# L'Aquila – Raccontiamo i documenti

## Documento di specifiche tecniche di elaborazione

Versione del documento: 1.0

Data: 28/08/2026

# Sommario

1. Obiettivi del Data Product ..... 4
2. Origine dei dati ed eventuali elaborazioni ..... 5
3. Tecnologie utilizzate all'interno del Data Product Lab ..... 10
3.1. Servizi abilitanti DPaaS ..... 11
3.2. Eventuali tecnologie esterne a DPaaS ..... 14
3.2.1. Modelli in locale ..... 14
4. Pipeline di elaborazione e trasformazione ..... 16
4.1. Pipeline di trascrizione, traduzione e semplificazione dei manoscritti storici 16
4.1.1. Acquisizione e segmentazione della pagina manoscritta ..... 16
4.1.2. Trascrizione automatica delle singole porzioni testuali ..... 16
4.1.3. Ricostruzione della trascrizione completa ..... 17
4.1.4. Revisione della trascrizione ..... 17
4.1.5. Generazione di spiegazioni semplificate del contenuto ..... 17
4.1.6. Produzione delle sintesi audio ..... 17
4.1.7. Traduzione del manoscritto in italiano moderno ..... 18
4.1.8. Generazione della traduzione inglese ..... 18
4.1.9. Elaborazione a livello di collezione documentale ..... 18
4.1.10. Pubblicazione degli artefatti ..... 19
5. Esposizione e distribuzione ..... 20
5.1. Canali di accesso ..... 20
5.2. Utenti e sistemi destinatari ..... 20
5.3. Autenticazione e autorizzazione ..... 21
5.4. Formati disponibili ..... 22
5.5. Modalità di aggiornamento ..... 22
5.6. Limiti di utilizzo ..... 23
5.7. Accesso alla documentazione ..... 23
5.8. Condizioni di licenza e riutilizzo ..... 23
5.9. Fruizione tramite API ..... 24

5.9.1. Tipologie di informazioni interrogabili ..... 24
5.9.2. Principali operazioni disponibili ..... 24
5.9.3. Ricerca, filtro, ordinamento e paginazione ..... 25
5.9.4. Recupero del dettaglio di una risorsa ..... 25
5.9.5. Servizi di aggregazione e analisi ..... 25
5.9.6. Formati delle richieste e delle risposte ..... 26
5.9.7. Modalità di autenticazione ..... 26
5.9.8. Limiti di chiamata ..... 26
5.9.9. Gestione degli errori ..... 27
6. Limitazioni ..... 27
6.1. Nota sulle condivisioni dei dati ..... 28
6.2. Nota sulle condivisioni dei dati ..... 29
7. Possibili evoluzioni future ..... 30

# 1.Obiettivi del Data Product

Il Data Product L'Aquila – Raccontiamo i documenti è finalizzato a rendere accessibili, in modo strutturato e automatizzato, fotografie e documenti manoscritti appartenenti all'Archivio di Stato dell'Aquila. Il progetto si concentra in particolare sull'elaborazione di manoscritti storici digitalizzati, con l'obiettivo di trasformare contenuti complessi e di difficile consultazione in artefatti fruibili da un'utenza ampia e diversificata, con particolare attenzione alle persone con disabilità visive, cognitive o con basso livello di alfabetizzazione.

All'interno della soluzione vengono raccolte le immagini digitali delle pagine dei manoscritti originali e i contenuti derivati generati automaticamente tramite tecnologie di Intelligenza Artificiale, con l'obiettivo di favorire l'accessibilità e la valorizzazione del patrimonio documentale. In particolare, a partire dalle immagini delle pagine, vengono prodotte la trascrizione testuale del contenuto manoscritto e la relativa traduzione nelle lingue italiana e inglese, rendendo i testi maggiormente comprensibili e fruibili da un pubblico più ampio. Inoltre, per ciascun manoscritto viene generato un riassunto complessivo dell'opera, trasformato anche in formato audio, così da consentire l'accesso ai contenuti attraverso modalità alternative di consultazione. Il Data Product conserva pertanto sia le immagini originali delle pagine sia tutti gli artefatti generati dall'AI, inclusi trascrizioni, traduzioni e sintesi audio nelle lingue supportate (ITA ed ENG), contribuendo a una fruizione più inclusiva, accessibile e immediata delle opere manoscritte.
Gli artefatti prodotti sono messi a disposizione attraverso un portale dedicato, che consente ai visitatori del museo, agli studiosi e agli altri utenti autorizzati di consultare sia le riproduzioni digitali dei manoscritti sia i contenuti arricchiti generati dall'Intelligenza Artificiale. Il portale prevede inoltre specifiche funzionalità di revisione e modifica degli artefatti, riservate a utenti abilitati, al fine di consentire attività di verifica, correzione e validazione delle trascrizioni, delle traduzioni e dei riassunti generati automaticamente. Tale approccio permette di integrare l'efficienza dei processi AI con il controllo esperto degli operatori, garantendo un miglioramento continuo della qualità dei contenuti e assicurando elevati livelli di accuratezza, affidabilità e accessibilità nella consultazione delle opere manoscritte.

# 2. Origine dei dati ed eventuali elaborazioni

### 1. Sistema, archivio, database, API o dataset di origine

La fonte originaria del Data Product è costituita dai documenti e manoscritti storici digitalizzati, dalle immagini delle relative pagine e dai metadati descrittivi conferiti dall'Archivio di Stato dell'Aquila nell'ambito del progetto "L'Aquila - Raccontiamo i documenti".

Il Data Product è identificato come DP00000084 e appartiene al dominio archivistico.

La fonte è organizzata secondo una struttura gerarchica: manoscritti, libri, documenti, fondi o collezioni costituiscono gli oggetti aggregatori, mentre le immagini digitali delle singole pagine ne rappresentano i contenuti elementari. Le relazioni tra i livelli consentono di ricostruire l'appartenenza di ciascuna pagina al proprio documento e il corretto ordine delle pagine.

Gli artefatti digitali, i metadati e il dataset strutturato che li descrive vengono depositati in uno spazio di archiviazione dedicato, punto di ingresso tecnico dei dati nel Data Product. Le strutture temporanee, il database applicativo e i processi interni di elaborazione sono componenti di supporto e non costituiscono fonti dati ulteriori: la fonte originaria resta quella conferita dall'Archivio di Stato dell'Aquila.

### 2. Soggetto responsabile o titolare della fonte

L'Archivio di Stato dell'Aquila è responsabile del conferimento dei documenti digitalizzati, delle immagini delle pagine e dei relativi metadati, nonché della disponibilità, correttezza e completezza dei materiali trasmessi e della corretta rappresentazione della struttura archivistica (appartenenza delle pagine, sequenza, associazione a fondi o collezioni).

La titolarità del Data Product e dei diritti su manoscritti, documenti, fondi, immagini e riproduzioni fa capo all'Archivio di Stato dell'Aquila (per la licenza si veda il punto 7).

Gli operatori autorizzati dall'ente possono correggere e validare trascrizioni e altri artefatti destinati alla pubblicazione, contribuendo al controllo scientifico e documentale dei risultati.

