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

ICCU

Il catalogo accessibile

Libri digitalizzati, riassunti e contenuti audio multilingua

Metadati • OCR • Riassunti • Audio

CONOSCENZA SENZA BARRIERE

Lo scopo

Sintesi bilingue

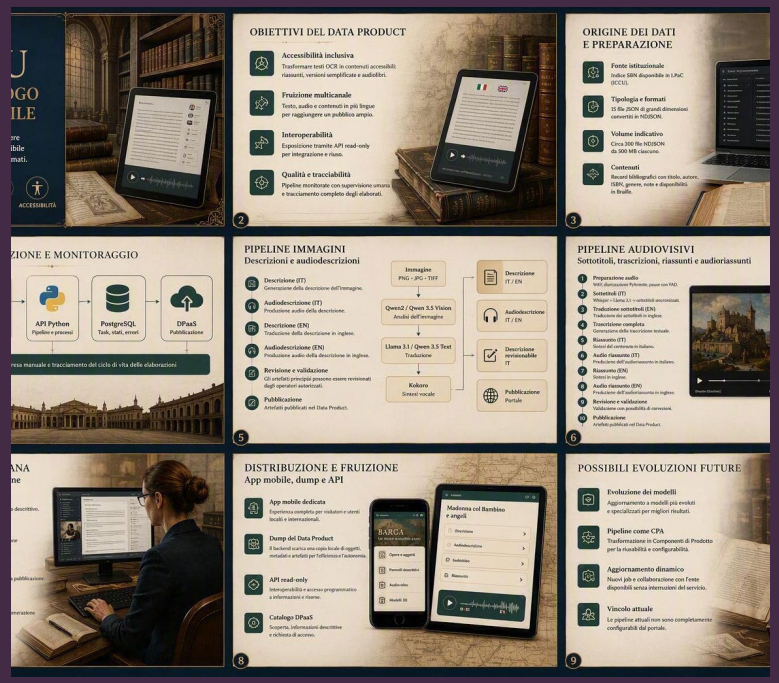
Generare riassunti dei libri in italiano e inglese.

Ascolto multilingua

Produrre versioni audio dei riassunti nelle due lingue supportate.

Catalogo integrato

Collegare metadati bibliografici, contenuti digitalizzati e artefatti accessibili tramite BID.



TESTI ACCESSIBILI

Trasformare testi OCR già disponibili in contenuti sintetici e riutilizzabili nella Digital Library.

CATALOGHI E TESTI DIGITALIZZATI

Tre fonti informative



Indice SBN tramite I.PaC

Record bibliografici originariamente UNIMARC, trasformati nel modello informativo I.PaC.



Aggregatore Digitale

Indice Excel e testi OCR su S3, associati ai record bibliografici mediante BID.



Liber Liber

Dump SQL con metadati e collegamenti alle pagine delle opere sul portale esterno.



Matching progressivo

Associazione tramite ISBN e corrispondenze esatte, normalizzate e fuzzy su titolo e autore.

DA JSON A DATI ELABORABILI

Data preparation e volumi

La fornitura comprende 15 file JSON di grandi dimensioni, identificati tramite BID.

FORNITURA BIBLIOGRAFICA



Conversione NDJSON

I dati originali vengono selezionati, convertiti e suddivisi in circa 300 file da circa 500 MB. I dati rispecchiano circa 25 milioni di record bibliotecari.



Contenuti esterno ad IPAC

Sono presenti 23.664 record OCR; 8.012 corrispondono al catalogo SBN utilizzato per l'output finale.
Sono stati forniti inoltre 5.156 record provenienti dal catalogo online LiberLiber, per cui è stato possible riconciliare circa 3.660 opere con quanto presente nella fornitura.



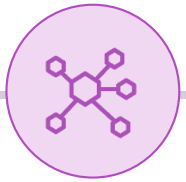
Controllo qualità

Circa 660 record OCR non hanno superato la validazione e non sono stati elaborati.

