ICCU – Il Catalogo Accessibile

Documento di specifiche tecniche di elaborazione

Versione del documento: 1.0

Data: 28/08/2026

Sommario

1. Obiettivi del Data Product 3
2. Origine dei dati ed eventuali elaborazioni 4
3. Tecnologie utilizzate all'interno del Data Product Lab 10
3.1. Servizi abilitanti DPaaS 11
3.2. Eventuali tecnologie esterne a DPaaS 13
3.2.1. Modelli in locale 13
4. Pipeline di elaborazione e trasformazione 15
4.1. Pipeline di generazione di riassunti accessibili per libri digitalizzati 15
4.1.1. Verifica preliminare della qualità del contenuto 15
4.1.2. Suddivisione del libro in sezioni elaborabili 15
4.1.3. Generazione dei riassunti dei singoli capitoli o sezioni 16
4.1.4. Riassunto del chunk 16
4.1.5. Costruzione del riassunto complessivo del libro 16
4.1.6. Generazione del riassunto in lingua inglese 16
4.1.7. Produzione della versione audio del riassunto in italiano 17
4.1.8. Produzione della versione audio del riassunto in inglese 17
4.1.9. Revisione e validazione degli artefatti 17
4.1.10. Pubblicazione degli artefatti 17
5. Esposizione e distribuzione 18
5.1. Canali di accesso 18
5.2. Contenuti resi disponibili 18
5.3. Fruizione tramite API 18
5.4. Utenti e sistemi destinatari 19
5.5. Modalità di aggiornamento 19
5.6. Condizioni di accesso, riutilizzo e condivisione 19
5.7. Limitazioni 20
6. Possibili evoluzioni future 21

1.Obiettivi del Data Product

Il Data Product ICCU ha l'obiettivo di trasformare testi OCR già disponibili su storage in contenuti accessibili, sintetici e riutilizzabili all'interno della Digital Library.

2. Origine dei dati ed eventuali elaborazioni

1. Catalogo bibliografico SBN tramite I.PaC

Sistema, archivio o dataset di origine

La fonte bibliografica principale è costituita dall'indice SBN disponibile all'interno di I.PaC.

I dati del catalogo SBN, originariamente rappresentati secondo il formato e il modello bibliografico UNIMARC, sono trasformati a monte nel modello informativo adottato da I.PaC. Il template bibliografico I.PaC comprende i metadati descrittivi delle opere.

I record bibliografici costituiscono la base rispetto alla quale vengono ricondotte e organizzate le differenti tipologie di risorse accessibili.

Soggetto responsabile o titolare della fonte

Il soggetto istituzionale di riferimento per il catalogo SBN è l'**Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche, ICCU**.

La fornitura utilizzata dal progetto è resa disponibile nell'ambito dell'ecosistema Digital Library attraverso I.PaC.

Tipologia e formato dei dati

La fornitura ricevuta è composta da **15 file JSON di grandi dimensioni**.

Ogni elemento rappresenta un record bibliografico identificato dal codice BID e può contenere, tra le altre, le seguenti informazioni:

- BID;
- titolo;
- autore e responsabilità;
- ISBN;
- genere;
- informazioni relative all'edizione;
- descrizioni e note;
- informazioni relative alla disponibilità di copie in Braille.

Il BID rappresenta l'identificativo utilizzato per riconoscere univocamente il record bibliografico all'interno del perimetro I.PaC/SBN.

Per facilitare il caricamento e il trattamento tecnico, i file JSON originali vengono ridotti di dimensione selezionando solo le informazioni necessarie e convertiti nel formato **NDJSON**, nel quale ogni riga rappresenta un record JSON autonomo.

Modalità di acquisizione

I file della fornitura vengono depositati nell'object storage S3 associato al Data Product.

L'acquisizione avviene tramite pipeline dedicate che:

1. individuano i file JSON presenti nel bucket;
2. selezionano le informazioni necessarie;
3. convertono i dati in formato NDJSON;
4. suddividono i file in unità di dimensioni elaborabili;
5. caricano i record nella tabella di staging dedicata applicando una logica di UPSERT sulla chiave BID.