## 3. Tipologia e formato dei dati

I dati sorgente sono costituiti principalmente da: documenti e manoscritti digitalizzati; immagini delle singole pagine; libri, documenti, fondi o collezioni come raggruppamenti archivistici; metadati descrittivi e tecnici; informazioni sulla sequenza delle pagine; relazioni di appartenenza tra pagine, manoscritti, libri, documenti e collezioni; descrizioni, parole chiave e informazioni contestuali.

I principali formati della fonte sono:

- **PNG, JPG/JPEG**, per le immagini digitali delle pagine;
- **TIFF/TIF**, per le riproduzioni digitali ad alta qualità;
- **XLSX**, per il dataset strutturato di riferimenti e metadati;
- **JSON**, per la rappresentazione strutturata nelle fasi successive di gestione.

Per ciascun oggetto o pagina possono essere disponibili: identificativo, nome file, estensione/formato, percorso di archiviazione, tipologia, titolo, descrizione, autore o fondo di provenienza, data, lingua, tag, classificazione, riferimento al livello superiore di appartenenza e ordine della pagina. Queste informazioni permettono di identificare ogni oggetto, collegarlo al file e ricostruire la struttura del documento, distinguendo il livello pagina da quello di manoscritto/libro/collezione.

## 4. Modalità di acquisizione

L'acquisizione avviene mediante caricamento del dataset strutturato, delle immagini e dei metadati nello spazio di archiviazione dedicato. Le immagini vengono associate ai record tramite identificativi condivisi (codice oggetto, nome file, formato, percorso, riferimento al documento di appartenenza).

Dopo il conferimento, la soluzione acquisisce dati e riferimenti ai file, identifica gli oggetti e ne registra le informazioni, transitando eventualmente per strutture temporanee di supporto. Le principali attività previste sono:

- selezione dei record validi ed esclusione di quelli incompleti;
- normalizzazione e conversione in formato compatibile;
- verifica dell'associazione tra record descrittivo e file digitale;
- classificazione per tipologia, lingua e stato;

- organizzazione delle relazioni gerarchiche e ricostruzione dell'ordine delle pagine;
- collegamento a tag e classificazioni;
- integrazione dei metadati tecnici;
- registrazione finale nel Data Product.

Le immagini possono successivamente essere elaborate per produrre trascrizioni, traduzioni, spiegazioni semplificate, riassunti e contenuti audio, anche a livello di intero manoscritto o raggruppamento. Questi sono artefatti derivati, non nuove fonti originarie.

### 5. Frequenza di aggiornamento

Il Data Product ha natura non real-time. La fonte si aggiorna quando l'Archivio di Stato dell'Aquila conferisce nuovi contenuti o modifica/sostituisce quelli esistenti: nuovi manoscritti o immagini, completamento di documenti parziali, sostituzione di immagini, integrazione dei metadati, correzione dell'ordine o delle relazioni tra pagine, revisione di trascrizioni con rigenerazione dei contenuti collegati.

Non è definita una frequenza periodica: l'aggiornamento è **non periodico**, legato al conferimento o alla modifica dei contenuti da parte dell'ente. Aggiornamenti di dati, immagini, metadati o artefatti rientrano nella versione corrente; un nuovo versionamento è necessario solo se le modifiche incidono sulla struttura dei dati o sulle caratteristiche sostanziali della soluzione.

### 6. Volume indicativo dei dati

Il volume della fonte originaria non è attualmente quantificato. Va determinato considerando: numero di manoscritti, libri, documenti e collezioni; numero complessivo di pagine e immagini; dimensione dei file; numero di record di metadati e relazioni; oggetti o cartelle ancora da completare; crescita in caso di nuovi conferimenti.

Il volume della fonte va distinto da quello del Data Product dopo le elaborazioni: da ogni pagina possono derivare trascrizioni, traduzioni, spiegazioni, riassunti, audio e contenuti aggregati a livello di manoscritto o collezione, per cui il volume complessivo del Data Product può risultare significativamente superiore a quello della sola fonte.

### 7. Eventuali vincoli di accesso, utilizzo o licenza

L'accesso tecnico allo spazio di archiviazione è riservato a utenti e componenti applicativi autorizzati tramite credenziali e profili dedicati; non c'è accesso diretto e indiscriminato alle risorse, che vengono fruite attraverso i meccanismi previsti dal Data Product. La parte pubblica del portale rende disponibili i contenuti dichiarati pubblicabili; revisione, modifica e validazione restano riservate agli operatori autorizzati.

La pubblicazione e il riuso dei dati devono essere disciplinati dalla licenza del Data Product, ancora da selezionare e formalizzare. Le condizioni di utilizzo dovranno considerare: titolarità di manoscritti, documenti e fondi; diritti sulle immagini e su eventuali diritti di terzi; obblighi di attribuzione; possibilità di copia, redistribuzione o modifica; eventuali restrizioni commerciali; condizioni per trascrizioni, traduzioni e altri contenuti derivati; distinzione tra pubblicazione di immagini, metadati o contenuti derivati.

Le autorizzazioni tecniche (credenziali e profili) e le condizioni giuridiche (licenza e atti di conferimento) restano distinte: le prime disciplinano l'accesso ai sistemi, le seconde l'elaborazione, la pubblicazione e il riuso dei contenuti.

## 8. Livello di qualità e completezza riscontrato

La fonte presenta una struttura organizzata per manoscritti, libri, documenti, collezioni e pagine, con identificativi, metadati e relazioni che permettono di ricostruirne il contesto archivistico e l'ordine.

I controlli di qualità devono riguardare:

- validità dell'identificativo e presenza effettiva del file;
- corrispondenza tra record descrittivo e file (nome, formato, percorso);
- leggibilità e integrità dell'immagine;
- presenza dei metadati minimi;
- corretta appartenenza gerarchica e ordinamento delle pagine;
- coerenza delle relazioni tra oggetti aggregatori e pagine;
- assenza di duplicati e individuazione di immagini mancanti o da sostituire;
- coerenza tra lingua, classificazione e tag.

La qualità delle immagini è particolarmente rilevante, poiché la leggibilità dei manoscritti può essere condizionata da calligrafie irregolari, degrado del supporto, variazioni cromatiche o impaginazioni complesse.

Non sono definiti KPI quantitativi, soglie di accettazione o una percentuale minima di completezza. La qualità della fonte può essere descritta come **adeguata** all'acquisizione dei documenti digitalizzati, a condizione che siano completati i controlli di coerenza tra immagini, metadati, struttura e relazioni documentali.

# 3. Tecnologie utilizzate all'interno del Data Product Lab

Il Data Product si avvale di alcuni componenti applicativi utilizzati per l'orchestrazione dei processi di elaborazione, la gestione dei flussi di lavoro, l'esecuzione delle pipeline di Intelligenza Artificiale e il monitoraggio dello stato di avanzamento delle elaborazioni. Tali componenti operano in integrazione con le funzionalità offerte dalla piattaforma e consentono l'automazione dell'intero ciclo di generazione degli artefatti di accessibilità.