ARCHITETTURA APPLICATIVA

Orchestrazione e monitoraggio

01/



S3

File e OCR.

02/



NiFi

Job schedulati.

03/



API Python

Traduzione e artefatti.

04/



PostgreSQL

Task, stati ed errori.

DALL'OCR AL RIASSUNTO ACCESSIBILE

Pipeline libri digitalizzati

Segmentazione

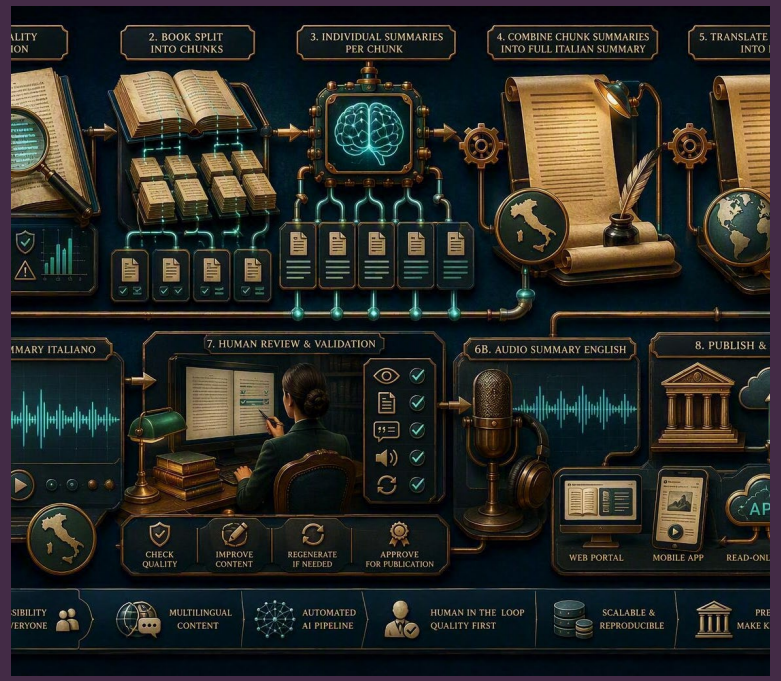
Suddivisione del libro in chunk da circa 8.000 parole.

Sintesi distribuita

Llama 3.1 e Qwen 3.5 Text producono i riassunti dei singoli blocchi.

Riassunto completo

I risultati vengono concatenati e rielaborati nel summary italiano del volume.



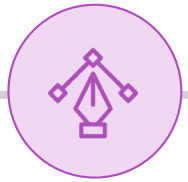
VALIDAZIONE OCR

Analisi delle prime 2.500 parole e interruzione in caso di testo non affidabile.

ITALIANO, INGLESE E AUDIO

Artefatti generati

01/



Validazione OCR

validazione_ocr

02/



Riassunto libro

Italiano e inglese.

03/



Audio riassunto

Italiano e inglese.

04/



Revisione umana

summary_it modificabile e validabile.

QUALITÀ PRIMA DELLA PUBBLICAZIONE

Supervisione editoriale

Riassunto principale

summary_it può essere modificato e validato dagli operatori autorizzati.

Controllo editoriale

La revisione corregge imprecisioni, integra informazioni e adatta il livello di dettaglio.

Input affidabile



OCR vuoti, illeggibili o eccessivamente rumorosi sono esclusi dalle elaborazioni.

Tracciabilità

Gli esiti di validazione e gli artefatti vengono memorizzati nel Data Product.

CATALOGO DPAAS E API READ-ONLY

Esposizione e fruizione

Contenuti consultabili

BID, metadati, disponibilità Braille, cover, preview, manifest, riassunti e audio.

API in sola lettura

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

Aggiornamento batch

Nuove forniture bibliografiche o OCR aggiornano i record tramite BID e avviano le elaborazioni.



CATALOGO ACCESSIBILE

Espone record bibliografici selezionati, riferimenti digitali e artefatti creati dagli OCR.