31 luglio 2026

Finanziato dall'Unione europea NextGenerationEU

MINISTERO DELLA CULTURA

DIGITAL LIBRARY
Istituto centrale per la digitalizzazione del patrimonio culturale

Ecomic ecosistema digitale per la cultura

DPaaS Piattaforma data product as a service

ORGANIZZARE, CATALOGARE, RENDERE ACCESSIBILE “LA BASILICATA NEL MONDO” E “IL FORO” MEDIANTE CREAZIONE DI UN VECTOR DB

Il Data Product realizza l’estrazione automatica del testo dalle scansioni delle riviste storiche “La Basilicata nel mondo” e “Il foro”, producendo contenuti strutturati e metadati semantici utili alla ricerca, all’analisi documentale e alla valorizzazione digitale del patrimonio culturale.

Quando le pagine storiche diventano testo interrogabile, l'archivio rivela persone, luoghi e conoscenze rimaste a lungo nascoste nelle scansioni.

Dalle scansioni a un patrimonio testuale ricercabile

Le riviste storiche **«La Basilicata nel mondo»** e **«Il foro»** conservano testimonianze sulla storia culturale e sociale del territorio: persone, istituzioni, luoghi, eventi e temi presenti nelle diverse edizioni.

Digitalizzate come immagini JPEG o TIFF, le pagine possono essere visualizzate, ma il loro contenuto non è direttamente ricercabile o analizzabile. Individuare una citazione o un argomento richiede quindi la consultazione manuale delle singole scansioni.

In che modo il Data Product genera valore?

Il Data Product analizza i contenuti e identifica emigranti, famiglie, luoghi, date, attività ed eventi come nascite, matrimoni, partenze e arrivi.

Le informazioni vengono organizzate in un catalogo relazionale che collega ogni entità ai documenti e alle testimonianze in cui compare.

Diventa così possibile ricostruire percorsi migratori, riunire materiali riferiti alla stessa persona o famiglia e sviluppare nuove narrazioni sulla memoria della comunità.

Ogni scansione diventa testo, metadati e conoscenza

Livello 1
Qualità tecnica dell'immagine

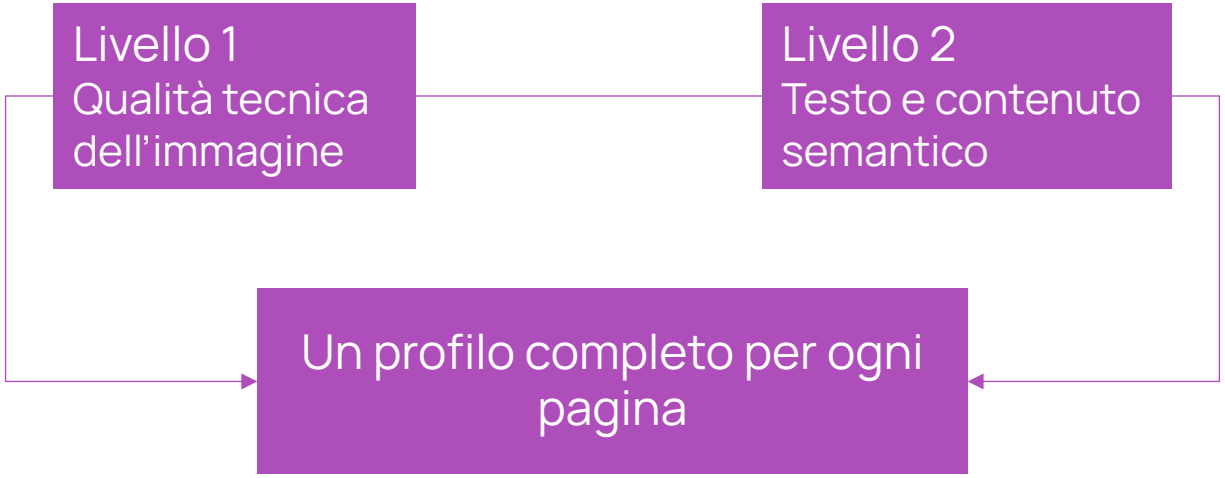
Il sistema estrae automaticamente:
- formato e dimensione del file;
- larghezza e altezza in pixel;
- risoluzione DPI;
- spazio colore e profondità;
- strumento o istituzione di acquisizione;
- data di caricamento nel sistema.

Queste informazioni permettono di verificare qualità, consistenza e caratteristiche del corpus digitalizzato.