In particolare, il processo è orchestrato tramite Apache NiFi, sul quale sono configurati job schedulati che effettuano periodicamente il monitoraggio delle cartelle S3 associate ai singoli Enti conferitori. I job verificano la presenza di nuovi contenuti e controllano, attraverso appositi meccanismi di tracciamento, se i file siano già stati elaborati dal sistema.

Nel caso in cui un contenuto risulti non ancora processato, NiFi attiva una sequenza di chiamate verso due API REST esposte sul Livello 1 dell'architettura applicativa. Entrambe le API sono state sviluppate in linguaggio Python e costituiscono componenti esterne rispetto alla piattaforma DPaaS.

La prima API è dedicata alla traduzione automatica dei metadati in lingua inglese, necessaria per garantire la disponibilità delle informazioni descrittive nelle lingue supportate dal Data Product. La seconda API implementa invece la pipeline di generazione degli artefatti, responsabile della creazione dei contenuti derivati quali descrizioni, audiodescrizioni, trascrizioni, sottotitoli, riassunti e ulteriori risorse di accessibilità.

Per ciascun Ente sono stati configurati specifici task di elaborazione, definiti sulla base dei requisiti funzionali e delle esigenze espresse durante la fase di raccolta dei fabbisogni. Ogni task identifica una particolare tipologia di elaborazione da eseguire sui contenuti conferiti dall'Ente e viene gestito in maniera indipendente all'interno della pipeline applicativa.

Lo stato di avanzamento delle elaborazioni viene registrato all'interno di un database PostgreSQL, anch'esso utilizzato come componente di supporto esterno alla piattaforma DPaaS. Per ogni elaborazione sono memorizzati identificativo del task, stato di esecuzione, informazioni di monitoraggio ed eventuali errori rilevati durante il processamento, consentendo il tracciamento completo del ciclo di vita delle lavorazioni.

A livello operativo, Apache NiFi implementa inoltre specifici meccanismi di retry automatico, che consentono di rieseguire le chiamate verso le API in caso di errori

temporanei o malfunzionamenti transitori, aumentando la resilienza complessiva del sistema e riducendo il rischio di perdita delle elaborazioni.

Al completamento di tutti i task previsti per uno specifico Ente, gli artefatti generati e i relativi metadati vengono considerati disponibili per la pubblicazione e possono essere successivamente fruiti dai portali applicativi che accedono al Data Product.

## 3.1. Servizi abilitanti DPaaS

Di seguito vengono riportati tutti i servizi abilitanti DPaas che sono stati associati dall'ente al Data Product e sono stati utilizzati al fine di generare gli artefatti:

- CPAI - OCR;
- CPAI - Speech2Text;
- CPAI - Text2Speech;
- CPAI - Ner;
- CPAI - Tagging;
- CPAI - Image Enhance;
- CPAI - Abstract;
- CPAI - Estrazione Entità;
- CPAI - Epigrafi Latine;
- CPAI - Abstract immagini;
- CPAI - Tagging Entità;
- CPAC - WhisperTTS;
- CPAC - Whisper;
- CPAC - Tagging;
- CPAC - Stable Diffusion;
- CPAC - Qwen2;
- CPAC - P2PaLa;
- CPAC - NERs;
- CPAC - Llama 31;
- CPAC - Kosmos2;
- CPAC - Keras;
- CPAC - HCR_Inference;
- CPAC - Gliner;

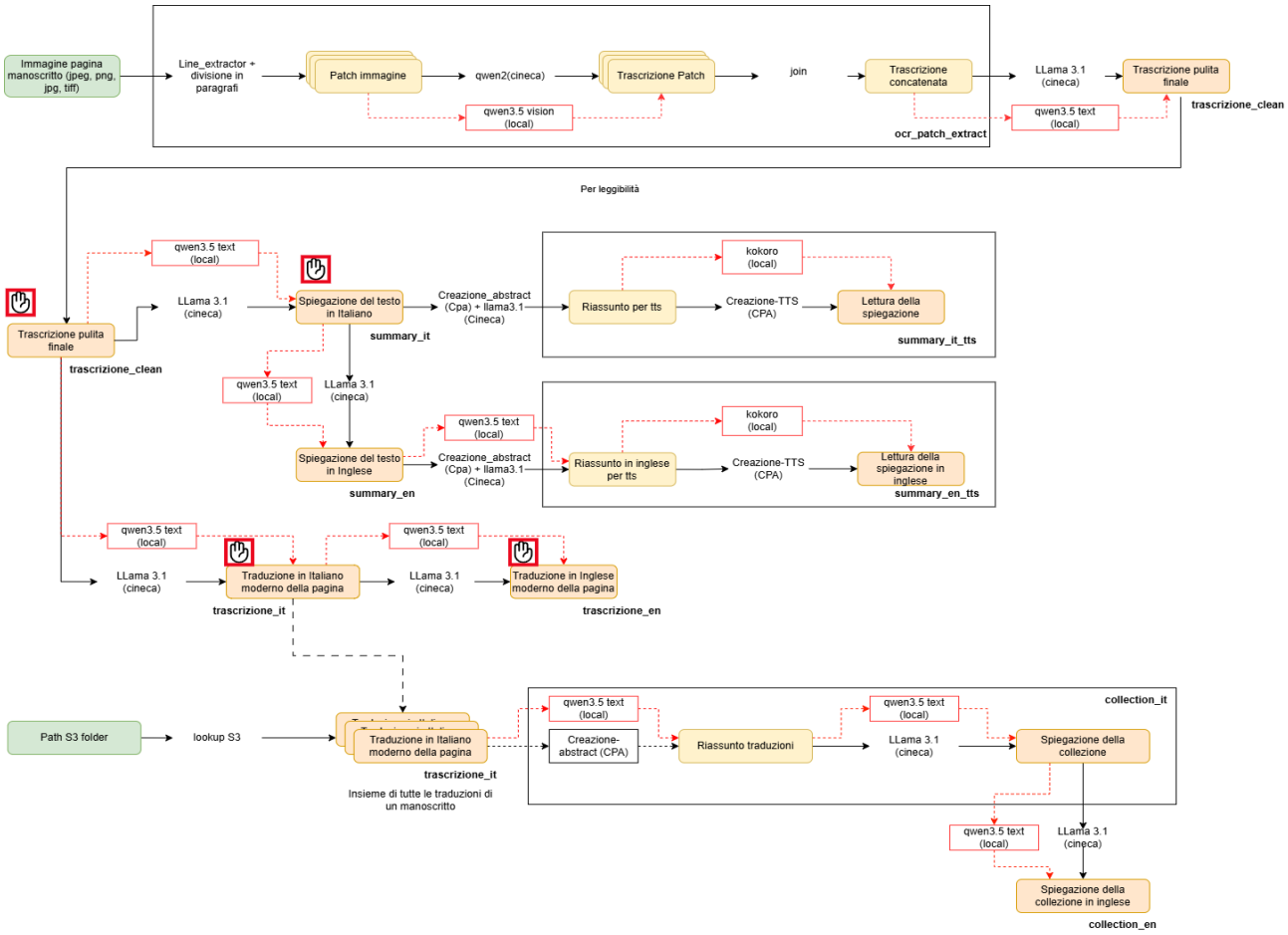
- HCR_DataPreparation;
- VM;
- JupyterLab;
- OpenSeaDragon;
- NiFi;
- Postgres;
- Elastic Search.

