EpigraFrame

Documento di specifiche tecniche di elaborazione

Autore: Università degli Studi di Napoli «L'Orientale»

Versione del documento: 1.3

Data: luglio 2026

Sommario

1. Scopo del documento 2
2. Obiettivi del Data Product 3
3. Origine dei dati ed eventuali elaborazioni 3
4. Tecnologie utilizzate all'interno del Data Product Lab 5
   4.1. Servizi abilitanti DPaaS 6
   4.2. Eventuali tecnologie esterne a DPaaS 7
5. Pipeline di elaborazione e trasformazione 7
   5.1. Fasi 8
   5.2. Controlli di qualità e gestione degli errori 9
   5.3. Modalità di esecuzione e valutazione 9
6. Esposizione e distribuzione 9
   6.1. Fruizione tramite dataset 10
   6.2. Limitazioni 10
   6.3. Nota sulle condivisioni dei dati 11
7. Possibili evoluzioni future 11
8. Componenti realizzate dal gruppo di lavoro e riuso del Data Product 11
   8.1. Il servizio di scioglimento delle abbreviazioni 11
   8.2. La piattaforma EpigraFrame 12
   8.3. Riuso del dataset 12

1. Scopo del documento

Il presente documento descrive gli obiettivi, le fonti dei dati, le tecnologie del Data Product Lab impiegate, i passaggi seguiti per la costruzione del dataset e le modalità di distribuzione del Data Product «EpigraFrame», realizzato all'interno del Data Product Lab (DP Lab) della piattaforma DPaaS dell'Ecosistema Digitale per la Cultura (Ecomic) del Ministero della Cultura.

Il Data Product è stato sviluppato nell'ambito del Piano Nazionale di Ripresa e Resilienza (PNRR) - Missione 1, Componente 3, Investimento 1.1 «Strategie e piattaforme digitali per il patrimonio culturale». Il documento è rivolto sia a utenti tecnici sia a utenti di dominio che intendano comprendere la natura del Data Product, il processo con cui è stato costruito e le sue modalità di fruizione.

I capitoli da 3 a 6 descrivono la costruzione del dataset così come è stata condotta con gli strumenti messi a disposizione dal DP Lab. Il capitolo 8 riporta una nota sulle componenti realizzate dal gruppo di lavoro nel corso del progetto e sulle possibilità di riuso del Data Product che ne derivano.

2. Obiettivi del Data Product

EpigraFrame è il Data Product realizzato nell'ambito del progetto attraverso l'utilizzo degli strumenti messi a disposizione dal DPLab. Il progetto nasce dall'esigenza di supportare e automatizzare alcune attività fondamentali dell'analisi delle epigrafi funerarie latine, quali la trascrizione del testo inciso, lo scioglimento delle abbreviazioni e il riconoscimento delle entità semantiche. Tali attività sono tradizionalmente svolte in modo manuale, richiedono competenze altamente specialistiche e comportano tempi di lavorazione significativi.

L'obiettivo principale di EpigraFrame è mettere a disposizione il risultato del sistema sviluppato sotto forma di un insieme di dati strutturati, normalizzati, interrogabili e interoperabili. Il Data Product raccoglie, per ciascuna epigrafe, la trascrizione diplomatica del testo inciso, l'edizione con lo scioglimento delle abbreviazioni e le entità semantiche riconosciute.

Gli obiettivi specifici del Data Product sono:

- **ridurre i tempi di lavorazione**, supportando l'automazione delle attività di trascrizione, scioglimento delle abbreviazioni e analisi semantica;
- **limitare la variabilità dei risultati**, mettendo a disposizione dati elaborati secondo criteri omogenei e normalizzati;
- **favorire la consultazione e la ricerca semantica**, consentendo di interrogare il corpus sulla base dei contenuti e delle entità riconosciute;
- **supportare la catalogazione e la valorizzazione digitale** del patrimonio epigrafico;
- **rendere i dati riusabili e interoperabili**, facilitandone l'integrazione con sistemi informativi, cataloghi e altri servizi digitali;

- **fornire un dataset di riferimento** per attività di ricerca e per l'addestramento, la sperimentazione e la valutazione di modelli di intelligenza artificiale;
- **abilitare ulteriori elaborazioni**, a partire dalle immagini, dalle trascrizioni e dalle informazioni semantiche associate alle epigrafi.