Frequenza di aggiornamento

L'aggiornamento avviene su nuova fornitura o nuova estrazione resa disponibile da I.PaC; periodicità operativa non formalizzata.

Il processo è predisposto per acquisire nuove estrazioni e aggiornare i record esistenti mediante UPSERT.

Volume indicativo dei dati

La fornitura originaria comprende 15 file JSON di grandi dimensioni.

Durante la preparazione tecnica dei dati è stata prevista la generazione di circa **300 file NDJSON da circa 500 MB ciascuno**. Tale dimensionamento ha evidenziato la necessità di:

- suddividere i file originali;
- evitare, ove possibile, il download locale tramite NiFi;
- gestire attentamente lo spazio del content repository;
- dimensionare opportunamente il database PostgreSQL utilizzato dal Data Product.

Vincoli di accesso, utilizzo o licenza

L'accesso ai file sorgente avviene attraverso il bucket S3 del Data Product e richiede le credenziali e le autorizzazioni previste per l'ambiente DP Lab.

Non risultano trattati dati personali nell'ambito della specifica fornitura bibliografica utilizzata per il Data Product.

I file originali sono utilizzati come sorgente tecnica interna alla pipeline. Il Data Product espone esclusivamente le informazioni selezionate, consolidate e necessarie alla consultazione del catalogo accessibile.

Livello di qualità e completezza riscontrato

La qualità dei metadati bibliografici è complessivamente adeguata alle finalità del Data Product. Sono state tuttavia riscontrate alcune condizioni che devono essere considerate nell'utilizzo dei dati:

- alcune informazioni presenti nei record UNIMARC originari potrebbero non essere state trasferite integralmente nel modello I.PaC;
- i file originali contengono numerosi nodi e proprietà non che non sono stati inclusi nel database finale del Data Product in quanto le informazioni contenute non erano strettamente necessarie. L'eventuale recupero di ulteriori attributi presenti nel modello I.PaC richiedono di effettuare una nuova creazione dei file NDJSON.

2. Aggregatore Digitale e testi OCR

Sistema o dataset di origine

La seconda fonte è costituita dall'elenco degli OCR realizzati nell'ambito dell'iniziativa Aggregatore Digitale e dagli OCR stessi. Esso fornisce l'associazione tra BID e oggetti digitalizzati. La tabella dedicata contiene, ove disponibili, identificativo digitale, cover, preview, manifest e riferimento al file. I testi OCR rappresentano il contenuto testuale già prodotto a partire dalle digitalizzazioni e sono utilizzati per generare gli arricchimenti accessibili.

Soggetto responsabile o titolare

I dati appartengono al perimetro ICCU/Digital Library e derivano dai cantieri di digitalizzazione conferiti nell'ecosistema I.PaC.

Tipologia e formato dei dati

La fornitura comprende:

- un file Excel OCR.xlsx, contenente l'elenco dei contenuti OCR e i relativi riferimenti;
- testi OCR memorizzati su object storage S3;
- riferimenti a cover, preview, manifest IIIF e file digitali;

Modalità di acquisizione

Il file Excel presente su S3 viene elaborato da una pipeline dedicata e caricato nella tabella di staging dedicata. I testi OCR vengono letti dalle relative posizioni su S3 quando devono essere sottoposti all'arricchimento AI.

Frequenza di aggiornamento

Non è indicata una frequenza automatica e formalizzata. La fornitura disponibile è stata acquisita come conferimento e può essere aggiornata attraverso nuove versioni del file di indice o nuove disponibilità sul bucket.

Volume indicativo

Nel catalogo fornito sono presenti 23664 record. Di questi, **8012** trovano una corrispondenza nel catalogo bibliografico SBN fornito e utilizzati per l'output finale del Data Product.

Vincoli di accesso e utilizzo

I contenuti vengono letti dal bucket S3 autorizzato. Il Data Product elabora i testi OCR e i relativi metadati, mentre l'accesso o la redistribuzione del file digitale originale dipendono dalle condizioni applicabili al singolo oggetto.