Si evidenzia che **non tutti i servizi, modelli e componenti riportati nell'elenco delle tecnologie disponibili sono stati effettivamente utilizzati nell'ambito della presente fase progettuale**. La loro inclusione è stata richiesta all'Ente al fine di consentire un'attività di confronto tra diverse soluzioni tecnologiche, sia in termini di prestazioni sia di qualità dei risultati ottenuti, nonché per valutare la fattibilità di specifici requisiti funzionali e non funzionali.

L'associazione delle tecnologie al Data Product ha pertanto avuto anche una finalità esplorativa e di assessment, volta a identificare gli strumenti maggiormente idonei per l'evoluzione della soluzione. Alcuni dei servizi elencati potrebbero quindi non essere stati utilizzati nelle pipeline attualmente implementate, ma rappresentano opzioni tecnologiche valutate durante la fase di analisi o candidate per future evoluzioni del sistema.

La presenza di tali componenti nella documentazione non implica pertanto il loro impiego operativo nella versione corrente del Data Product, ma ne documenta la disponibilità e la potenziale adozione in successive fasi di sviluppo, ottimizzazione o ampliamento delle funzionalità previste.

Di seguito viene riportato l'albero elaborativo che descrive, per ogni tipologia di input, il flusso che è stato costruito e i servizi utilizzati per le varie trasformazioni.




Come è possibile vedere dall'immagine, ciò che viene generato è:

INPUT: *immagine*

OUTPUT

- Trascrizione della pagina
- Traduzione in Italiano e in inglese
- Riassunto della pagina in Italiano e in inglese
- Audio riassunto della pagina in Italiano e in inglese
- Riassunto dell'intero manoscritto in Italiano e in inglese

La descrizione approfondita delle pipeline viene riportata di seguito, nel paragrafo 4.

# 3.2. Eventuali tecnologie esterne a DPaaS

## 3.2.1. Modelli in locale

Durante la fase di realizzazione del Data Product sono stati valutati e testati i servizi di Intelligenza Artificiale messi a disposizione dall'ecosistema DPaaS, verificandone l'applicabilità rispetto agli specifici casi d'uso, agli obiettivi progettuali e ai requisiti di accessibilità definiti dagli Enti coinvolti. Per alcune funzionalità, al fine di supportare le attività di sviluppo e collaudo e di assicurare la necessaria continuità nella sperimentazione, si è ritenuto opportuno affiancare ai servizi disponibili modelli eseguiti su infrastrutture esterne gestite dal fornitore. Tale configurazione ha mantenuto la piena integrazione con i workflow di orchestrazione e con i meccanismi di pubblicazione della piattaforma DPaaS.
Per la maggior parte delle funzionalità sono stati adottati modelli funzionalmente equivalenti ai servizi presenti sulla piattaforma DPaaS. L'impiego di istanze locali durante la fase di sviluppo ha consentito di configurare, verificare e ottimizzare i modelli in relazione agli specifici contenuti e requisiti del Data Product. Per la componente di elaborazione multimodale è stato utilizzato **Qwen 3.5 Vision**, mentre la piattaforma DPaaS rende attualmente disponibile la versione **Qwen2**.
La scelta di utilizzare Qwen 3.5 Vision è stata determinata dalla necessità di elaborare contenuti multimodali, in particolare immagini di opere, fotografie, pagine manoscritte e documenti digitalizzati. Rispetto a Qwen2, la versione 3.5 introduce capacità più evolute di interpretazione visiva, una migliore integrazione tra informazioni testuali e contenuti grafici, una maggiore accuratezza nell'estrazione del contenuto documentale e una qualità superiore nella generazione di descrizioni contestualizzate. Tali caratteristiche si sono dimostrate particolarmente rilevanti per la realizzazione degli artefatti di accessibilità previsti dal progetto, quali descrizioni di immagini, audiodescrizioni, trascrizioni di manoscritti e contenuti derivati destinati alla consultazione museale.
Considerata la rilevanza di tale componente per il corretto funzionamento delle pipeline implementate, è stato richiesto ai responsabili della piattaforma DPaaS di rendere disponibile e ospitare una versione del modello **Qwen 3.5 Vision** all'interno dell'infrastruttura ufficiale. L'adozione del medesimo modello consentirebbe infatti di mantenere allineati i risultati prodotti durante la fase di sviluppo con quelli generati successivamente in esercizio.
Si evidenzia pertanto che, a valle della messa in produzione del Data Product, tutti i nuovi contenuti caricati o modificati verranno elaborati mediante i modelli disponibili sulla piattaforma DPaaS. Qualora il modello **Qwen 3.5 Vision** non venga reso disponibile all'interno dell'infrastruttura, le elaborazioni che richiedono capacità di analisi visuale verranno eseguite utilizzando **Qwen2**. In tale scenario, pur mantenendo invariata la

funzionalità complessiva del sistema, gli artefatti generati potrebbero presentare un livello di dettaglio, accuratezza e qualità descrittiva differente rispetto a quello ottenuto durante la fase di realizzazione e collaudo del Data Product, nella quale è stato utilizzato Qwen 3.5 Vision come modello di riferimento.

# 4. Pipeline di elaborazione e trasformazione

## 4.1. Pipeline di trascrizione, traduzione e semplificazione dei manoscritti storici

La pipeline è progettata per trasformare immagini digitali di manoscritti storici in contenuti testuali accessibili, comprensibili e fruibili da un pubblico ampio. Il processo integra tecniche avanzate di riconoscimento ottico della scrittura, modelli multimodali e Large Language Models per produrre trascrizioni, traduzioni in linguaggio moderno, sintesi, audiodescrizioni e spiegazioni contestuali delle collezioni documentarie. Tutti gli artefatti generati vengono archiviati all'interno del Data Product e resi disponibili attraverso il portale di consultazione. Gli elementi contrassegnati con l'icona della mano possono essere revisionati e modificati dagli operatori autorizzati per garantirne la correttezza e la qualità scientifica.

### 4.1.1. Acquisizione e segmentazione della pagina manoscritta

La pipeline prende in ingresso immagini di pagine manoscritte nei formati **PNG, JPG e TIFF**.

Poiché i manoscritti possono presentare impaginazioni complesse, differenti colonne e variazioni nella distribuzione del testo, la prima fase prevede l'utilizzo di un componente di **Line Extractor**, incaricato di identificare e segmentare automaticamente righe e paragrafi presenti nella pagina.

L'immagine viene quindi suddivisa in una serie di porzioni più piccole (*patch*), ciascuna contenente una specifica area testuale che potrà essere elaborata in modo indipendente.

### 4.1.2. Trascrizione automatica delle singole porzioni testuali

Le patch generate vengono elaborate mediante i modelli **Qwen 2** eseguito presso **CINECA** e **Qwen 3.5 Vision** eseguito localmente.

L'obiettivo è effettuare il riconoscimento del testo manoscritto e produrre una prima trascrizione di ciascun frammento documentale (*Trascrizione Patch*). Questa strategia consente di aumentare la qualità dell'estrazione testuale anche in presenza di manoscritti caratterizzati da calligrafie irregolari, degrado del supporto o strutture particolarmente complesse.