EpigraFrame si rivolge a istituzioni culturali e Pubbliche Amministrazioni, università e centri di ricerca, epigrafisti, studiosi di storia antica e sviluppatori di servizi digitali dedicati al patrimonio culturale.

Attraverso il Data Product è quindi possibile migliorare e velocizzare i processi di catalogazione, ricerca e studio delle epigrafi, mettendo a disposizione in forma strutturata e riusabile il risultato del sistema realizzato mediante gli strumenti del DPLab.

3. Origine dei dati ed eventuali elaborazioni

Le immagini e i dati epigrafici provengono da banche dati epigrafiche pubbliche di riferimento, con attribuzione alle rispettive fonti: in particolare Epigraphic Database Roma (EDR, edr-edr.it) ed Epigraphic Database Heidelberg (EDH), i cui dati testuali sono distribuiti in licenza CC BY-SA 4.0, oltre a ulteriori raccolte per la parte fotografica. Le immagini restano di proprietà dei rispettivi titolari e sono utilizzate con attribuzione.

Per le fonti utilizzate si indicano, ove applicabile, gli aspetti seguenti:

- Sistema di origine e titolarità: EDR ed EDH, banche dati epigrafiche gestite dalle rispettive istituzioni scientifiche di riferimento; per la parte fotografica, raccolte ulteriori con attribuzione ai titolari delle immagini.
- Tipologia e formato: immagini raster (JPG) delle lastre epigrafiche e metadati testuali (identificativo di banca dati, provenienza).
- Modalità di acquisizione: selezione e raccolta a cura del gruppo scientifico dell'Università «L'Orientale»; acquisizione dalle rispettive banche dati con attribuzione.
- Frequenza di aggiornamento: le fonti non sono interrogate in modo continuativo; l'acquisizione è avvenuta una sola volta, in fase di costituzione del corpus.
- Volume: 560 epigrafi selezionate, con 559 immagini disponibili.

- Vincoli e licenza: i dati EDR/EDH sono distribuiti in **CC BY-NC-SA 4.0** (**Creative Commons Attribution-NonCommercial - ShareAlike 4.0 International**); le immagini restano di proprietà dei rispettivi titolari e sono utilizzate con attribuzione.
- Qualità e completezza: le schede di origine sono corredate di edizione critica e costituiscono il riferimento per la verifica delle letture; in un numero limitato di casi l'immagine è assente oppure di qualità insufficiente per la lettura automatica, e tali casi sono segnalati nel dataset.

Elaborazioni svolte all'origine, a monte del DP Lab: selezione delle epigrafi per tipologia (funeraria), lingua (latino) e leggibilità dell'immagine; normalizzazione degli identificativi con assegnazione di un codice progressivo interno (G001, G002, …); associazione univoca tra immagine e scheda della banca dati di origine; uniformazione del formato delle immagini. Non è stata effettuata alcuna attività di anonimizzazione, non essendo presenti dati personali di persone viventi.

Le elaborazioni successive - trascrizione, scioglimento delle abbreviazioni e riconoscimento delle entità semantiche - sono state svolte interamente all'interno del DP Lab e sono descritte al capitolo 5.

Il dataset consolidato, esito del processo descritto al capitolo 5, comprende 2.174 entità semantiche validate, distribuite nelle otto categorie seguenti:

Categoria NER	Entità annotate
Persona	988
Relazione	337
Formula iniziale	262
Verbo di dedica	214
Filiazione / patronato	183
Formula biometrica	143
Titolatura	38
Luogo	14

4. Tecnologie utilizzate all'interno del Data Product Lab

La costruzione del Data Product è stata realizzata con gli strumenti messi a disposizione dal DP Lab: le componenti di elaborazione automatica (CPA) esposte dalla piattaforma tramite API REST, lo storage a oggetti per la persistenza delle immagini e degli output

intermedi e l'ambiente di lavoro assegnato al Data Product. Le elaborazioni sul dato - trascrizione, scioglimento e riconoscimento delle entità - sono state eseguite esclusivamente dai servizi abilitanti della piattaforma; l'apporto del gruppo di lavoro si è concentrato sulla preparazione del corpus e sulla validazione degli output, come descritto al capitolo 5.

