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

Documento di specifiche tecniche di elaborazione

Sommario

1. Obiettivi del Data Product 3
2. Origine dei dati ed eventuali elaborazioni 5
3. Tecnologie utilizzate all'interno del Data Product Lab 6
3.1. Servizi abilitanti DPaaS 7
3.2. Eventuali tecnologie esterne a DPaaS 7
4. Pipeline di elaborazione e trasformazione 9
5. Esposizione e distribuzione 12
5.1. Nota sulle condivisioni dei dati 15
6. Possibili evoluzioni future 16

Indice delle figure

Figura 1 ..... **Errore. Il segnalibro non è definito.**

Indice delle tabelle

Tabella 1 ..... **Errore. Il segnalibro non è definito.**
Tabella 2 ..... **Errore. Il segnalibro non è definito.**

1.Obiettivi del Data Product

Nell'ambito del Lotto 4 del Dialogo Competitivo (PNRR – Missione 1, Componente 3), il Data Product è stato sviluppato per la Biblioteca Nazionale Centrale di Roma (BNCR) allo scopo di migliorare e colorizzare automaticamente, mediante una pipeline di intelligenza artificiale, i manoscritti digitalizzati su microfilm conservati in bianco e nero. In questo modo se ne incrementano la leggibilità, la fruizione pubblica e il supporto alla ricerca paleografica.



Figura 1 Obiettivi del Data Product

Censito all'interno del catalogo della piattaforma DPaaS, il DP opera nel perimetro delle seguenti macro-categorie:

- Archivistica & Biblioteconomia Digitale: gestione e arricchimento di corpus documentali storici;
- Diagnostica, Restauro & Monitoraggio: strumenti applicati alla conservazione preventiva e al restauro digitale;
- Gestione e arricchimento dei dati: data augmentation e generazione di metadati descrittivi e visivi evoluti per la ricerca avanzata.

Coerentemente con queste macro-aree, il Data Product si rivolge a istituzioni culturali come la BNCR, a studiosi e a sviluppatori di servizi digitali sul patrimonio.

Per quanto riguarda i sistemi, il DP alimenta principalmente la piattaforma Chromos, sviluppata anch'essa nell'ambito del medesimo progetto, che ne costituisce il principale canale di fruizione e integrazione. Il dataset prodotto, tuttavia, non è vincolato a tale piattaforma: qualora se ne rispetti la struttura dei dati, immagini e metadati possono essere inseriti anche in altri sistemi, terzi rispetto a Chromos, per essere visualizzati e resi fruibili agli utenti finali.

Partendo da immagini digitalizzate di microfilm in bianco e nero relativi a manoscritti storici, la pipeline genera immagini migliorate e colorizzate. Gli output prodotti sono:

- immagini colorizzate e migliorate (denoise, normalizzazione del contrasto, super-resolution);
- varianti plausibili di colorizzazione per la medesima immagine sorgente;
- un file manifest.json, che contiene i log decisionali e i dati di riconoscimento automatico (es. rilevamento di miniature);
- eventuali file .xml, con le schede descrittive dei manoscritti estratte automaticamente da ManusOnline.

Due sono i principali casi d'uso individuati:

- Caso d'uso 1 – Arricchimento e fruizione del patrimonio manoscritto: attraverso la pipeline AI, i microfilm di manoscritti antichi digitalizzati di cui dispone la BNCR vengono migliorati, colorati e arricchiti di metadati e schede descrittive; il dataset così ottenuto potrà essere fruito anche da utenti generici tramite piattaforme dedicate alla consultazione di immagini.
- Caso d'uso 2 – Supporto alla ricerca paleografica e storico-linguistica: mediante piattaforme specialistiche dedicate, i dataset arricchiti possono essere interrogati e ulteriormente arricchiti, oltre a essere utilizzati per addestrare nuove pipeline AI di riconoscimento automatico o di colorizzazione.