### 4.1.3. Ricostruzione della trascrizione completa

Le trascrizioni ottenute dalle singole patch vengono successivamente concatenate per ricostruire il contenuto integrale della pagina.

Il risultato di questa fase è una trascrizione unica (*ocr_patch_extract*) che rappresenta il testo completo estratto dal documento storico.

La trascrizione concatenata viene quindi sottoposta a una fase di revisione automatica mediante **Llama 3.1** eseguito presso CINECA e **Qwen 3.5 Text** eseguito localmente, con l'obiettivo di correggere errori di riconoscimento, uniformare la punteggiatura e migliorare la leggibilità del testo.

L'artefatto risultante prende il nome di **Trascrizione pulita finale** (*trascrizione_clean*).

### 4.1.4. Revisione della trascrizione

La trascrizione pulita rappresenta il principale risultato del processo OCR ed è resa modificabile attraverso il portale.

Gli operatori autorizzati possono correggere eventuali errori residui, validare l'accuratezza della trascrizione e integrare manualmente parti del testo che potrebbero risultare difficili da riconoscere automaticamente.

Questa fase garantisce l'elevata qualità scientifica del contenuto e costituisce il punto di partenza per tutte le successive elaborazioni.

### 4.1.5. Generazione di spiegazioni semplificate del contenuto

A partire dalla trascrizione validata, la pipeline genera una **spiegazione del testo in italiano** (*summary_it*) utilizzando il modello **Llama 3.1**, supportato da **Qwen 3.5 Text**.

L'obiettivo è produrre una descrizione più chiara e facilmente comprensibile rispetto al testo originale, mantenendo il significato storico e culturale del documento ma adattandolo a un pubblico non specialistico.

Questo artefatto è modificabile dal portale e può essere rivisto dagli operatori per migliorarne chiarezza, accuratezza e livello divulgativo.

Successivamente viene prodotta anche la corrispondente **spiegazione in lingua inglese** (*summary_en*), destinata agli utenti internazionali.

### 4.1.6. Produzione delle sintesi audio

Le spiegazioni testuali vengono ulteriormente elaborate per creare versioni ottimizzate per l'ascolto.

Attraverso il servizio **Creazione Abstract (CPA)** basato su **Llama 3.1**, viene prodotto un riassunto sintetico specificamente progettato per la successiva conversione vocale.

Tali contenuti vengono convertiti in audio mediante il servizio **Text-to-Speech CPAI**, che utilizza il modello **Kokoro** eseguito localmente.

Vengono così generati:

- **summary_it_tts**, ovvero la lettura audio della spiegazione italiana;
- **summary_en_tts**, ovvero la lettura audio della spiegazione inglese.

Questi artefatti consentono l'accesso ai contenuti anche tramite modalità di ascolto, ampliando le possibilità di fruizione del patrimonio documentale.

### 4.1.7. Traduzione del manoscritto in italiano moderno

Parallelamente al processo di sintesi, la trascrizione pulita viene elaborata mediante **Llama 3.1** e **Qwen 3.5 Text** per produrre una **traduzione in italiano moderno** (*trascrizione_it*).

L'obiettivo di questa fase è rendere comprensibili testi redatti in forme linguistiche arcaiche, dialettali o caratterizzate da una sintassi non più in uso, mantenendo il contenuto informativo originale.

Anche questo artefatto è modificabile direttamente dal portale e può essere validato dagli esperti prima della pubblicazione.

### 4.1.8. Generazione della traduzione inglese

A partire dalla traduzione in italiano moderno viene prodotta una **versione in inglese** (*trascrizione_en*).

Il risultato consente la consultazione dei contenuti da parte di utenti internazionali, preservando il significato storico e documentario del manoscritto.

### 4.1.9. Elaborazione a livello di collezione documentale

La pipeline non si limita alla singola pagina ma consente anche una visione aggregata dell'intera collezione.

Partendo da tutte le traduzioni moderne associate alle pagine di uno stesso manoscritto o fondo archivistico, il sistema recupera automaticamente i documenti tramite il repository S3 e costruisce un corpus documentale unico.

Attraverso il servizio **Creazione Abstract (CPA)** viene generato un **riassunto complessivo delle traduzioni**, che sintetizza i principali temi e contenuti presenti nell'intera raccolta.

Successivamente **Llama 3.1** genera una **spiegazione della collezione in italiano** (*collection_it*), descrivendo in maniera organica:

- i temi principali trattati;
- il contesto storico-documentario;
- gli argomenti ricorrenti;
- il valore culturale dell'insieme dei documenti.

La spiegazione italiana viene infine tradotta in inglese, generando l'artefatto **collection_en**.

### 4.1.10. Pubblicazione degli artefatti

Al termine dell'elaborazione, il Data Product rende disponibili i seguenti artefatti:

- Trascrizione OCR pulita della pagina (*trascrizione_clean*);
- Spiegazione semplificata in italiano (*summary_it*);
- Spiegazione semplificata in inglese (*summary_en*);
- Lettura audio della spiegazione italiana (*summary_it_tts*);
- Lettura audio della spiegazione inglese (*summary_en_tts*);
- Traduzione in italiano moderno (*trascrizione_it*);
- Traduzione in inglese (*trascrizione_en*);
- Spiegazione dell'intera collezione in italiano (*collection_it*);
- Spiegazione dell'intera collezione in inglese (*collection_en*).

Attraverso questo processo il manoscritto storico viene trasformato da semplice immagine digitale a un insieme strutturato di contenuti testuali, multilingue e accessibili, favorendo la divulgazione e la valorizzazione del patrimonio culturale verso pubblici eterogenei, studiosi, ricercatori e visitatori non specialisti.

# 5. Esposizione e distribuzione

Il Data Product "L'Aquila – Raccontiamo i documenti" viene pubblicato nel catalogo della piattaforma DPaaS e reso fruibile attraverso un portale web dedicato e mediante API REST in sola lettura.

Il Data Product consente la consultazione strutturata delle immagini delle pagine manoscritte appartenenti all'Archivio di Stato dell'Aquila, insieme ai metadati disponibili e agli artefatti di accessibilità generati dalle pipeline di elaborazione: trascrizioni, versioni in italiano moderno, traduzioni in inglese, spiegazioni sintetiche, contenuti audio e descrizioni complessive dei manoscritti o delle collezioni.

L'esposizione è articolata su due livelli funzionali:

- fruizione tramite portale web, destinata agli utenti finali e agli operatori autorizzati;
- fruizione machine-to-machine tramite API, destinata al portale, al Backend for Frontend e agli altri sistemi consumer autorizzati.

Il portale rappresenta il principale canale di consultazione per gli utenti finali. Consente di esplorare il patrimonio documentale, visualizzare le immagini delle pagine, consultare i metadati e accedere alle trascrizioni, alle traduzioni, alle spiegazioni e alle versioni audio prodotte dal Data Product. Per i profili autorizzati mette inoltre a disposizione le funzioni di revisione e validazione previste.