4.1. Servizi abilitanti DPaaS

Per la realizzazione del Data Product sono stati utilizzati i servizi abilitanti (CPA) e di piattaforma seguenti:

Servizio	Funzione	Input → Output
OCR - «Epigrafi Latine»	Riconoscimento del testo inciso dall'immagine (trascrizione diplomatica).	Immagine → testo
Scioglimento - CPAC	Espansione delle abbreviazioni e normalizzazione del testo.	Trascrizione diplomatica → edizione
NER - «NERs»	Riconoscimento delle entità semantiche in 8 categorie.	Testo → entità (JSON)
Storage a oggetti (S3)	Persistenza di immagini e output intermedi tra le fasi della pipeline.	File → file

Modalità di integrazione: i servizi sono stati invocati tramite le API REST della piattaforma dedicate al Data Product, autenticate con la chiave di consumo assegnata; la sottomissione delle elaborazioni avviene in modalità asincrona e il relativo stato viene interrogato periodicamente fino al completamento. I risultati prodotti da ciascuna fase - trascrizioni, testi con abbreviazioni sciolte ed entità semantiche - vengono acquisiti dalla pipeline e utilizzati come input della fase successiva. PostgreSQL è stato utilizzato per l'archiviazione e la gestione dei dati destinati alle attività di addestramento e validazione.

Configurazioni adottate: la lavorazione è stata organizzata a lotti, per contenere i tempi di coda dei servizi e rendere gestibile la validazione. Per la fase di riconoscimento e trascrizione del testo sono stati utilizzati i servizi CPAI - Epigrafi Latine, CPAI - OCR e CPAC - TrOCR. Il servizio CPAC - Qwen2 è stato impiegato per lo scioglimento delle abbreviazioni e la normalizzazione del testo, mentre CPAC - NERs è stato utilizzato per il riconoscimento e la classificazione delle entità semantiche.

4.2.Eventuali tecnologie esterne a DPaaS

Nessuna tecnologia esterna alla piattaforma ha concorso all'elaborazione dei dati: trascrizione, scioglimento e riconoscimento delle entità sono stati eseguiti dai servizi abilitanti descritti al paragrafo precedente. Al di fuori della piattaforma sono stati impiegati soltanto strumenti di supporto al lavoro redazionale, che non trasformano il dato:

- **Fogli di calcolo (formato .xlsx):** raccolta degli output di ciascuna fase, annotazione delle correzioni degli esperti e consolidamento del dataset in una riga per epigrafe.
- **Registro delle revisioni:** tracciamento, per ciascuna epigrafe, dell'output automatico e del testo approvato, così da rendere misurabile lo scostamento tra i due.
- **Script di appoggio:** sottomissione dei lotti alle API dei servizi abilitanti e raccolta degli output dallo storage a oggetti; sono eseguiti nell'ambiente di lavoro del DP Lab.

Motivazione: gli strumenti elencati assolvono funzioni redazionali e di controllo non fornite direttamente dalla piattaforma e non introducono dipendenze da servizi cloud di terze parti; nessun dato è stato trasmesso a sistemi esterni. Le componenti applicative realizzate dal gruppo di lavoro e integrate all'interno dell'ecosistema Ecomic (DPaaS) nel corso del progetto, tra cui il servizio di scioglimento delle abbreviazioni impiegato nella seconda fase, sono descritte al capitolo 8.

5. Pipeline di elaborazione e trasformazione

Il dataset è stato costruito con una lavorazione in tre fasi di elaborazione automatica, ciascuna affidata a un servizio abilitante della piattaforma e ciascuna seguita dalla validazione degli esperti di dominio. L'ingresso e l'uscita delle fasi seguono una regola fissa: la prima fase riceve l'immagine dell'epigrafe, le fasi successive lavorano sul testo prodotto e approvato nella fase precedente. Nessun output automatico entra nel dataset senza essere stato verificato; ogni output, automatico e validato, è persistito sullo storage a oggetti.