2. Origine dei dati ed eventuali elaborazioni

Il Data Product ha come input immagini digitali di microfilm, in diversi formati (es. .jpg, .tiff, .png) e con dimensioni variabili a seconda della campagna di digitalizzazione di provenienza. Tali immagini sono conservate su uno storage a oggetti (S3). Accanto alle immagini, possono essere opzionalmente presenti file .xml descrittivi, contenenti i riferimenti alle schede catalografiche dei manoscritti pubblicate sulla piattaforma ManusOnline; tali file, quando disponibili, vengono associati alle rispettive immagini sorgente e utilizzati per arricchire i metadati del corpus in output.

Il volume dei dati in ingresso attualmente gestito dal DP è pari a circa 50 GB. Un volume comparabile caratterizza anche i dati in uscita, gli output prodotti, come: immagini colorizzate, varianti di colorizzazione e file accessori; vengono sottoposti a compressione al fine di contenere l'occupazione dello storage e ottimizzare i successivi tempi di trasferimento e consultazione da parte dei sistemi a valle.

Il flusso di elaborazione si articola in due fasi principali. Nella prima, le immagini sorgente prelevate dal bucket S3 vengono sottoposte a un processo di preprocessing che ne cura il ridimensionamento e la normalizzazione, trasformandole in tensori compatibili con il formato di input atteso dal modello di intelligenza artificiale. Questa fase comprende, tra le altre operazioni, l'uniformazione della risoluzione, la gestione dell'aspect ratio e la normalizzazione dei valori pixel, necessarie a garantire la coerenza dei dati in ingresso al modello e la stabilità delle successive fasi di inferenza.

Nella seconda fase, i tensori così ottenuti vengono elaborati dal modello AI, che produce gli output di colorizzazione e miglioramento descritti in precedenza. È importante sottolineare che né il preprocessing né l'elaborazione AI vengono eseguiti direttamente dal Data Product: entrambe le attività sono delegate al servizio CPA di colorizzazione Chromos, un componente esterno al perimetro del DP e sviluppato nell'ambito del medesimo progetto. Il servizio Chromos riceve in input le immagini sorgente (e gli eventuali metadati associati), esegue le elaborazioni sui propri ambienti di calcolo e restituisce gli output nuovamente sullo storage S3, secondo la struttura dati concordata.

Parallelamente alla produzione degli output, il servizio Chromos traccia l'esito di ciascuna elaborazione su un database relazionale collegato al Data Product, che ne costituisce il registro di stato e consente il monitoraggio dell'intero processo, dall'ingestione delle immagini sorgente fino alla disponibilità degli output finali.

3. Tecnologie utilizzate all'interno del Data Product Lab

Le tecnologie e i componenti impiegati per la realizzazione del Data Product coprono in modo integrato l'intero ciclo di vita dei dati, dall'orchestrazione della pipeline all'esposizione degli output verso i sistemi consumatori. In sintesi, i principali componenti tecnologici sono:

- Python: linguaggio di sviluppo su cui è costruito il Data Product, utilizzato per realizzare la pipeline che genera i dati di output;
- FastAPI: framework tramite cui sono state esposte le API REST del DP, che rendono disponibili le immagini in bianco e nero e colorizzate tramite URL firmati (presigned) e consentono la consultazione dello stato di progetti ed elaborazioni;
- Real-ESRGAN: servizio AI dedicato alla fase di super-resolution delle immagini digitalizzate;
- YOLO-segmentation: servizio AI per il riconoscimento e la localizzazione automatica di elementi visivi di interesse (es. miniature, apparati decorativi);
- EfficientNet: modello di classificazione utilizzato a supporto dell'analisi automatica delle immagini;
- Algoritmo di colorizzazione Chromos: modello centrale della pipeline, responsabile della generazione delle immagini colorizzate e delle relative varianti plausibili;
- PostgreSQL (con accesso asincrono tramite SQLAlchemy): database relazionale che custodisce i metadati di progetti, manoscritti e dataset, oltre ai record di elaborazione (Task) con log e statistiche di processo.