La read-only API costituisce il canale di interoperabilità per il portale e per gli altri sistemi applicativi. Espone gli oggetti documentali e i relativi metadati, le aggregazioni, le relazioni tra manoscritti e pagine e i riferimenti agli artefatti accessibili generati. Le operazioni disponibili sono esclusivamente di consultazione e non consentono ai consumer di modificare direttamente i dati esposti.

## 5.1. Canali di accesso

I canali di accesso previsti sono:

- portale web pubblico dedicato al Data Product L'Aquila;
- API REST read-only accessibili tramite protocollo HTTPS.

La parte pubblica del portale permette la consultazione dei contenuti senza autenticazione. Le funzionalità riservate, in particolare quelle destinate alla revisione, modifica e validazione degli artefatti, richiedono autenticazione e un profilo autorizzativo adeguato.

## 5.2.Utenti e sistemi destinatari

Il Data Product è destinato principalmente a:

- cittadini, visitatori e utenti interessati al patrimonio storico e documentale dell'Aquila;
- persone con disabilità visive o cognitive;
- utenti che beneficiano di contenuti semplificati o di modalità alternative di consultazione;
- utenti internazionali, grazie alla disponibilità di contenuti in lingua inglese;
- studiosi, ricercatori e utenti non specialisti interessati alla consultazione dei manoscritti;
- operatori dell'Archivio di Stato dell'Aquila e operatori incaricati della revisione e validazione degli artefatti;
- portale L'Aquila e relativo Backend for Frontend;
- sistemi applicativi autorizzati che necessitano di interrogare i contenuti tramite API;
- enti culturali e sviluppatori interessati al riutilizzo dei dati e dei metadati, nel rispetto delle condizioni di licenza applicabili.

## 5.3. Autenticazione e autorizzazione

La consultazione dei contenuti pubblici attraverso il portale non richiede autenticazione.

Le funzionalità riservate agli operatori, quali revisione, modifica e validazione degli artefatti generati, sono accessibili tramite autenticazione SPID/CIE e sulla base dei ruoli applicativi assegnati.

In particolare, possono essere sottoposti a revisione da parte degli operatori autorizzati:

- la trascrizione pulita della pagina manoscritta;
- la traduzione in italiano moderno;
- la spiegazione sintetica in italiano;
- gli altri artefatti esplicitamente resi modificabili dal portale.

La revisione di un contenuto può determinare la rigenerazione degli artefatti derivati, quali traduzioni, spiegazioni in lingua inglese e tracce audio, per mantenerli coerenti con il contenuto validato.

Per l'accesso machine-to-machine ai servizi del Data Product, le API sono protette mediante una chiave identificativa dell'utente o dell'applicazione e il codice identificativo del Data Product.

Le credenziali tecniche necessarie per accedere alle componenti di persistenza e allo storage sono utilizzate esclusivamente dai componenti backend e non sono esposte ai consumer. Le modalità definitive di gestione delle chiavi, la loro durata, la rotazione e le

eventuali restrizioni di rete sono definite nell'ambiente di esercizio e configurate sul livello infrastrutturale di esposizione.

## 5.4.Formati disponibili

I dati e i metadati restituiti dalle API sono disponibili in formato JSON.

Le risorse gestite dal Data Product comprendono:

- immagini di pagine manoscritte nei formati PNG, JPG e TIFF;
- trascrizioni testuali delle pagine;
- traduzioni in italiano moderno e in lingua inglese;
- spiegazioni sintetiche in italiano e in inglese;
- letture audio delle spiegazioni;
- spiegazioni complessive dei manoscritti o delle collezioni in italiano e in inglese;
- metadati tecnici e descrittivi strutturati;
- informazioni relative a documenti, pagine, aggregazioni, classificazioni e relazioni.

Le risorse binarie, quali immagini e tracce audio, sono conservate in uno storage privato. I client non accedono direttamente allo storage: l'applicazione recupera le risorse e le restituisce al consumer in streaming attraverso operazioni dedicate.

## 5.5.Modalità di aggiornamento

Il Data Product ha natura non real-time. L'aggiornamento avviene attraverso pipeline eseguite in modalità pianificata oppure attivate a seguito della presa in carico di nuovi contenuti o della revisione di artefatti esistenti.

Lo stato di ogni elaborazione viene registrato. Per ciascuna esecuzione vengono tracciati:

- identificativo della lavorazione;
- task eseguiti;
- stato di avanzamento;
- esito delle singole elaborazioni;
- eventuali errori;
- informazioni necessarie al monitoraggio e alla diagnosi.

Gli artefatti prodotti vengono associati alla pagina o all'aggregazione documentale sorgente e resi disponibili al completamento delle elaborazioni previste. Sono previsti meccanismi di nuovo tentativo per gli errori temporanei e una modalità di ripresa successiva a una correzione del contenuto sorgente o di un artefatto revisionabile.

Quando l'aggiornamento riguarda esclusivamente i dati o gli artefatti, senza modificare il modello dati, la struttura delle API o la pipeline applicativa, non è necessario introdurre una nuova versione del Data Product.

## 5.6. Limiti di utilizzo

Le funzioni di ricerca adottano un sistema di paginazione basato sul numero di pagina e sul numero di risultati restituiti per pagina. L'utilizzo della paginazione consente di limitare la dimensione delle risposte e di gestire in modo efficiente la consultazione di insiemi numerosi di oggetti.

Gli eventuali limiti relativi al numero di richieste per unità di tempo, alle richieste concorrenti, alla dimensione massima delle risposte e della pagina, ai timeout e alle soglie di throttling sono definiti a livello di ambiente di esercizio e applicati dal livello infrastrutturale di esposizione.

## 5.7. Accesso alla documentazione

Il Data Product è corredato dalla documentazione prevista per la pubblicazione nel catalogo DPaaS, comprendente:

- descrizione pubblica del Data Product;
- manuale utente;
- manuale operativo;
- specifiche tecniche di elaborazione;
- documentazione tecnica delle API;
- specifica OpenAPI;
- interfaccia di consultazione e test delle API.

La documentazione pubblica descrive a livello funzionale le modalità di fruizione del Data Product. La documentazione tecnica di dettaglio, comprendente operazioni, parametri, modelli dati, formati delle risposte ed esempi di chiamata, è resa disponibile secondo le modalità previste dal catalogo DPaaS e dal processo di richiesta di consumo.

## 5.8. Condizioni di licenza e riutilizzo

I dati, i metadati e gli artefatti sono riutilizzabili secondo la licenza indicata nella scheda del Data Product pubblicata nel catalogo DPaaS.

Restano applicabili gli eventuali diritti e vincoli associati alle singole immagini, ai manoscritti, ai fondi archivistici, agli asset messi a disposizione dall'Archivio di Stato dell'Aquila e agli artefatti derivati prodotti dalle elaborazioni automatiche.

Le condizioni di riutilizzo devono specificare almeno:

- possibilità di copia e redistribuzione;
- possibilità di modifica o elaborazione;
- eventuali obblighi di attribuzione;
- eventuali restrizioni all'utilizzo commerciale;
- modalità di citazione della fonte;
- eventuali vincoli derivanti da diritti di terzi.

