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

Corpus di immagini da microfilm di manoscritti storici, restaurate e colorizzate mediante AI

Obiettivi del Data Product

Il Data Product, sviluppato per la BNCR nell'ambito del PNRR (Lotto 4 – Dialogo Competitivo), migliora e colorizza automaticamente, tramite intelligenza artificiale, i manoscritti digitalizzati su microfilm in bianco e nero, aumentandone leggibilità e supporto alla ricerca paleografica. La pipeline produce immagini colorizzate e migliorate, varianti di colorizzazione e metadati descrittivi, distribuiti principalmente tramite la piattaforma Chromos.



A chi si rivolge il Data Product?

Un ponte tra patrimonio storico e strumenti digitali: destinato a istituzioni culturali, studiosi e sviluppatori digitali, il Data Product supporta fruizione pubblica e ricerca paleografica.

Origine dei dati ed eventuali elaborazioni

Il Data Product riceve in input 2550 immagini di microfilm digitalizzati, conservate su storage S3, che vengono elaborate tramite un servizio AI dedicato alla colorizzazione.

- Input e volumi

  immagini in formato jpg, tiff, png (con schede XML opzionali da ManusOnline), per un volume di circa 50 GB sia in ingresso che in uscita; gli output vengono compressi per ottimizzare storage e trasferimento.

- Elaborazione

  Le immagini sono sottoposte a preprocessing (resize, normalizzazione, conversione in tensori) e successivamente colorizzate tramite modello AI.

- Delega e tracciamento

  Preprocessing e colorizzazione sono delegati al servizio esterno Chromos, che riceve e restituisce i dati su S3 e ne registra l'esito su un DB relazionale collegato al DP.

Tecnologie utilizzate all'interno del Data Product Lab

Python
Linguaggio di sviluppo su cui è costruito il Data Product, utilizzato per realizzare la pipeline che genera i dati di output

Esterno

FastAPI
Framework tramite cui sono esposte le API REST del DP, per l'accesso alle immagini via URL presigned e la consultazione dello stato di progetti ed elaborazioni

Esterno

Real-ESRGAN
Servizio AI dedicato alla fase di super-resolution delle immagini digitalizzate

Servizio CPA

YOLO-segmentation ed EfficientNet
Servizi AI di supporto all'analisi automatica delle immagini, per la localizzazione di elementi di interesse e per la loro classificazione

Esterno

Algoritmo di colorizzazione Chromos
Modello centrale della pipeline, responsabile della generazione delle immagini colorizzate e delle relative varianti plausibili

Esterno

PostgreSQL
Database relazionale che custodisce i metadati di progetti, manoscritti e dataset, oltre ai record di elaborazione con log e statistiche di processo

Servizio CPA

Pipeline di elaborazione e trasformazione

La pipeline trasforma le immagini sorgente dei microfilm in un pacchetto di dati arricchito per il Data Product. Ogni fase produce le proprie immagini di output in una cartella con un log JSON degli esiti. Al termine, gli output vengono raccolti in un'unica cartella riepilogativa insieme a immagini sorgente e metadati, pronta per il Data Product.

Workflow:

- Miglioramento delle immagini
  Le immagini sorgente vengono elaborate dal modello Real-ESRGAN per aumentarne la qualità e la risoluzione
- Segmentazione
  Individuazione degli elementi visivi rilevanti (miniature, piedi di mosca, iniziali, capolettera) tramite bounding box in formato YOLO
- Classificazione
  Riconoscimento della tipologia di inchiostro e del supporto del manoscritto
- Colorizzazione
  Generazione delle immagini colorizzate a partire dalle immagini originali e migliorate, tramite il modello cGAN
- Composizione del pacchetto finale
  Raccolta di immagini sorgente, output delle fasi precedenti e metadati del manoscritto in un'unica cartella riepilogativa, pronta per l'inserimento nel Data Product

Esposizione e distribuzione

- Fruizione tramite API: accesso programmatico via API REST per consultare progetti, Task e relative metriche; le immagini vengono recuperate direttamente da S3 tramite URL presigned a validità limitata, senza esporre il bucket.
- Fruizione tramite download manuale: dataset organizzato su tre entità PostgreSQL (Project, Task, Manuscript) referenziate rispetto agli oggetti S3; il pacchetto completo è scaricabile dal catalogo DPaaS come archivio compresso con immagini, colorizzazioni e metadati.



Limitazioni

01 **Formati supportati**
Solo JPG, PNG, TIFF in ingresso

02 **Risoluzione fissa del modello**
Input/output vincolati a 512×512 px, con downscaling e successivo riporto alla scala originale

03 **Perdita di dettaglio**
La riduzione forzata degrada nitidezza e informazione visiva, specialmente su TIFF ad alta risoluzione (da ~5000 DPI a ~96 DPI)

04 **Impatto sui casi d'uso**
Per analisi che richiedono alto dettaglio l'immagine colorizzata va intesa come supporto interpretativo, non come sostituto della digitalizzazione originale

Possibili evoluzioni future

Guardando avanti, due direzioni possibili per far evolvere la pipeline:

- Ampliamento dataset di addestramento: evoluzione dei modelli AI tramite più coppie immagine-ground truth per casi meno rappresentati e raccolta strutturata del feedback dei validatori umani, per orientare il fine-tuning verso i casi realmente problematici.
- Evoluzione del modulo di enhancement: nuovi modelli AI per il miglioramento dei microfilm sorgente, con contrasto e saturazione più accurati oltre a super-resolution e denoise, per migliorare la leggibilità e mitigare il limite di risoluzione fissa (512×512 px).