Qualità e completezza

La qualità degli OCR non è uniforme. Prima dell'arricchimento viene eseguita una validazione sul testo; i record vuoti, illeggibili o eccessivamente rumorosi sono esclusi dalle elaborazioni successive. Circa 660 record non hanno superato la validazione e per tanto non sono stati elaborati.

Elaborazioni effettuate nel DP Lab

Nel DP Lab vengono svolte:

- acquisizione dell'indice Excel;
- associazione dei testi OCR ai record bibliografici mediante BID;
- verifica automatica della leggibilità del testo sulle prime 2.500 parole;
- interruzione del flusso in caso di OCR non valido;
- generazione del riassunto italiano;

- generazione della versione inglese;
- conversione dei riassunti in audio italiano e inglese;
- memorizzazione degli esiti di validazione e degli artefatti generati nel Data Product.

3. Liber Liber

Sistema o dataset di origine

La fonte è il catalogo digitale di Liber Liber (https://liberliber.it). La fornitura ricevuta comprende un dump SQL del catalogo del portale Liber Liber. Per l'integrazione è stata utilizzata la porzione del dump contenente i dati necessari al matching.

Soggetto responsabile o titolare

Il soggetto responsabile della fonte è Liber Liber.

Tipologia e formato dei dati

I dati comprendono metadati bibliografici, titolo, autore, ISBN o altri codici quando presenti e tipologia della risorsa. I materiali consegnati sono un dump SQL.

Modalità di acquisizione

Dal dump del database sono stati selezionati la tabella e i campi necessari, caricati nella tabella di staging dedicata. L'importazione avviene in modalità manuale, in attesa che venga definita un'interfaccia con Liber Liber che possa rendere automatica l'importazione di aggiornamenti e nuovi record.

Frequenza di aggiornamento

L'acquisizione è basata su fornitura puntuale dei file e richiede un nuovo caricamento in caso di aggiornamento del catalogo.

Volume indicativo

5.156 opere Liber Liber di cui 5055 libri e 101 audiolibri.

Vincoli di accesso, utilizzo o licenza

Il Data Product non replica il contenuto integrale della risorsa: viene esposto il collegamento (URL) alla pagina del portale Liber Liber relativa all'opera. I contenuti rimangono quindi a disposizione e fruibili tramite il portale Liber Liber alle condizioni previste dall'Organizzazione titolare dei contenuti.

Qualità e completezza

La disponibilità di ISBN e autore non è uniforme e i dati possono presentare differenze di accenti, maiuscole, punteggiatura, ordine dei nomi, refusi e varianti del titolo. Questa variabilità limita la diretta associazione con le opere presenti nel catalogo SBN. Dove non è possibile un collegamento diretto (tramite codice ISBN della fonte) è necessario applicare un algoritmo allo scopo di effettuare la normalizzazione di titolo e autore per associare correttamente le voci dei due cataloghi oltre che applicare in ultima istanza una logica fuzzy per il matching delle fonti.

Essendo il matching solo in parte deterministico, è necessaria una validazione del risultato delle associazioni effettuate dall'algoritmo.

Elaborazioni effettuate prima dell'ingresso nel DP Lab

La preparazione e la fornitura del dump del database del portale Liber Liber è effettuata dal titolare della fonte. Da questo dump è necessario estrarre il dato di interesse da importare nel DP Lab.

Elaborazioni effettuate nel DP Lab

Nel DP Lab vengono effettuate:

- caricamento nella staging dedicata;
- normalizzazione di ISBN, titolo e autore;
- neutralizzazione di differenze formali quali accenti, maiuscole, punteggiatura e ordine dei nomi;
- matching deterministico mediante ISBN, codici disponibili e corrispondenza esatta di titolo;
- matching su titolo e autore normalizzati;
- matching fuzzy, applicato ai record residui;
- memorizzazione dei match validi e dei candidati sotto soglia;
- associazione del link Liber Liber al record I.PaC.