La licenza definitiva applicabile al Data Product e ai suoi contenuti è quella selezionata e pubblicata dal titolare nella relativa scheda del catalogo DPaaS.

## 5.9. Fruizione tramite API

L'API di L'Aquila fornisce accesso programmatico, in sola lettura, agli oggetti documentali del Data Product e ai relativi contenuti accessibili. Consente alle applicazioni consumer di ricercare, recuperare e navigare manoscritti, pagine e collezioni, insieme ai metadati descrittivi e tecnici, alle informazioni di classificazione e agli artefatti generati dalle pipeline di accessibilità.

### 5.9.1. Tipologie di informazioni interrogabili

Le principali tipologie di informazioni interrogabili comprendono:

- oggetti digitali e relativi identificativi;
- manoscritti, documenti, collezioni e singole pagine;
- riferimenti ai file e agli asset associati;
- titolo, descrizione, autore o fondo di provenienza, quando disponibili nei metadati;
- datazione e copertura temporale, quando disponibili;
- tipologia, sottotipo e stato del ciclo di vita;
- lingua, metadati descrittivi e metadati tecnici;
- classificazioni, tag e parole chiave, quando presenti;
- relazioni tra manoscritti, pagine, collezioni e artefatti derivati;
- trascrizioni, traduzioni, spiegazioni sintetiche e contenuti complessivi in italiano e inglese;
- tracce audio associate alle spiegazioni.

### 5.9.2. Principali operazioni disponibili

Le principali operazioni messe a disposizione dall'API sono:

- ricerca degli oggetti digitali;
- recupero del dettaglio di un singolo oggetto;
- recupero dei metadati completi di una risorsa;
- recupero degli oggetti collegati e consultazione delle relazioni;
- consultazione di stati, lingue, tipi, sottotipi, tag e tipi di relazione;

- consultazione delle aggregazioni documentali disponibili;
- recupero delle immagini e delle risorse audio.

Le operazioni di consultazione non modificano il contenuto del Data Product. Le funzioni di revisione e validazione sono gestite dal portale e dai servizi applicativi riservati agli operatori autorizzati e non fanno parte dell'interfaccia pubblica read-only.

### 5.9.3. Ricerca, filtro, ordinamento e paginazione

L'API mette a disposizione una funzione di ricerca che consente di combinare più criteri. In base ai metadati effettivamente disponibili, la ricerca può essere filtrata rispetto a:

- codice o identificativo dell'oggetto;
- titolo;
- autore, soggetto produttore o fondo di provenienza;
- data di caricamento;
- lingua;
- stato;
- tipologia e sottotipo;
- manoscritto, documento o collezione di appartenenza;
- parole chiave o tag;
- presenza di relazioni con un oggetto padre o con pagine collegate.

Per la specificità del patrimonio dell'Aquila assumono particolare rilevanza i criteri relativi al manoscritto o fondo archivistico, alla pagina, alla datazione, alla lingua e alla disponibilità degli artefatti accessibili.

I risultati possono essere ordinati in modalità ascendente o discendente rispetto ai campi supportati. La paginazione viene gestita attraverso il numero della pagina richiesta e il numero di risultati da restituire per pagina.

### 5.9.4. Recupero del dettaglio di una risorsa

Il dettaglio di una risorsa viene recuperato attraverso il relativo identificativo univoco. La risposta contiene le informazioni principali dell'oggetto, tra cui identificativo, codice, classificazione, tipologia e sottotipo, stato, riferimento al file e appartenenza al documento o alla collezione, quando disponibili.

I metadati completi possono comprendere titolo, autore o soggetto produttore, datazione, descrizione, note, lingua, metadati tecnici e descrittivi, parole chiave e artefatti di accessibilità disponibili. Le relazioni della risorsa consentono di individuare le pagine appartenenti allo stesso manoscritto, gli oggetti collegati e gli artefatti generati a partire dalla risorsa originale.

### 5.9.5. Servizi di aggregazione e analisi

L'API può esporre aggregazioni funzionali alla navigazione del patrimonio, ad esempio per manoscritto, raccolta o altra classificazione disponibile nei metadati. Gli aggregati possono includere il numero di oggetti associati a ciascun raggruppamento, consentendo al portale di costruire percorsi di consultazione coerenti con la struttura documentale.

Le API non costituiscono un sistema di Business Intelligence e non forniscono elaborazioni analitiche avanzate, KPI o reporting multidimensionale. Le funzioni di aggregazione sono finalizzate principalmente alla navigazione del patrimonio e alla costruzione dell'esperienza di consultazione.

### 5.9.6. Formati delle richieste e delle risposte

Le richieste utilizzano il protocollo HTTPS. Le operazioni di consultazione sono effettuate principalmente tramite metodo HTTP GET e utilizzano parametri nel percorso per identificare una specifica risorsa e parametri di interrogazione per definire filtri, lingua, ordinamento e paginazione.

Le risposte sono restituite in formato JSON. I contenuti binari, quali immagini e tracce audio, vengono restituiti in streaming con il formato coerente con la risorsa. La struttura delle richieste, delle risposte e dei modelli informativi è descritta nella specifica OpenAPI.

### 5.9.7. Modalità di autenticazione

L'accesso machine-to-machine alle API avviene mediante una chiave identificativa dell'utente o dell'applicazione e il codice del Data Product. La chiave permette di identificare il consumer e di applicare le regole di accesso previste dall'ambiente di esercizio.

Le credenziali per l'accesso alle componenti di persistenza e allo storage non sono trasferite ai consumer e rimangono confinate nei componenti backend. L'accesso pubblico tramite portale non richiede autenticazione per i contenuti dichiarati pubblici; le funzionalità riservate agli operatori richiedono autenticazione SPID/CIE e un ruolo applicativo autorizzato.

### 5.9.8. Limiti di chiamata

L'API utilizza la paginazione per controllare la quantità di dati restituita nelle operazioni di ricerca. Gli eventuali limiti relativi a frequenza delle chiamate, quota giornaliera o oraria, chiamate concorrenti, dimensione massima della pagina, timeout e dimensione massima delle risposte sono definiti nella configurazione dell'ambiente di esercizio e applicati dal livello infrastrutturale di esposizione.

### 5.9.9. Gestione degli errori

L'API utilizza codici di stato HTTP standard per comunicare l'esito delle richieste. Le principali categorie di errore riguardano:

- parametri mancanti o non validi;
- formato non corretto di un parametro;
- identificativo della risorsa non valido o risorsa non esistente;
- autenticazione assente o non valida;
- accesso non autorizzato;
- superamento degli eventuali limiti di utilizzo;
- indisponibilità temporanea delle componenti di persistenza o dello storage;
- errore interno durante l'elaborazione della richiesta.

Le risposte di errore sono restituite in formato JSON e contengono le informazioni necessarie per identificare la tipologia e la causa del problema. Gli errori di validazione dei parametri vengono distinti dagli errori applicativi o infrastrutturali.

I controlli tecnici sullo stato e sulla disponibilità del servizio consentono ai sistemi di monitoraggio di verificare separatamente che il processo applicativo sia attivo e che il servizio sia pronto a gestire le richieste.