I servizi AI esterni sono orchestrati in un'unica pipeline che riceve in input le immagini sorgente e restituisce i pacchetti elaborati costituenti l'output del DP. Le sezioni seguenti approfondiscono nel dettaglio il funzionamento e le configurazioni dei principali componenti sopra descritti.

3.1. Servizi abilitanti DPaaS

Il servizio abilitante messo a disposizione da DPaaS utilizzato è Postgres che all'interno del processo svolge funzione di database relazionale garantendo la persistenza strutturata dei metadati. Il database viene interrogato in modalità asincrona, tramite l'ORM SQLAlchemy, dalle API FastAPI del DP, ed è inoltre aggiornato dal servizio CPA Chromos al termine di ciascuna elaborazione, ai fini del tracciamento dello stato di avanzamento. In ingresso, PostgreSQL riceve i metadati relativi ai manoscritti, oltre agli aggiornamenti di stato dei singoli task; in uscita, fornisce alle API REST le informazioni richieste in fase di consultazione, quali lo stato dei progetti, i log e le statistiche delle elaborazioni.

3.2. Eventuali tecnologie esterne a DPaaS

Esternamente a DPaaS, per la creazione degli output sono state utilizzate le seguenti tecnologie esterne, integrate in un'unica pipeline dedicata alla produzione degli output del Data Product:

- Real-ESRGAN: è il servizio AI dedicato alla fase di super-resolution delle immagini digitalizzate. Interviene nelle prime fasi della pipeline, a valle dell'acquisizione delle immagini sorgente da microfilm, con l'obiettivo di incrementarne la risoluzione e la nitidezza, compensando la perdita di dettaglio e il rumore tipicamente introdotti dal processo di digitalizzazione da supporto analogico. L'output prodotto da questa fase costituisce l'input di partenza, qualitativamente migliorato, per le successive elaborazioni di riconoscimento e colorizzazione;
- YOLO-segmentation è il servizio AI impiegato per il riconoscimento e la localizzazione automatica degli elementi visivi di interesse presenti nella pagina manoscritta, quali miniature, piedi di mosca, iniziali e capolettera. Il modello individua la posizione e l'estensione di ciascun elemento all'interno dell'immagine, generando le informazioni di localizzazione che vengono successivamente riportate nel file manifest.json che possono essere utilizzate a fini descrittivi;
- EfficientNet è impiegato all'interno della pipeline in due distinte modalità di classificazione, ciascuna basata su un modello dedicato e specificamente addestrato.
  Il primo modello è stato personalizzato per la classificazione del supporto scrittorio del manoscritto, distinguendo tra supporto pergamenaceo e supporto cartaceo. Questa informazione, derivata dall'analisi automatica dell'immagine, arricchisce i metadati descrittivi del manoscritto. Il secondo modello, sviluppato successivamente, è invece dedicato alla classificazione del colore dell'inchiostro

rilevato sui microfilm, a partire dall'analisi delle aree di testo individuate nell'immagine.

- L'algoritmo di colorizzazione rappresenta il componente centrale della pipeline ed è basato su un'architettura di tipo cGAN (Conditional Generative Adversarial Network), addestrata per generare la colorizzazione plausibile delle immagini di manoscritto a partire dal corrispondente microfilm. Il modello riceve in input l'immagine preprocessata e restituisce in output una o più immagini colorizzate, corrispondenti a varianti plausibili della medesima immagine sorgente. La natura generativa e condizionata dell'architettura consente di produrre risultati coerenti con le caratteristiche specifiche di ciascun manoscritto, pur mantenendo la possibilità di generare interpretazioni cromatiche alternative per una medesima pagina.

Real-ESRGAN, EfficientNet e YOLO-segmentation si basano su architetture open source con licenza permissiva, senza vincoli contrattuali rilevanti; L'algoritmo di colorizzazione cGAN, sviluppato internamente e addestrato su un dataset proprietario di coppie bianco/nero-colore, non comporta vincoli di licenza esterni ma dipende fortemente dalla qualità del dataset e dalla disponibilità di risorse GPU per training e inferenza.