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 - Tagging Entità;
- CPAC – WhisperTTS;
- CPAC – Whisper;
- CPAC – Tagging;
- CPAC - Stable Diffusion;
- CPAC - Qwen2;
- CPAC - P2PaLa;
- CPAC – NERs;
- CPAC - Llama 31;
- CPAC – Keras;
- CPAC- Gliner;
- VM;
- JupyterLab;
- LPW_DataPreparation;
- Neo4j;
- Ontotext Refine;
- OpenSeaDragon;
- NiFi;

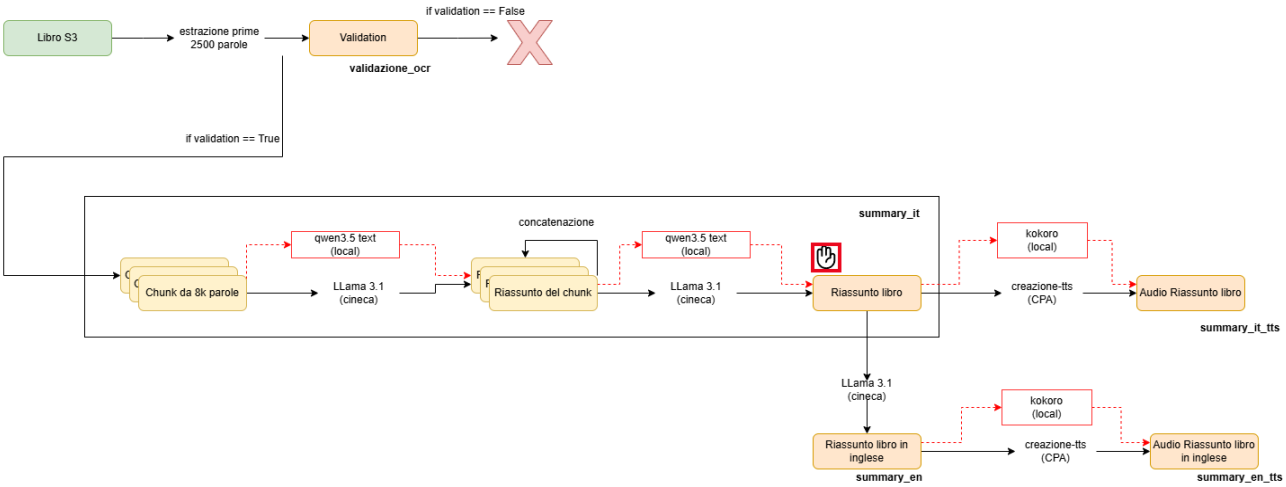
- Postgres;
- Elastic Search;
- GraphDB.

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 dagli alberi elaborativi, al fine di garantire l'accessibilità delle risorse vengono generati:

INPUT: *OCR di un libro*

OUTPUT :

- Riassunto del libro in Italiano e in inglese
- Audio riassunto del libro 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 generazione di riassunti accessibili per libri digitalizzati

La pipeline è progettata per produrre riassunti strutturati e contenuti audio accessibili a partire da libri digitalizzati archiviati all'interno del repository documentale. Il processo combina tecniche di validazione preliminare della qualità del testo, elaborazione distribuita di grandi quantità di contenuto e modelli linguistici avanzati per generare sintesi complete dell'opera in italiano e in inglese. Tutti gli artefatti prodotti vengono archiviati all'interno del Data Product e resi disponibili tramite il portale. Gli elementi contrassegnati dall'icona della mano possono essere revisionati e modificati dagli utenti autorizzati prima della pubblicazione.

4.1.1. Verifica preliminare della qualità del contenuto

La pipeline prende in ingresso un libro precedentemente archiviato in un bucket S3.

Prima dell'avvio delle elaborazioni viene effettuata una fase di controllo automatico della qualità del testo estratto. A tale scopo vengono analizzate le prime **2.500 parole** del documento, allo scopo di verificare che il contenuto sia sufficientemente leggibile e che le precedenti attività di digitalizzazione e riconoscimento ottico dei caratteri abbiano prodotto risultati adeguati.