Le pipeline di alimentazione gestiscono separatamente gli errori di elaborazione. Lo stato della lavorazione, la fase interessata e il messaggio di errore vengono registrati per consentire il monitoraggio, il nuovo tentativo automatico o la ripresa manuale dell'elaborazione.

# 6. Limitazioni

Il Data Product L'Aquila è costituito dall'insieme delle immagini digitali, dei metadati e degli artefatti relativi ai manoscritti conferiti nell'ambito del progetto. Non rappresenta pertanto l'intero patrimonio documentale dell'Archivio di Stato dell'Aquila e la sua completezza dipende dalla disponibilità, dalla qualità e dalla corretta associazione dei materiali ricevuti.

Il Data Product ha natura prevalentemente statica e non real-time. Gli aggiornamenti non avvengono in modo continuo, ma sono effettuati:

- manualmente, a seguito della consegna o della revisione di immagini e metadati;
- mediante esecuzioni pianificate delle pipeline, secondo la configurazione dell'ambiente;
- tramite eventuali elaborazioni di arricchimento successive al caricamento.

Di conseguenza, tra la modifica della fonte e la pubblicazione del contenuto aggiornato può intercorrere un intervallo legato alla frequenza di esecuzione, al completamento delle pipeline e alle eventuali attività di revisione.

La qualità degli artefatti generati dipende dalla leggibilità del manoscritto, dallo stato di conservazione e dalla qualità dell'immagine digitalizzata. Calligrafie irregolari, parti deteriorate, annotazioni sovrapposte, impaginazioni complesse o pagine incomplete possono determinare trascrizioni o interpretazioni non pienamente accurate. Per questo motivo gli artefatti modificabili devono essere considerati contenuti derivati da sottoporre, quando necessario, alla verifica degli operatori autorizzati.

## 6.1. Nota sulle condivisioni dei dati

La condivisione, la pubblicazione e il riutilizzo dei dati e degli artefatti del Data Product sono regolati dalle condizioni indicate nella scheda pubblicata nel catalogo DPaaS e dagli eventuali diritti associati ai manoscritti, alle immagini digitali e ai fondi archivistici di provenienza.

La fruizione pubblica riguarda i contenuti dichiarati pubblicabili dal titolare. Le funzionalità di revisione, modifica e validazione sono riservate agli utenti autenticati e autorizzati. L'accesso machine-to-machine è subordinato alle credenziali e alle regole configurate nell'ambiente di esercizio.

La condivisione può riguardare le immagini originali, i metadati, una selezione di attributi o i soli artefatti derivati, secondo quanto definito dal titolare e riportato nella licenza applicabile. Gli obblighi di attribuzione, gli eventuali divieti di modifica o riutilizzo commerciale e i vincoli derivanti da diritti di terzi devono essere rispettati dai consumer.

Le credenziali tecniche e le informazioni necessarie all'accesso diretto alle componenti backend non rientrano nei dati condivisi. I consumer accedono esclusivamente attraverso i canali di esposizione previsti dal Data Product.

## 6.2. 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.

# 7. Possibili evoluzioni future

L'architettura realizzata è stata concepita in modo modulare, al fine di consentire future evoluzioni sia dal punto di vista tecnologico sia dal punto di vista funzionale. Una prima area di miglioramento riguarda la possibilità di sostituire o aggiornare i modelli di Intelligenza Artificiale attualmente impiegati con versioni più recenti o con modelli differenti, qualora disponibili sulla piattaforma DPaaS o ritenuti maggiormente performanti rispetto agli specifici casi d'uso. Tale approccio consentirebbe di beneficiare dei continui progressi tecnologici nel campo dell'AI senza richiedere modifiche significative ai processi di orchestrazione e pubblicazione del Data Product.

Un'ulteriore evoluzione riguarda la trasformazione delle pipeline attualmente sviluppate nell'ambito del progetto in veri e propri **servizi CPA (Common Processing Activities)** integrati all'interno dell'ecosistema DPaaS. In questo scenario, le logiche di generazione degli artefatti verrebbero rese disponibili come servizi condivisi e riutilizzabili, consentendo agli Enti di invocare un unico servizio applicativo per ottenere direttamente gli artefatti di accessibilità richiesti. Questa evoluzione permetterebbe di semplificare i processi di integrazione, ridurre la complessità operativa e favorire il riuso delle componenti sviluppate anche in altri Data Product o progetti futuri.

Dal punto di vista funzionale, potrebbe inoltre essere introdotto un meccanismo di **aggiornamento dinamico del Data Product**, che consenta l'acquisizione automatica di nuovi contenuti successivamente alla pubblicazione iniziale. In tale scenario, l'aggiunta di nuovi file di input da parte degli Enti attiverebbe automaticamente le pipeline di elaborazione e la successiva produzione degli artefatti derivati, mantenendo il Data Product costantemente aggiornato senza richiedere interventi manuali di riconfigurazione.

Tale funzionalità non è stata implementata nella presente versione della soluzione a causa di alcune limitazioni tecnologiche e operative riscontrate nell'attuale piattaforma DPaaS, che non consentono una gestione completamente automatizzata dell'evoluzione del contenuto del Data Product una volta completato il ciclo di pubblicazione iniziale. Qualora tali limitazioni venissero superate in future versioni della piattaforma, sarebbe possibile estendere l'architettura per supportare scenari di aggiornamento continuo e incrementale dei contenuti, migliorando ulteriormente la capacità del sistema di adattarsi all'evoluzione del patrimonio digitale degli Enti aderenti.

Un'ulteriore area di evoluzione riguarda i processi di riconoscimento e trascrizione dei manoscritti storici e medievali. Attualmente le pipeline implementate si basano su

modelli general-purpose di riconoscimento documentale e analisi multimodale, in quanto non risultano disponibili modelli specificamente addestrati sulle differenti tipologie di scritture manoscritte presenti nelle collezioni trattate dagli Enti coinvolti. Sebbene i risultati ottenuti siano complessivamente soddisfacenti, le prestazioni potrebbero essere ulteriormente migliorate attraverso attività di **fine tuning** dei modelli impiegati.

A tal fine sarebbe necessario disporre di una quantità significativa di dati di addestramento validati, costituiti da immagini dei manoscritti e relative trascrizioni corrette (ground truth). In particolare, **per ciascuna tipologia documentaria o classe omogenea di manoscritti**, sarebbe opportuno raccogliere un campione di circa **150 pagine trascritte e validate manualmente**, così da consentire l'addestramento supervisionato dei modelli sulle specifiche caratteristiche paleografiche e linguistiche dei documenti.

La disponibilità di tale patrimonio di dati permetterebbe di valutare la realizzazione di un processo di fine tuning dedicato, con l'obiettivo di aumentare l'accuratezza della trascrizione automatica, ridurre gli errori di riconoscimento e migliorare la qualità complessiva degli artefatti derivati generati dalle pipeline di accessibilità. Questa evoluzione potrebbe risultare particolarmente efficace per fondi archivistici caratterizzati da grafie ricorrenti, scritture medievali specialistiche o collezioni documentarie omogenee, nelle quali i modelli potrebbero apprendere caratteristiche specifiche non adeguatamente rappresentate nei dataset di addestramento generici disponibili oggi.