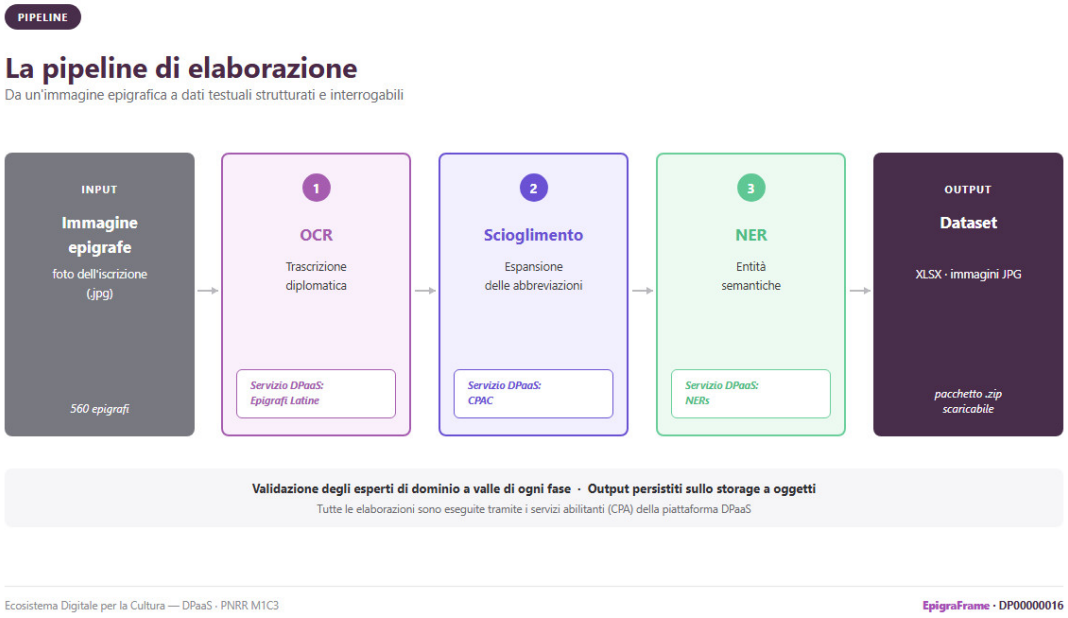


Figura 1 - Le fasi di elaborazione seguite per la costruzione del dataset.

5.1. Fasi

- **Fase preliminare - Preparazione del corpus.** Input: immagini e metadati acquisiti dalle fonti. Operazioni: verifica della leggibilità, assegnazione dell'identificativo progressivo, caricamento delle immagini sullo storage a oggetti nell'area del Data Product. Output: lotto di immagini indicizzato e pronto per la sottomissione.
- **Fase 1 - Trascrizione (OCR).** Input: immagine dell'epigrafe. Servizio: CPA di riconoscimento del testo epigrafico. Operazioni: lettura del testo inciso, riga per riga. Output: trascrizione diplomatica, ossia il testo «come inciso», con sigle e abbreviazioni non sciolte. Validazione: gli esperti confrontano la trascrizione con la fotografia e con l'edizione della scheda di origine, correggono le letture errate e approvano il testo.
- **Fase 2 - Scioglimento delle abbreviazioni.** Input: trascrizione diplomatica approvata. Servizio: CPA di scioglimento. Operazioni: espansione delle sigle e delle abbreviazioni e normalizzazione della grafia. Output: edizione con le parti integrate racchiuse tra parentesi tonde, secondo le convenzioni epigrafiche. Validazione: gli esperti verificano lo scioglimento proposto, con particolare attenzione alle formule ricorrenti e ai casi ambigui, e approvano l'edizione.
- **Fase 3 - Riconoscimento delle entità (NER).** Input: edizione approvata. Servizio: CPA di riconoscimento delle entità. Operazioni: individuazione e classificazione delle porzioni di testo significative. Output: entità nelle otto categorie del dataset (formula iniziale, persona, filiazione/patronato, relazione, titolatura, formula biometrica, verbo

di dedica, luogo). Validazione: gli esperti verificano i confini e la categoria di ciascuna entità e integrano quelle non riconosciute.

- **Fase conclusiva - Consolidamento e confezionamento.** Input: testi ed entità approvati. Operazioni: composizione di una riga per epigrafe con identificativo, provenienza, trascrizione, edizione ed entità; verifica del collegamento tra riga e file immagine; segnalazione dei valori mancanti. Output: dataset in formato Excel e cartella delle immagini, confezionati nel pacchetto distribuito dal Catalogo pubblico.