L'esito della verifica viene registrato nell'artefatto **validazione_ocr**.

Nel caso in cui la validazione non venga superata, il processo viene interrotto per evitare la generazione di contenuti derivati basati su dati non affidabili. Se invece la verifica produce esito positivo, il documento viene inoltrato alle successive fasi di elaborazione.

4.1.2. Suddivisione del libro in sezioni elaborabili

Poiché i libri possono contenere quantità molto elevate di contenuto testuale, il documento viene suddiviso in blocchi di dimensione controllata.

In particolare, il testo viene segmentato in **chunk da circa 8.000 parole**, consentendo un'elaborazione efficiente mediante i modelli linguistici e riducendo il rischio di perdita di informazioni rilevanti.

Ogni sezione viene trattata come un'unità autonoma e sottoposta a processi di sintesi indipendenti.

4.1.3. Generazione dei riassunti dei singoli capitoli o sezioni

Ciascun chunk viene elaborato mediante il modello **Llama 3.1**, eseguito presso l'infrastruttura di calcolo del **CINECA**, con il supporto del modello **Qwen 3.5 Text** eseguito localmente.

L'obiettivo di questa fase è produrre una sintesi coerente dei contenuti presenti nel blocco documentale, individuando gli elementi principali, gli argomenti trattati e le informazioni più rilevanti.

Per ogni sezione viene pertanto generato un artefatto di tipo:

4.1.4. Riassunto del chunk

capace di condensare il contenuto originale mantenendone il significato e il contesto.

4.1.5. Costruzione del riassunto complessivo del libro

I riassunti prodotti per i singoli chunk vengono successivamente concatenati e aggregati in un'unica rappresentazione sintetica dell'opera.

Il corpus risultante viene nuovamente elaborato mediante **Llama 3.1** e **Qwen 3.5 Text**, che generano un riassunto organico dell'intero volume.

Il risultato di questa fase è l'artefatto:

- **summary_it** – Riassunto del libro in italiano.

Questo contenuto rappresenta una sintesi completa dell'opera e consente una consultazione rapida dei temi principali trattati nel testo.

Poiché il riassunto costituisce uno degli artefatti più rilevanti della pipeline, il sistema consente agli operatori autorizzati di modificarlo e validarlo attraverso il portale. In questo modo è possibile correggere eventuali imprecisioni, integrare informazioni ritenute importanti o adattare il livello di dettaglio alle esigenze editoriali del progetto.

4.1.6. Generazione del riassunto in lingua inglese

A partire dal riassunto italiano validato, il modello **Llama 3.1** genera una corrispondente versione in lingua inglese.

L'artefatto prodotto è:

- **summary_en** – Riassunto del libro in inglese.

La traduzione mantiene la struttura e i contenuti fondamentali del riassunto originale, consentendo la consultazione dell'opera anche da parte di utenti internazionali.

4.1.7. Produzione della versione audio del riassunto in italiano

Il riassunto italiano viene convertito in contenuto audio mediante il servizio **Creazione TTS (CPA)** basato sul motore di sintesi vocale **Kokoro**, eseguito localmente.

Da questo processo viene generato l'artefatto:

- **summary_it_tts** – Audio del riassunto del libro in italiano.

L'audio consente agli utenti di accedere rapidamente ai contenuti del volume anche attraverso una modalità di ascolto, ampliando le possibilità di fruizione.

4.1.8. Produzione della versione audio del riassunto in inglese

In maniera analoga, il riassunto inglese viene elaborato dal servizio di sintesi vocale CPA per generare:

- **summary_en_tts** – Audio del riassunto del libro in inglese.

Anche questo contenuto viene prodotto utilizzando il motore **Kokoro**, garantendo uniformità qualitativa e coerenza tra le diverse lingue supportate.

4.1.9. Revisione e validazione degli artefatti

La pipeline prevede una fase di supervisione umana sul riassunto principale del volume. In particolare può essere modificato e validato:

- **summary_it** – Riassunto del libro in italiano.

Questo approccio consente di integrare il controllo editoriale umano con l'efficienza dei modelli di Intelligenza Artificiale, migliorando la qualità complessiva dei contenuti pubblicati.

