# ELA: fasi del progetto, bilanci e prospettive

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

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

## Abstract

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

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

## 1 Introduzione

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

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

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

## 2 Metodologia di progetto

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

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

## 3 Analisi del corpus e creazione della piattaforma

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

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

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

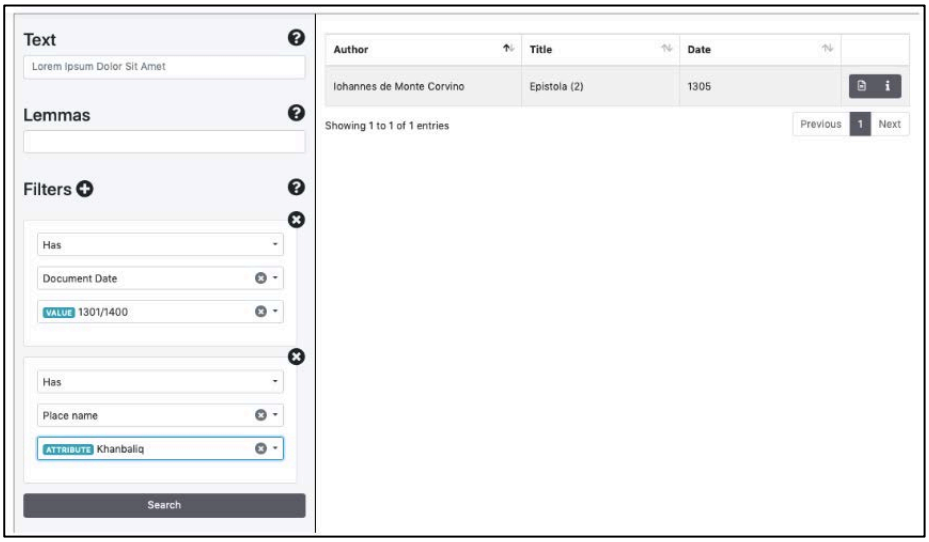


Figura 1. Il prototipo della piattaforma.

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

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

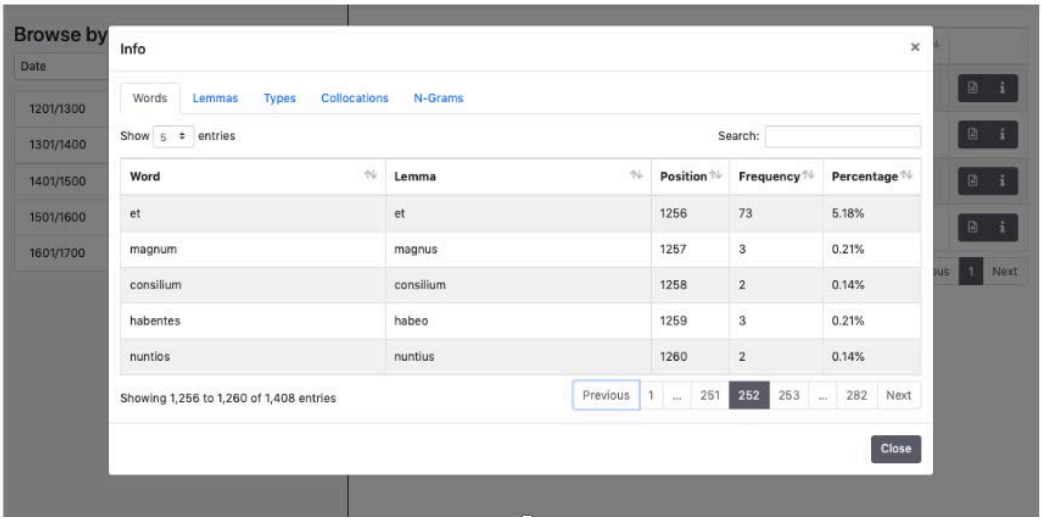


Figura 2. Visualizzazione dei risultati processati dal tool di analisi linguistica.

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

## 4 Bilanci, prospettive

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

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

## Bibliografia


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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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