5.2.Controlli di qualità e gestione degli errori

- Validazione esperta al termine di ciascuna fase: l'output automatico non prosegue verso la fase successiva prima di essere stato verificato e, ove necessario, corretto.
- Confronto con la fonte: nei casi dubbi la lettura è verificata sull'edizione della banca dati di origine e, quando disponibile, su una seconda immagine dell'epigrafe.
- Gestione delle elaborazioni in errore: le sottomissioni non andate a buon fine o rimaste in coda oltre la soglia attesa sono state risottomesse; l'esito di ogni sottomissione è tracciato.
- Gestione degli output assenti: la lavorazione non prosegue in silenzio - il caso è segnalato e il campo viene compilato manualmente dall'esperto oppure marcato come non disponibile nella colonna Note.
- Tracciamento delle correzioni: per ciascuna epigrafe sono conservati sia l'output automatico sia il testo approvato, così da poter misurare l'accuratezza dei servizi.

5.3.Modalità di esecuzione e valutazione

L'elaborazione è stata eseguita su richiesta, a lotti: per ciascun lotto le immagini o i testi sono stati sottomessi alle CPA con elaborazione asincrona e interrogazione periodica dello stato fino al completamento. Non è prevista un'esecuzione pianificata o in tempo reale: il dataset è versionato e aggiornato a rilascio.

La qualità è misurata confrontando l'output automatico con il testo approvato dagli esperti: CER e WER per la trascrizione e per lo scioglimento, Precision, Recall e F1 per il riconoscimento delle entità. Il dataset validato costituisce il Gold Standard di riferimento del Data Product ed è la base rispetto alla quale sono valutate le elaborazioni automatiche.

6. Esposizione e distribuzione

Il Data Product è pubblicato sul Catalogo pubblico dei Data Product della piattaforma DPaaS come prodotto di tipo file: un pacchetto .zip scaricabile che contiene il dataset

delle epigrafi e le relative immagini. La descrizione seguente si mantiene a livello funzionale; il dettaglio dei contenuti è riportato nel manuale utente del Data Product.

6.1. Fruizione tramite dataset

- Formato: pacchetto .zip contenente il dataset in formato Excel (.xlsx) consultabile e le immagini JPG delle epigrafi.
- Organizzazione logica: una riga per epigrafe.
- Chiavi e relazioni: la colonna ID_Epigrafe (es. G001) collega la riga alla relativa immagine (/img/G001.jpg).
- Gestione dei valori mancanti: segnalati nella colonna Note (es. immagine non disponibile).
- Volume e aggiornamento: 560 epigrafi, 559 immagini, 2.174 entità; aggiornamento a rilascio (versionato).

Il dataset è organizzato nelle colonne seguenti:

Colonna	Contenuto
ID_Epigrafe	Identificativo univoco dell'epigrafe (es. G001).
File_Immagine	Nome del file foto nella cartella /img.
Provenienza / Nr_Database	Fonte della scheda e identificativo nel database di origine.
Trascrizione_OCR	Trascrizione diplomatica (testo come inciso).
Trascrizione_Sciolta	Edizione con scioglimento delle abbreviazioni (grafia normalizzata).
8 colonne di entità NER	Formula iniziale, persona, filiazione/patronato, relazione, titolatura, formula biometrica, verbo di dedica, luogo.
Note	Annotazioni (es. immagine non disponibile).

6.2. Limitazioni

- Copertura: il dataset copre 560 epigrafi funerarie latine selezionate e non è esaustivo dell'intero corpus epigrafico.
- Accuratezza: i valori riportati nel dataset sono stati validati da esperti di dominio; restano possibili sviste, in particolare su epigrafi lacunose, usurate o con impaginazioni complesse.
- Dipendenze: la costruzione del dataset dipende dalla disponibilità dei servizi abilitanti DPaaS; i tempi di coda dei servizi incidono sui tempi di lavorazione.

- Immagini: soggette ai diritti dei rispettivi titolari; non ridistribuibili con licenza diversa da quella della fonte.
- Casi non supportati: lingue diverse dal latino ed epigrafi non funerarie non costituiscono l'oggetto del Data Product.

