In viaggio su ali di Carta

Archivio di Stato di Genova

Sommario

1. Obiettivi del Data Product 3
2. Origine dei dati ed eventuali elaborazioni 4
3. Tecnologie utilizzate all'interno del Data Product Lab 5
   3.1. Servizi abilitanti DPaaS 5
   3.2. Eventuali tecnologie esterne a DPaaS 5
4. Pipeline di elaborazione e trasformazione 6
5. Esposizione e distribuzione 7
   5.1. Nota sulle condivisioni dei dati 9
6. Possibili evoluzioni future 10

1. Obiettivi del Data Product

Il Data Product nasce dall'esigenza di rendere accessibile e fruibile, in forma digitale strutturata e arricchita, il corpus di documenti archivistici medievali conservati presso l'Archivio di Stato di Genova, valorizzando un patrimonio documentario altrimenti consultabile solo tramite schede catalografiche tradizionali.
Il dataset è inoltre utilizzato, in parte, per alimentare le schede informative dell'applicazione immersiva 3D per tablet e VR "In viaggio su ali di carta".

Categorie di utenti destinatari:

- Enti pubblici e istituzioni accademiche del settore beni culturali
- Centri di ricerca storica e archivistica sul patrimonio documentario genovese
- Sviluppatori di applicazioni digitali per la cultura
- Operatori del settore archivistico e museale

Risultati attesi: un dataset multilingue e multimediale (testo, immagini, audio), consultabile tramite API, che ne consenta il riuso in nuovi contesti applicativi.

Principali casi d'uso:

- Consultazione e ricerca scientifica sul patrimonio archivistico medievale genovese
- Integrazione in sistemi di catalogazione MiC
- Sviluppo di applicazioni immersive e di valorizzazione digitale
- Produzione di audio-guide e contenuti multimediali accessibili

2. Origine dei dati ed eventuali elaborazioni

Fonte: schede catalografiche fornite dall'Archivio di Stato di Genova, trasformate in formato JSON strutturato.

Titolare della fonte: Archivio di Stato di Genova.

Struttura dei dati in ingresso: per ciascuna scheda sono forniti: identificativo, titolo, descrizione narrativa del documento e del suo contesto storico, un array di immagini associate (dove presenti) e un blocco di metadati strutturati (extra) con regesto archivistico, edizione bibliografica e testo divulgativo.

Volume indicativo: 21 schede (documenti archivistici dal 958 al 1462, secc. X-XV), 19 immagini, 17 regesti archivistici e 19 edizioni bibliografiche disponibili, 20 testi divulgativi.

Modalità di acquisizione: conferimento del dataset strutturato da parte dell'ente; non risulta una connessione diretta o schedulata verso un sistema sorgente esterno.

Elaborazioni precedenti l'ingresso nel DP Lab: nessuna; i dati giungono come schede già strutturate e compilate dall'ente, è stato processato il contenuto per avere un JSON strutturato come input del DP.

Elaborazioni realizzate all'interno del DP Lab:

- Indicizzazione delle schede a partire dal manifest e materializzazione in un archivio di lavoro locale, con tracciamento indipendente di ogni scheda e di ogni immagine
- Segmentazione delle descrizioni lunghe in blocchi testuali coerenti, unità di riferimento per l'allineamento tra testo e traccia audio
- Normalizzazione delle immagini per l'inferenza (ridimensionamento entro 1000×1000 px), senza alterazione degli originali

Il dato originale in italiano resta la fonte di verità e non viene mai sovrascritto dai contenuti generati, viene arricchito dalle informazioni extra elaborate.

3. Tecnologie utilizzate all'interno del Data Product Lab

3.1. Servizi abilitanti DPaaS

Servizio	Funzione	Fase della pipeline
Qwen2 (modalità testuale, pipeline cloud)	Traduzione di titolo, descrizione e metadati strutturati (regesto, edizione, testo divulgativo) da IT a EN, FR, DE, ES, PT, con protezione di codici, identificativi, date e nomi propri	Generazione contenuti testuali
Qwen2 (modalità multimodale)	Descrizione visiva delle immagini dei documenti, multilingue	Generazione contenuti da immagine
Kokoro TTS (esposto via IPAC)	Sintesi vocale della descrizione in italiano, generata per singolo blocco testuale	Generazione audio
DP LAB	Gestione del ciclo di vita delle lavorazioni (stati: in attesa → completata → approvata, con stato di errore) e dell'export versionato	Approvazione e pubblicazione

3.2. Eventuali tecnologie esterne a DPaaS

Servizio	Funzione	Fase della pipeline
Storage S3 (bucket cin-mic-dpaas-prod)	Persistenza degli snapshot versionati del dataset e degli asset binari (immagini, audio)	Bucket privato; accesso solo tramite l'API applicativa
SQLite (locale al container)	Indicizzazione runtime dello snapshot attivo per servire le query dell'API	Ricreato ad ogni avvio del container, non persistente
Docker / Kubernetes	Containerizzazione e orchestrazione dell'API di esposizione	Immagine unica, nessun servizio stateful esterno