4. Pipeline di elaborazione e trasformazione

La pipeline di creazione dei dati viene avviata dal data producer che, tramite apposite API, richiede l'elaborazione di un dato pacchetto di microfilm, fornendo come parametro il riferimento S3 alle immagini sorgente.

La prima fase è quella di miglioramento delle immagini: le immagini originali vengono scaricate dal bucket S3 e passate al modello Real-ESRGAN. L'output di questa fase è una cartella contenente le immagini migliorate e un file di log manifest.json, che riporta l'elenco delle immagini processate e, in caso di problemi, il relativo messaggio di errore.

Segue la fase di segmentazione, che opera anch'essa sulle immagini originali scaricate da S3 per individuare gli elementi visivi rilevanti (miniature, piedi di mosca, iniziali, capolettera). L'output è una cartella con un file manifest.json che, oltre a eventuali errori, riporta per ciascuna immagine gli elementi rilevati sotto forma di bounding box in formato YOLO.

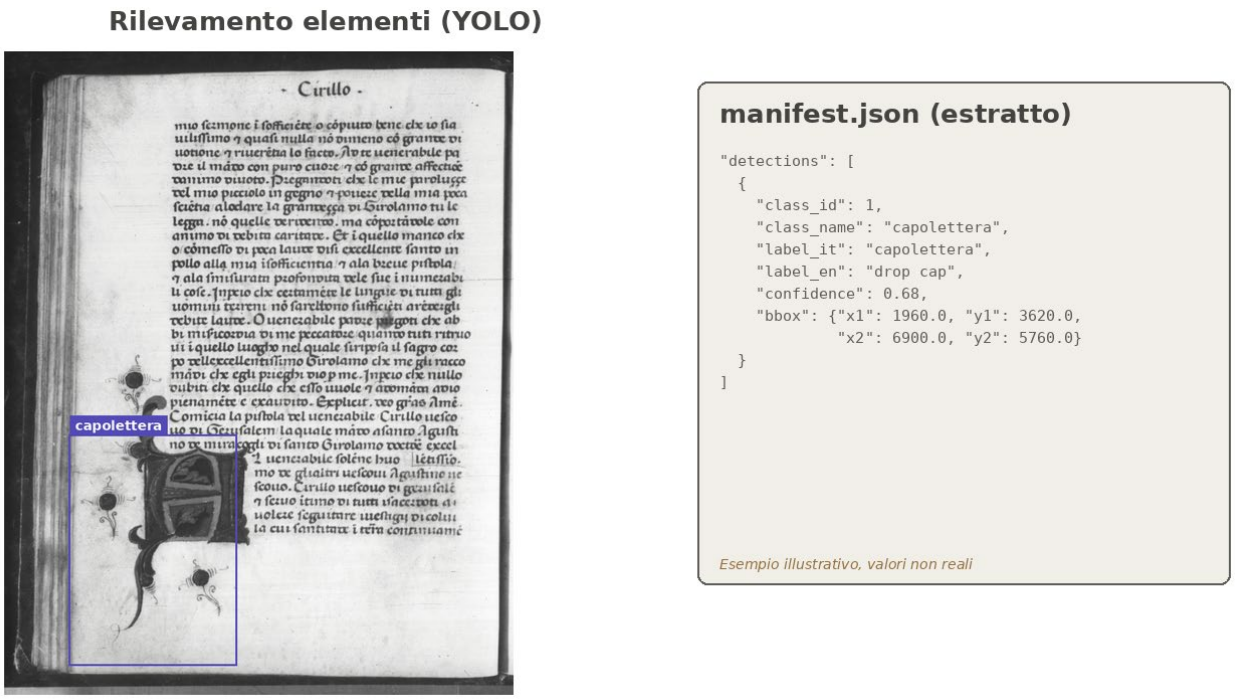


Figura 2 Esempio di output della segmentazione AI su un'immagine

Vengono quindi eseguite le fasi di classificazione, distinte per tipologia di inchiostro e per tipologia di supporto del manoscritto: ciascuna classificazione produce una propria cartella con un file manifest.json in cui, oltre a eventuali errori, è indicata la classe rilevata per ogni immagine.

