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

L'AQUILA

Raccontiamo i documenti

Manoscritti storici accessibili, comprensibili e multilingua

Trascrizioni • Traduzioni • Spiegazioni • Audio

OBIETTIVI

Manoscritti accessibili a tutti

Pubblico diversificato

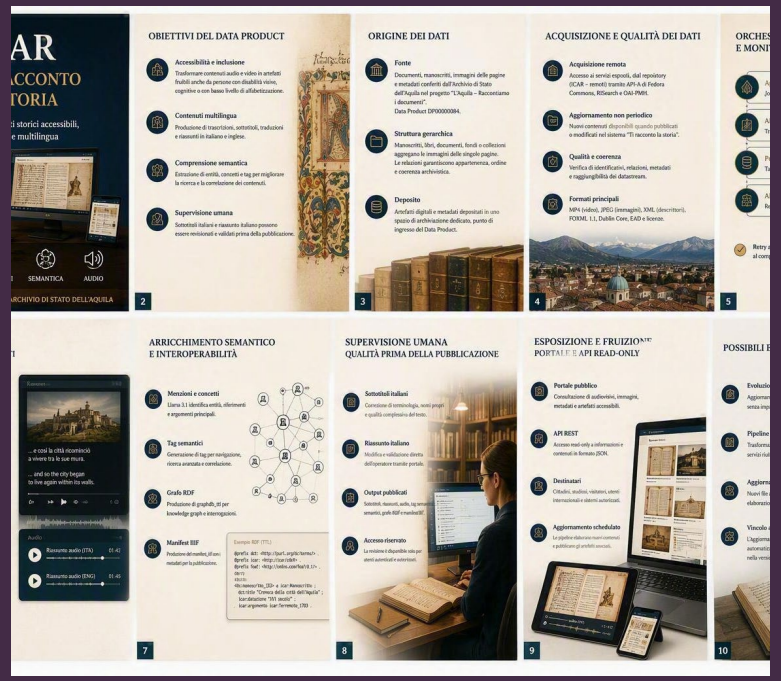
I contenuti complessi sono adattati a utenti con profili cognitivi e livelli di alfabetizzazione differenti.

Fruizione bilingue

Gli artefatti principali sono prodotti in italiano e inglese.

Controllo esperto

Gli operatori possono verificare, correggere e validare gli artefatti generati.



ACCESSIBILITÀ STRUTTURATA

Le immagini delle pagine diventano trascrizioni, traduzioni moderne, spiegazioni e contenuti audio.

ORIGINE DEI DATI

Archivio e struttura gerarchica

Fonte istituzionale

Documenti e manoscritti digitalizzati conferiti dall'Archivio di Stato dell'Aquila.

Gerarchia

Fondi e collezioni aggregano manoscritti, libri e documenti; le pagine sono i contenuti elementari.

Formati



PNG, JPG/JPEG, TIFF/TIF, XLSX e JSON con metadati e ordine delle pagine.

Identificativo

Il Data Product archivistico è identificato come DP00000084.

DATA PREPARATION

Conferimento e qualità

“Dataset, immagini e metadati sono caricati nello spazio di archiviazione dedicato.”

ACQUISIZIONE



Preparazione

Selezione, normalizzazione, associazione record-file e ricostruzione delle gerarchie.



Aggiornamento

Non periodico, legato a nuovi conferimenti o modifiche da parte dell’ente.



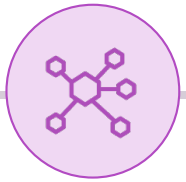
Controlli

Leggibilità, metadati, ordine delle pagine, duplicati, relazioni e immagini mancanti.

ARCHITETTURA

Orchestrazione e monitoraggio

01/



Apache NiFi

Job schedulati individuano contenuti non elaborati e attivano i task.

02/



API Python

Traduzione dei metadati e generazione degli artefatti accessibili.

03/



PostgreSQL

Task, avanzamento, esiti ed errori sono tracciati nel database.

04/



Resilienza

Retry automatico e pubblicazione al completamento delle elaborazioni.

PIPELINE MANOSCRITTI

Dalla pagina alla trascrizione pulita

Trascrizione patch

Qwen2 e Qwen 3.5 Vision riconoscono il testo delle porzioni manoscritte.

Ricostruzione

Le patch vengono concatenate per produrre il testo completo della pagina.

Pulizia finale

Llama 3.1 e Qwen 3.5 Text correggono errori e migliorano leggibilità e punteggiatura.



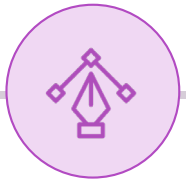
SEGMENTAZIONE

Line Extractor individua righe e paragrafi e suddivide la pagina in patch.