4.1.10. Pubblicazione degli artefatti

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

- Esito della validazione OCR (*validazione_ocr*);
- Riassunto del libro in italiano (*summary_it*);
- Riassunto del libro in inglese (*summary_en*);
- Audio del riassunto in italiano (*summary_it_tts*);
- Audio del riassunto in inglese (*summary_en_tts*).

Attraverso questa pipeline il libro digitalizzato viene trasformato in un insieme di contenuti sintetici, multilingua e accessibili, consentendo agli utenti di comprenderne rapidamente i temi principali e di accedere alle informazioni anche tramite modalità di ascolto, favorendo una fruizione più inclusiva del patrimonio documentale

5. Esposizione e distribuzione

Il Data Product "ICCU – Il Catalogo Accessibile" rende disponibili i record bibliografici selezionati, i riferimenti ai contenuti digitalizzati e gli artefatti accessibili generati a partire dai testi OCR. L'esposizione è orientata alla consultazione del catalogo e al riutilizzo controllato delle informazioni da parte dei sistemi consumer autorizzati.

5.1. Canali di accesso

Le modalità di fruizione documentate sono:

- catalogo della piattaforma DPaaS, per la descrizione e la scoperta del Data Product e per l'avvio del processo di richiesta di consumo;
- API in sola lettura, per l'accesso applicativo ai dati e agli artefatti del catalogo accessibile.

Non sono previste operazioni pubbliche di modifica attraverso l'interfaccia di consultazione. Le attività di caricamento, aggiornamento, revisione e validazione restano separate dalle funzioni di lettura rese disponibili ai consumer.

5.2. Contenuti resi disponibili

Il perimetro informativo esposto comprende, in funzione della disponibilità nella fonte:

- identificativo BID e metadati bibliografici selezionati, inclusi titolo, autore e responsabilità, genere, editore, descrizioni;
- informazioni relative alla disponibilità di copie in Braille, quando presenti nel record, con riferimento alla biblioteca;
- riferimenti ai contenuti digitalizzati associati, quali url per il reperimento della copia su portali esterni, cover, preview, manifest degli OCR, quando disponibili;
- riassunti in italiano e in inglese dei testi per cui è disponibile un OCR;
- versioni audio dei riassunti in italiano e in inglese.

Il Data Product espone esclusivamente le informazioni selezionate, consolidate e necessarie alla consultazione del catalogo accessibile. Non replica automaticamente tutti gli attributi presenti nelle forniture originarie.

5.3. Fruizione tramite API

L'API consente la consultazione programmatica, in sola lettura, degli oggetti del catalogo accessibile e delle informazioni a essi collegate. A livello funzionale permette di:

- ricercare gli oggetti del catalogo utilizzando i criteri disponibili;

- recuperare il dettaglio di un singolo oggetto e i relativi metadati;
- consultare classificazioni, lingue, tipologie, tag e relazioni disponibili;
- navigare i collegamenti tra record bibliografici, contenuti digitalizzati e artefatti derivati;
- recuperare gli artefatti accessibili associati al record, quando presenti;
- gestire insiemi numerosi di risultati mediante filtri, ordinamento e paginazione, secondo le funzioni rese disponibili dall'interfaccia.

Le risposte informative sono rese in formato JSON. L'accesso applicativo è soggetto alle credenziali e alle autorizzazioni previste dall'ambiente di esercizio. La documentazione di dettaglio dell'interfaccia, comprensiva di operazioni, parametri e modelli delle risposte, è resa disponibile secondo il processo di consumo del Data Product.

5.4. Utenti e sistemi destinatari

Il Data Product è destinato principalmente a sistemi applicativi autorizzati dell'ecosistema Digital Library, operatori abilitati alla verifica dei contenuti e consumer autorizzati che necessitano di consultare o riutilizzare i metadati bibliografici e gli artefatti accessibili. Le modalità di accesso effettive dipendono dal profilo autorizzativo e dal canale di fruizione utilizzato.