L'ultima fase di elaborazione è quella di colorizzazione, che utilizza sia le immagini originali sia quelle migliorate come input del modello cGAN. Per ciascuna elaborazione viene creata una cartella contenente le immagini colorizzate prodotte e un file JSON dedicato alla registrazione di eventuali errori.

A completamento del pacchetto da inserire nel Data Product viene infine composta una cartella riepilogativa, contenente: le immagini sorgente originali; il file manus.xml con le informazioni del manoscritto estratte da ManusOnline; un file skipped.json, che elenca le eventuali immagini non elaborate a causa di problemi riscontrati; e il file pipeline_manifest.json, riepilogo complessivo di tutte le operazioni effettuate su ciascuna immagine del pacchetto di input.

L'intera pipeline viene eseguita per ogni pacchetto di immagini destinato all'inserimento nel Data Product.

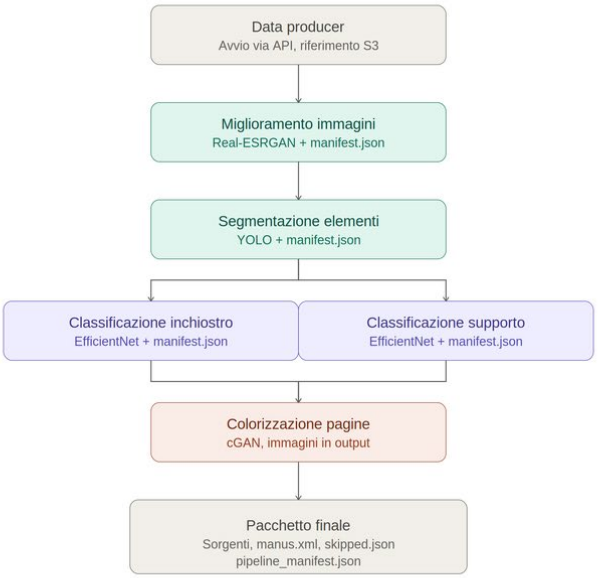


Figura 3 Workflow creazione dei risultati del Data Product

Il diagramma sopra riassume visivamente il flusso: dall'avvio via API, passando per miglioramento (Real-ESRGAN), segmentazione (YOLO) e le due classificazioni parallele (inchiostro e supporto, entrambe su EfficientNet), fino alla colorizzazione (cGAN) e alla composizione del pacchetto finale destinato al Data Product.

5. Esposizione e distribuzione

L'accesso al Data Product avviene tramite API REST e il download manuale di pacchetti strutturati direttamente dal catalog. Entrambe le modalità espongono gli stessi dati: immagini migliorate e colorizzate, varianti di colorizzazione, metadati e schede descrittive. In entrambi i casi, gli asset veri e propri (immagini, archivi, manifest) risiedono sullo storage a oggetti S3: l'API e il pacchetto scaricabile ne costituiscono le due porte d'accesso, non una copia separata del dato.

I destinatari dell'esposizione comprendono sia utenti umani sia sistemi terzi: Data Consumer e utenti business (professionisti, ricercatori, enti autorizzati) che consultano, riutilizzano o integrano il dato nei propri flussi di lavoro; archivisti, curatori e conservatori che validano le ipotesi di colorazione tramite l'applicazione remota; Product Manager e amministratori di piattaforma che supervisionano configurazioni e accessi. L'autenticazione e l'autorizzazione all'accesso sono gestite a livello di piattaforma DPaaS, che regola le richieste di consumo dal catalogo pubblico; il dettaglio delle credenziali e dei meccanismi di autorizzazione applicativa è riportato nel manuale del Data Product, disponibile previa richiesta di consumo dal catalogo.