Livello 2
Testo e contenuto semantico

L'OCR trasforma la scansione in testo ricercabile. I servizi di intelligenza artificiale producono inoltre:
- una breve descrizione della pagina;
- persone e organizzazioni citate;
- luoghi e riferimenti geografici;
- date e periodi storici;
- eventi e opere menzionate;
- argomenti e parole chiave tematiche.

Relazioni – La conoscenza che collega i contenuti



Metadati tecnici, testo OCR ed entità vengono riuniti nel profilo della scansione e collegati all'immagine originale.
Ogni pagina diventa così ricercabile, analizzabile ed esportabile, mantenendo il rapporto tra il contenuto estratto e la fonte digitalizzata.

La pipeline traduce l'eterogeneità in conoscenza condivisa

1

INPUT

01 · Acquisizione e descrizione

Immagine importata dallo storage, verificata nel formato e controllata per evitare duplicazioni. Apache Tika ne estrae dimensioni, risoluzione, spazio colore e altri metadati tecnici.

2

ANALISI

02 · Lettura e Interpretazione

L'OCR trasforma la scansione in testo ricercabile. I servizi AI generano una descrizione della pagina e identificano persone, organizzazioni, luoghi, date, eventi e temi.

3

OUTPUT

03 · Pubblicazione

Le informazioni vengono normalizzate secondo Dublin Core, ArCo e il Nuovo Soggettario, archiviate in PostgreSQL e rese disponibili tramite profili, API ed export CSV/JSON.

Un servizio asincrono per elaborare le collezioni digitalizzate

Il Data Product separa le operazioni immediate dalle elaborazioni più complesse.

Dopo una prima verifica tecnica, OCR e analisi semantica proseguono in background senza bloccare il servizio.



Acquisizione e risposta immediata

L'immagine viene acquisita dallo storage cloud e verificata nel formato. Il sistema calcola l'impronta SHA-256, estrae i primi metadati tecnici e restituisce un identificativo per monitorare la lavorazione.



OCR e arricchimento in background

Apache Tika rileva le caratteristiche tecniche della scansione.
Successivamente, i servizi AI eseguono:

- OCR e pulizia del testo;
- descrizione automatica della pagina;
- riconoscimento di persone, luoghi, enti e date;
- classificazione degli argomenti.



Normalizzazione semantica

Dublin Core + ArCo + Nuovo Soggettario

Le informazioni estratte vengono ricondotte a categorie e vocabolari condivisi, rendendo coerenti i dati prodotti da pagine e fascicoli differenti.



Persistenza e accesso

Metadati tecnici, testo OCR ed entità vengono archiviati nel database e resi disponibili attraverso profili, viste analitiche, API REST ed export CSV/JSON.

Quattro scenari applicativi

01/ Ricerca nel patrimonio editoriale

Individuare le pagine che citano una persona, un luogo, un'organizzazione, un evento o uno specifico argomento.

02/ Analisi storica del corpus

Esportare testo ed entità per studiare ricorrenze, periodi, temi e reti di relazioni tra i soggetti presenti nelle riviste.

03/ Verifica della digitalizzazione

Consultare il testo OCR di una singola scansione e controllarne qualità, completezza e corrispondenza con l'immagine originale.

04/ Nuovi servizi di accesso

Utilizzare API ed export CSV/JSON per alimentare motori di ricerca, portali documentali, mappe, visualizzazioni e applicazioni semantiche.

Riusabilità

Il Data Product può essere adottato da biblioteche, archivi, fondazioni e istituzioni culturali che conservano riviste, giornali, periodici e documenti disponibili prevalentemente come immagini.

Condizioni per il riuso: l'ente deve selezionare le collezioni, verificare qualità e formati delle immagini e definire le entità e i vocabolari più rilevanti.

Rendere ricercabili le scansioni

JPEG, TIFF e PNG vengono trasformati in testo interrogabile, riducendo la necessità di trascrizione manuale e facilitando l'accesso a collezioni estese.

Arricchire il patrimonio documentale

Il sistema identifica persone, organizzazioni, luoghi, date, eventi e temi, collegandoli alle pagine originali e organizzandoli secondo standard condivisi.

Supportare ricerca e gestione

Bibliotecari e ricercatori possono verificare il testo OCR, consultare i profili delle scansioni, analizzare il corpus ed esportare dati in formato CSV o JSON.

Alimentare nuovi servizi

Testi e metadati possono essere integrati in motori di ricerca, portali, mappe, visualizzazioni, strumenti didattici e applicazioni semantiche.

La digitalizzazione conserva le pagine; i dati strutturati le trasformano in nuove possibilità di ricerca, accesso e conoscenza.