4. Pipeline di elaborazione e trasformazione

1. **Ingestione** — le schede vengono indicizzate dal manifest ingestion.json e materializzate in un archivio di lavoro locale; scheda e immagine diventano unità tracciate indipendentemente.
2. **Segmentazione** — le descrizioni lunghe vengono suddivise in blocchi testuali coerenti, unità su cui si allineeranno traduzioni e audio.
3. **Normalizzazione immagini** — ridimensionamento entro 1000×1000 px per l'inferenza, in un'area di lavoro separata dagli originali.
4. **Generazione contenuti testuali** (Qwen2) — traduzione di titolo, descrizione e metadati strutturati (regesto, edizione, testo divulgativo) in EN/FR/DE/ES, con strategia adattiva (inferenza multilingue unica per testi brevi, inferenza dedicata per lingua per testi lunghi).
5. **Generazione contenuti da immagine** (vision AI) — descrizione visiva multilingue delle immagini dei documenti.
6. **Sintesi vocale** (Kokoro TTS) — generazione dell'audio per blocco testuale, in italiano.
7. **Controllo e approvazione** — ogni operazione segue il ciclo di vita in attesa → completata → approvata (o errore); solo i contenuti approvati da un operatore umano procedono verso la pubblicazione. Il controllo è granulare per operazione e per lingua; le lavorazioni non riuscite sono rieseguibili in modo selettivo, senza rigenerare l'intera scheda.
8. **Export e versionamento** — produzione di uno snapshot versionato su storage S3; ogni versione è distinta e non sovrascrivibile.
9. **Pubblicazione** — l'API di fruizione legge automaticamente il nuovo snapshot, senza richiedere intervento tecnico o interruzione di servizio.

Modalità di esecuzione: la pipeline di arricchimento opera su richiesta/per lotti sulle schede in ingresso; l'aggiornamento della versione servita dall'API avviene invece per sostituzione dello snapshot, tramite riavvio del container con la nuova versione configurata.

5. Esposizione e distribuzione

Fruizione tramite API

Il Data Product è esposto tramite un'**API REST read-only** (containerizzata, FastAPI su porta 8000), pensata per essere integrata da sviluppatori e applicazioni terze.

- **Canale di accesso:** endpoint HTTP, esposti pubblicamente tramite un layer di ingress/reverse proxy (non incluso nel container)
- **Utenti/sistemi destinatari:** sviluppatori e applicazioni che integrano il dataset (es. app di valorizzazione culturale, sistemi di catalogazione)
- **Autenticazione e autorizzazione:** non implementate a livello applicativo; il controllo accessi è demandato al layer di esposizione (ingress, reverse proxy, API gateway)
- **Formati disponibili:** JSON
- **Principali funzionalità:**
  - GET /v1/items — lista paginata degli item, con filtri su disponibilità audio/immagini e versione del dataset
  - GET /v1/items/{id} — dettaglio completo di un item: titolo, descrizioni multilingua, immagini con metadati estratti, audio
  - Paginazione tramite page / page_size
  - Consultazione di versioni storiche del dataset tramite il parametro version
- **Modalità di aggiornamento:** automatica al momento della pubblicazione di un nuovo snapshot, senza interruzione di servizio
- **Limiti di utilizzo:** non risultano rate limit o quote documentati
- **Livelli di servizio:** singola replica; disponibilità dei due endpoint di health check (/health, /health/ready) per il monitoraggio
- **Documentazione:** Manuale Utente (per chi integra/consuma le API) e Manuale Operativo (per il team tecnico responsabile di deploy e gestione)
- **Condizioni di licenza:** da definire nei metadati DPaaS del Data Product (categoria: Valorizzazione Immersiva & Storytelling)

6. Limitazioni

- Il corpus copre **21 schede** su un patrimonio documentario più ampio conservato presso l'Archivio di Stato di Genova: non rappresenta l'intera collezione disponibile
- **4 schede su 21** risultano prive di immagini a corredo;
- Alcune schede risultano prive di regesto (4 su 21) o di edizione bibliografica (2 su 21);
- Le traduzioni e le descrizioni generate automaticamente (Qwen2) possono contenere imprecisioni; sono per questo soggette a un ciclo di approvazione umana prima della pubblicazione
- Non risulta documentata alcuna ridondanza applicativa: il deployment di riferimento prevede una singola replica del container

5.1. Nota sulle condivisioni dei dati

- L'accesso ai dati avviene tramite il Catalogo pubblico DPaaS, previa richiesta di consumo dell'API
- I dati condivisi comprendono schede documentarie arricchite con regesto archivistico, edizione bibliografica, testo divulgativo, traduzioni multilingue e immagini; non risultano presenti dati personali, riservati o soggetti a restrizioni particolari
- L'ente titolare del dato di partenza è l'Archivio di Stato di Genova

6. Possibili evoluzioni future

- Estensione del corpus ad altri documenti archivistici conservati presso l'Archivio di Stato di Genova
- Completamento dei regesti archivistici e delle edizioni bibliografiche attualmente non valorizzati per migliorare la qualità del dato e l'affidabilità dell'apparato critico
- Completamento del corredo fotografico per le 4 schede attualmente prive di immagini
- Introduzione di ridondanza/alta disponibilità per il servizio API