L'aggiornamento del dataset avviene per lotti, in corrispondenza dell'elaborazione di nuovi pacchetti di microfilm da parte del Data Producer; non è previsto un aggiornamento incrementale continuo, bensì il progressivo arricchimento del corpus a mano a mano che nuovi lotti vengono processati e validati.

Il Data Product è distribuito con licenza CC BY-NC-SA 4.0, che ne consente il riutilizzo per finalità non commerciali con obbligo di attribuzione e condivisione allo stesso modo.

Fruizione tramite dataset

Il dataset esposto dal Data Product è organizzato attorno a tre entità principali, persistite su un database relazionale PostgreSQL e referenziate rispetto ai relativi oggetti conservati su storage S3:

- Project: contenitore logico che raggruppa l'insieme delle operazioni relative a un dato lotto di microfilm;
- Task: singola esecuzione di elaborazione AI, che referenzia i percorsi S3 dei file prodotti (manifest_path, archive_path) e ne registra le principali metriche;
- Manuscript: informazioni storiche e codicologiche del manoscritto (dimensioni, foratura, rigatura, segnature) e riferimenti ai microfilm scansionati.

Ogni Task espone un insieme di metadati e metriche associate: identificativi di configurazione ed esecuzione (task_id, mode, status, con valori ammessi in corso, completato, fallito); metriche temporali (started_at, completed_at, total_elapsed_time, espresso in secondi); metriche volumetriche (total_input, total_processed, total_outputs, total_skipped); flag booleani di processamento AI (enhanced, upscaled); e percorsi di tracciabilità verso gli oggetti S3 (source_folder, output_folder, manifest_path, archive_path). I campi relativi a percorsi di output e archivio possono risultare assenti finché il Task non raggiunge lo stato completato; il file manifest.json, anch'esso conservato su S3, costituisce in ogni caso la fonte autorevole per riconciliare i file effettivamente prodotti rispetto a quelli attesi, incluse le eventuali immagini saltate per anomalie.
Il download del pacchetto strutturato avviene direttamente dal catalogo DPaaS, sotto forma di archivio compresso contenente l'insieme coordinato degli output generati per un dato lotto: immagini originali e migliorate, varianti plausibili di colorizzazione, file manifest.json e, ove disponibili, le schede descrittive .xml estratte da ManusOnline.

Fruizione tramite API
L'accesso programmatico al Data Product avviene tramite un'API REST che consente di consultare, a livello funzionale, i progetti e lo stato delle relative elaborazioni. Le principali funzionalità esposte comprendono il recupero della lista paginata dei progetti disponibili e la consultazione del dettaglio di un singolo progetto, con l'elenco completo dei Task associati e delle relative metriche di elaborazione. Le richieste supportano parametri di paginazione (numero di pagina e dimensione della pagina) per la navigazione efficiente di cataloghi di progetti estesi.

Le risposte dell'API sono restituite in formato JSON e includono, per ciascun progetto, l'identificativo univoco, il nome, la data di creazione e il numero di Task associati; per ciascun Task, lo stato dell'elaborazione, le tempistiche di esecuzione, le metriche volumetriche e i riferimenti ai percorsi S3 degli asset prodotti. L'accesso effettivo alle immagini non avviene tramite il dataset esposto dall'API, bensì direttamente sullo storage S3: l'API restituisce URL firmati (presigned), generati dinamicamente verso i singoli oggetti S3 e con validità temporale limitata, che consentono al client di recuperare il file richiesto senza esporre pubblicamente il bucket sottostante.

Limitazioni

I limiti di utilizzo del Data Product sono principalmente riconducibili ai vincoli tecnici del modello AI impiegato nella fase di colorizzazione. In primo luogo, il DP supporta esclusivamente immagini nei formati JPG, PNG e TIFF; formati diversi da questi non sono accettati in ingresso dalla pipeline.

