandazione.

I moduli interagiranno tra loro attraverso tecniche e formati standard: il singolo item, costituito da un documento XML in una codifica comune e documentata, dotato di identificatore univoco, sarà il nucleo su cui basare la comunicazione stessa – eventualmente anche da e verso agenti esterni – e la persistenza dei dati. Questa scelta può consentire da una parte il riutilizzo di eventuali dati già esistenti attraverso semplici trasformazioni automatiche (ad esempio parte dei documenti pubblicati in ELA), dall'altra la condivisione in evoluzioni o aggregazioni future, nella convinzione che l'obiettivo primario di ogni progetto digitale, ancora prima della costruzione di uno strumento gestionale, sia la preservazione e l'evoluzione del dato, nonché la sua documentazione.

In base all'analisi dei requisiti è in corso di realizzazione un prototipo operativo della piattaforma, realizzato da un gruppo di sviluppatori dell'azienda Net7, concentrato primariamente sul gestionale, sul quale i ricercatori sperimenteranno il processo di creazione di una risorsa all'interno della Digital Library. I quattro moduli saranno dunque accessibili attraverso un unico portale, che presenterà diversi livelli di operatività a seconda della tipologia di utente, mantenendo lo stesso linguaggio grafico e la stessa interfaccia.

4. RINGRAZIAMENTI

Il progetto è finanziato dal Ministero dell'Università e della Ricerca (MUR), nel quadro del programma FISR 2019 - Grant No. 03602.

BIBLIOGRAFIA

- [1] Balbo, Andrea, e Ahn Jaewon. 2019. *Confucius and Cicero: Old Ideas for a New World, New Ideas for an Old World*. Vol. 1. Roma: Sinica.
- [2] Bird, Steven, Klein Ewan, e Edward Loper. 2009. *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. www.nltk.org/book/.
- [3] Burns, Patrick J. 2019. «Building a Text Analysis Pipeline for Classical Languages». In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 159–76. Boston: Berti, M., and De Grutyer, B.

- [4] Carbé, Emmanuela, e Francesco Garosi. 2019. «Ela Tools». GitHub. 2019. <https://github.com/EurasianLatinArchive/ELA>.
- [5] Carbé, Emmanuela, e Nicola Giannelli. 2020. «ELA: fasi del progetto, bilanci e prospettive». In *La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*, 61–66.
- [6] Carbé, Emmanuela, e Nicola Giannelli. 2021. «First steps in building the Eurasian Latin Archive». *QQML Journal* 9: 577–89.
- [7] Stella, Francesco. 2020. «“Global Latin”: Testi latini d’america e d’Asia fra Medioevo e prima età moderna». In 3–298. Siena: SISMELE- Edizioni del Galluzzo.
- [8] ———. 2020. «The Eurasian Latin Archive and the ‘Long’ Latin Middle Ages». In *Digital Philology and Quantitative Criticism of Medieval Literature: Unconventional Approaches to Medieval Latin Literature*, II:219–35. Brepols Publishers.