ARTEFATTI ACCESSIBILI

Traduzioni, spiegazioni e audio

01/



Trascrizione

La trascrizione validata genera spiegazioni comprensibili in italiano e inglese.

QWEN 3.5

02/



Italiano moderno

Il testo storico viene trasformato in una versione italiana contemporanea.

LLAMA 3.1

03/



Traduzione inglese

La versione moderna alimenta la traduzione destinata agli utenti internazionali.

LLAMA 3.1

04/



Sintesi audio

Kokoro produce letture audio delle spiegazioni italiana e inglese.

LLAMA 3.1 + KOKORO

LIVELLO DI COLLEZIONE

Una visione complessiva del manoscritto

Corpus documentale

Le traduzioni delle pagine dello stesso manoscritto vengono recuperate e aggregate.

Riassunto complessivo

Creazione Abstract sintetizza temi e contenuti dell’intera raccolta.



Spiegazione italiana

collection_it descrive contesto, temi ricorrenti e valore culturale.

Versione inglese

collection_en rende la visione d’insieme disponibile anche in inglese.

SUPERVISIONE UMANA

Controllo scientifico

Italiano moderno

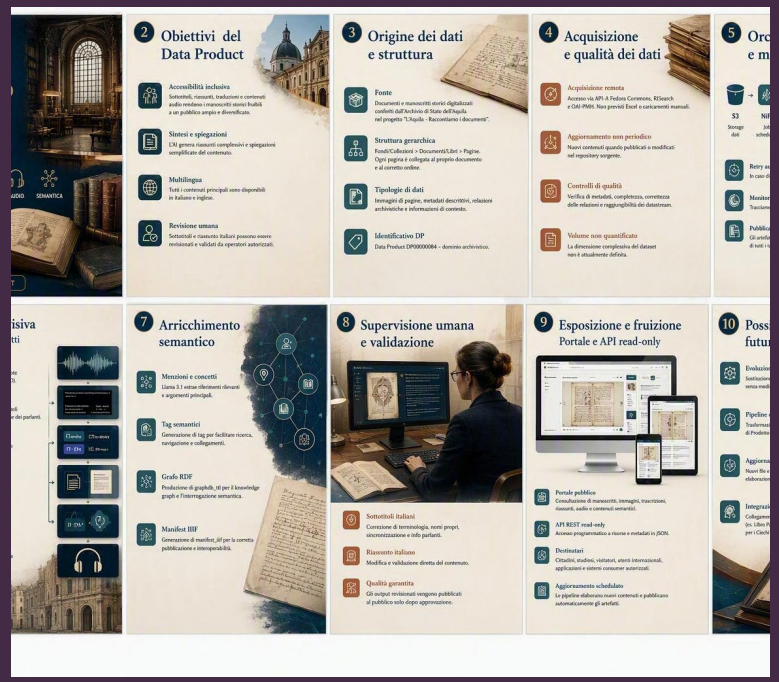
La traduzione contemporanea è modificabile dal portale.

Spiegazione italiana

Il contenuto semplificato può essere migliorato per chiarezza e accuratezza.

Rigenerazione

Le revisioni possono aggiornare traduzioni inglesi e tracce audio dipendenti.



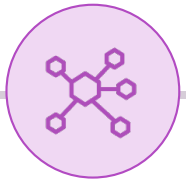
TRASCRIZIONE PULITA

Può essere corretta, integrata e validata dagli esperti.

ESPOSIZIONE E FRUIZIONE

Portale e API read-only

01/



Portale pubblico

Immagini, metadati, trascrizioni, traduzioni, spiegazioni e audio.

02/



Funzioni riservate

SPID/CIE e ruolo autorizzato per revisione, modifica e validazione.

03/



API REST

Ricerca, dettaglio, relazioni, filtri e paginazione con risposte JSON.

04/



Storage protetto

Immagini e audio sono restituiti in streaming senza accesso diretto allo storage.

POSSIBILI EVOLUZIONI

Modelli specializzati sui manoscritti

“Aggiornamento o sostituzione dei modelli AI senza modifiche sostanziali all’orchestrazione.”

EVOLUZIONE DEI MODELLI



Pipeline come CPA

Le logiche possono diventare servizi condivisi e riutilizzabili nella DPaaS.



Aggiornamento dinamico

Nuove pagine potrebbero attivare automaticamente elaborazione e pubblicazione.



Fine tuning dedicato

Per ogni classe omogenea servirebbero circa 150 pagine trascritte e validate come ground truth.