Il vincolo più significativo riguarda tuttavia la risoluzione di elaborazione: il modello AI accetta in input, e restituisce in output, immagini con risoluzione fissa pari a 512×512 px. Le immagini sorgente, spesso di dimensioni ben superiori, devono pertanto essere preventivamente ridotte (downscaling) a tale risoluzione prima dell'elaborazione, per poi essere riportate alla scala originale al termine del processo. Questo adattamento forzato introduce una perdita di informazione tra la dimensione nativa dell'immagine e quella effettivamente elaborata dal modello, con un conseguente impatto sulla nitidezza e sul dettaglio dell'immagine in output.

La degradazione risulta particolarmente evidente nel caso di immagini in formato TIFF ad alta risoluzione, tipiche delle campagne di digitalizzazione da microfilm: è stato osservato un passaggio da valori di circa 5000 DPI in input a valori effettivi prossimi ai 96 DPI in output, con una perdita sostanziale di informazioni visive. Questo limite va tenuto in considerazione soprattutto nei casi d'uso che richiedono un elevato livello di dettaglio (es. analisi paleografica di dettagli minuti), per i quali l'immagine colorizzata dovrebbe essere considerata come supporto interpretativo piuttosto che come sostituto della digitalizzazione originale ad alta risoluzione.

5.1. Nota sulle condivisioni dei dati

Descrivere le condizioni che regolano la condivisione, la pubblicazione e il riutilizzo dei dati contenuti nel Data Product.

Specificare:

- chi è autorizzato ad accedere ai dati;
- quali dati possono essere condivisi;
- con quali soggetti possono essere condivisi;
- attraverso quali canali;
- con quali livelli di autorizzazione;
- se sono presenti dati personali, riservati, classificati o soggetti a restrizioni;
- le eventuali attività di anonimizzazione o pseudonimizzazione;
- la licenza applicata ai dati e ai metadati;
- gli obblighi di attribuzione;
- i divieti di modifica, redistribuzione o utilizzo commerciale;
- le eventuali condizioni definite dal titolare o dal responsabile del dato.

Indicare inoltre se la condivisione riguarda il dato completo, una sua versione aggregata, una selezione di attributi, i soli metadati o esclusivamente i risultati delle elaborazioni.

Qualora siano previsti diversi livelli di accesso, descrivere chiaramente le differenze tra fruizione pubblica, riservata, autenticata o subordinata a specifica autorizzazione.

6. Possibili evoluzioni future

Durante la realizzazione del Data Product si sono individuate alcune opportunità di sviluppo, finalizzate a migliorare la qualità degli output prodotti. Tra queste, l'ampliamento del dataset di addestramento dei modelli AI e l'evoluzione del modulo di miglioramento delle immagini rappresentano le direttrici principali su cui intervenire per innalzare ulteriormente la resa qualitativa delle immagini colorizzate.