6.3.Nota sulle condivisioni dei dati

- Accesso: dataset e metadati sono pubblici e distribuiti tramite il Catalogo pubblico DPaaS, dal quale il pacchetto .zip è liberamente scaricabile.
- Dati personali: le epigrafi riguardano persone decedute in epoca antica; non sono presenti dati personali di persone viventi.
- Licenza: CC BY 4.0 (Creative Commons Attribuzione 4.0 International) per dati e metadati, con obbligo di attribuzione della fonte.
- Immagini: restano dei rispettivi titolari (EDR, EDH e altre fonti) e vanno attribuite alle fonti di origine.
- Oggetto della condivisione: dato completo (trascrizioni, scioglimenti ed entità) e relativi metadati.

7. Possibili evoluzioni future

Le evoluzioni seguenti rappresentano opportunità di sviluppo individuate durante la realizzazione del Data Product e non costituiscono un impegno progettuale vincolante:

- Ampliamento del corpus e ulteriore addestramento con nuove epigrafi e nuove fonti.
- Estensione a tipologie epigrafiche diverse da quelle funerarie e ampliamento del range cronologico all'età altomedievale e bassomedievale.
- Integrazione con altri Data Product e con i sistemi di catalogazione del Ministero della Cultura.
- Riduzione dell'intervento manuale attraverso il miglioramento dell'automazione e dell'esperienza di revisione.

8. Componenti realizzate dal gruppo di lavoro e riuso del Data Product

I capitoli precedenti descrivono la costruzione del dataset dal punto di vista di chi utilizza gli strumenti messi a disposizione dal DP Lab. Questo capitolo dà conto, per completezza, delle componenti realizzate dal gruppo di lavoro nel corso del progetto e delle possibilità di riuso del Data Product che ne derivano.

8.1. Il servizio di scioglimento delle abbreviazioni

Il servizio impiegato nella seconda fase della lavorazione (capitolo 5) è stato sviluppato nell'ambito del progetto dall'Università degli Studi di Napoli «L'Orientale» - Dipartimento Asia, Africa e Mediterraneo, con SPICI S.r.l. - Officina Digitale in qualità di partner tecnologico, ed è reso disponibile come servizio abilitante della piattaforma DPaaS. Si tratta di un modello multimodale sottoposto a fine-tuning sul dominio epigrafico latino, addestrato su coppie «trascrizione diplomatica → edizione con scioglimento» validate dagli esperti.

Il lavoro di validazione descritto al capitolo 5 ha quindi una doppia funzione: garantire la qualità del dataset distribuito e fornire il materiale di riferimento per la messa a punto e la valutazione del servizio.

8.2.La piattaforma EpigraFrame

Il metodo di lavoro descritto in questo documento - tre elaborazioni in sequenza, ciascuna con la propria validazione - è stato successivamente reso disponibile in un'applicazione web, denominata EpigraFrame come il Data Product che ne ha originato lo sviluppo. La piattaforma invoca i medesimi servizi abilitanti descritti al capitolo 4 ed esegue le medesime elaborazioni; vi aggiunge l'interfaccia di revisione assistita, la conservazione delle correzioni verificate e la gestione della lavorazione per lotti.

8.3.Riuso del dataset

Un dataset costruito secondo i passaggi del capitolo 5 può essere consumato direttamente dalla piattaforma a partire da qualunque dei suoi tre livelli informativi: l'immagine dell'epigrafe, la trascrizione diplomatica o l'edizione con le abbreviazioni sciolte. Poiché le elaborazioni sono le medesime, è possibile:

- partire dall'immagine e rieseguire l'intera sequenza, ottenendo trascrizione, edizione ed entità;
- partire da una trascrizione già disponibile e ottenerne l'edizione con lo scioglimento delle abbreviazioni;
- partire dall'edizione e ottenerne le entità semantiche;
- confrontare l'esito automatico con il valore validato presente nel dataset, per misurare l'accuratezza dei servizi su un riferimento certo.

Il pacchetto pubblicato nel Catalogo pubblico è quindi utilizzabile sia come dataset di riferimento, sia come insieme di dati di ingresso per ulteriori elaborazioni: ogni epigrafe è corredata dell'immagine e dei testi corrispondenti, e ciascuno di questi elementi costituisce un punto di ingresso valido per le medesime elaborazioni impiegate nella sua costruzione.