5.5. Modalità di aggiornamento

L'aggiornamento avviene in modalità batch quando viene resa disponibile una nuova fornitura bibliografica, una nuova versione dell'indice OCR o nuovi contenuti sullo storage. I record bibliografici del catalogo SBN sono aggiornati mediante BID; i contenuti OCR associati possono quindi essere validati ed elaborati per produrre o aggiornare gli artefatti accessibili. Non è documentata una frequenza automatica formalizzata.

5.6. Condizioni di accesso, riutilizzo e condivisione

L'accesso ai dati sorgente e alle funzioni di gestione è riservato ai componenti e agli operatori autorizzati nell'ambiente DP Lab. I consumer ricevono esclusivamente i dati e gli artefatti previsti dal Data Product attraverso i canali di esposizione autorizzati.

Non risultano trattati dati personali nella specifica fornitura bibliografica descritta. L'accesso o la redistribuzione dei file digitali originali resta subordinato alle condizioni applicabili al singolo oggetto. Le condizioni definitive di licenza, attribuzione e riutilizzo devono pertanto essere quelle indicate dal titolare e pubblicate nella scheda del Data Product.

5.7. Limitazioni

Il Data Product presenta le seguenti limitazioni operative e informative:

- la copertura è limitata ai record e ai contenuti effettivamente presenti nelle forniture acquisite e correttamente associati mediante BID o in alternativa tramite ISBN, titolo e autore o logica fuzzy;
- le informazioni non trasferite nel modello I.PaC o non selezionate durante la preparazione dei dati non sono disponibili nel database finale;
- la qualità dei riassunti e delle relative versioni audio dipende dalla leggibilità e dalla completezza del testo OCR in ingresso;
- i testi OCR vuoti, illeggibili o eccessivamente rumorosi vengono esclusi dalle elaborazioni successive;
- gli aggiornamenti non sono continui né real-time, ma dipendono dalla disponibilità di nuove forniture e dall'esecuzione delle pipeline batch;
- l'eventuale disponibilità del file digitale originale e il relativo riutilizzo dipendono dalle condizioni applicabili alla singola risorsa;

le interfacce di consultazione sono in sola lettura e non consentono ai consumer di modificare direttamente i dati o gli artefatti pubblicati.

6. 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 evoluzione potrà riguardare l'ampliamento delle fonti documentarie elaborate dal Data Product. Nella presente fase progettuale, infatti, le attività di

acquisizione ed elaborazione sono state limitate alle risorse provenienti dal catalogo **Liber Liber**, principalmente a causa di vincoli tecnologici e autorizzativi emersi durante le attività di integrazione con le diverse piattaforme sorgente. Tale scelta ha consentito di validare le pipeline di accessibilità su un insieme omogeneo di contenuti liberamente utilizzabili e disponibili in formato digitale strutturato.

In prospettiva futura, una volta superati gli attuali vincoli tecnici e completate le eventuali attività autorizzative necessarie, le pipeline sviluppate potranno essere estese all'elaborazione di ulteriori cataloghi e raccolte documentarie dedicate all'accessibilità. A titolo esemplificativo, potranno essere valutate integrazioni con collezioni quali **Libro Parlato**, i cataloghi dell'**Unione Italiana dei Ciechi e degli Ipovedenti**, la **Biblioteca Italiana per i Ciechi "Regina Margherita" di Monza** e ulteriori repository digitali nazionali o territoriali che rendano disponibili contenuti compatibili con le modalità di acquisizione previste dall'architettura del Data Product.

Tale estensione consentirebbe di ampliare significativamente il patrimonio informativo accessibile attraverso la piattaforma, aumentando il numero di opere disponibili e favorendo la diffusione di contenuti inclusivi verso un pubblico sempre più ampio. Inoltre, l'applicazione delle medesime pipeline a cataloghi differenti consentirebbe di valorizzare ulteriormente gli investimenti effettuati, favorendo il riuso delle componenti tecnologiche già sviluppate e l'estensione delle funzionalità di accessibilità a nuove tipologie di contenuti e collezioni digitali.