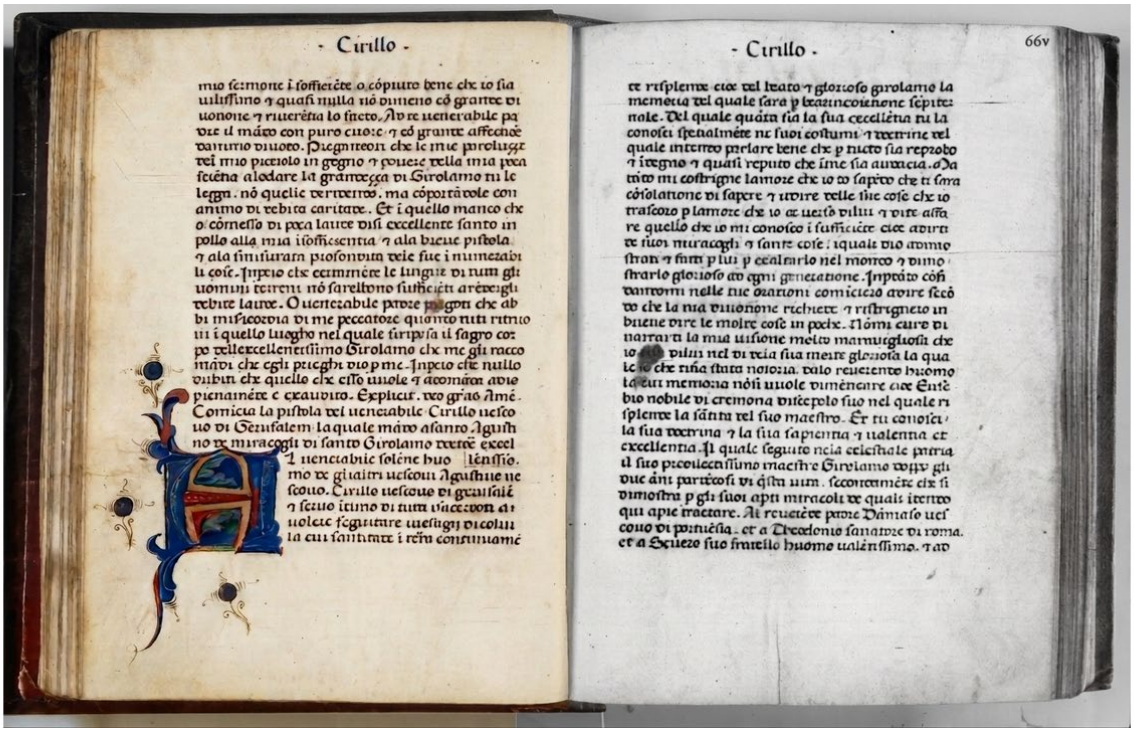


Figura 4 Esempio di colorizzazione AI

Ampliamento del dataset di addestramento

Una prima direzione di miglioramento riguarda l'evoluzione dei modelli AI della pipeline (cGAN di colorizzazione e modelli di classificazione e segmentazione), da conseguire attraverso due leve complementari.

La prima consiste nell'ampliamento del dataset di addestramento, integrando un numero maggiore di coppie immagine-ground truth, con particolare attenzione a esempi significativi per varietà di supporto, tipologia di inchiostro e apparati decorativi. Un dataset più ampio e rappresentativo consentirebbe al modello di generalizzare meglio su manoscritti con caratteristiche meno frequenti nel dataset attuale, riducendo gli errori sistematici di colorizzazione.

La seconda leva riguarda la raccolta strutturata del feedback fornito dai validatori umani (archivisti, curatori e conservatori) in fase di revisione delle varianti di colorizzazione prodotte. Le scelte di validazione, ad esempio quali varianti vengono accettate, corrette o scartate, costituiscono infatti un segnale utile per individuare in modo mirato le tipologie di manoscritto o gli elementi (es. inchiostri, decorazioni) su cui il modello mostra le maggiori criticità, orientando così le successive campagne di raccolta dati e di fine-tuning verso i casi realmente problematici, anziché verso un ampliamento indistinto del dataset.

Evoluzione del modulo di miglioramento immagini (enhancement)

Una seconda linea di sviluppo riguarda l'introduzione di nuovi modelli AI dedicati al miglioramento della qualità dei microfilm sorgente, in affiancamento o in sostituzione di Real-ESRGAN. Oltre alla super-resolution e al denoise già presenti nella pipeline, tali modelli dovrebbero occuparsi di un bilanciamento più accurato di contrasto e saturazione, con l'obiettivo di rendere maggiormente leggibili i microfilm più problematici. Un enhancement più efficace, oltre a migliorare la leggibilità delle immagini in bianco e nero, fornirebbe anche un input di partenza più pulito e uniforme al modello di colorizzazione, con benefici indiretti sulla qualità della resa cromatica finale. Questa evoluzione contribuirebbe inoltre a mitigare, almeno parzialmente, il limite strutturale legato alla risoluzione fissa di elaborazione del modello di colorizzazione (512×512 px), oggi causa della perdita di dettaglio osservata sulle immagini TIFF ad alta risoluzione.