Condition Report: dataset integrato a supporto del prestito

Documento di specifiche tecniche di elaborazione

Sommario

Sommario 2
1. Obiettivi del Data Product 3
2. Origine dei dati ed eventuali elaborazioni 6
3. Servizi abilitanti DPaaS 7
4. Pipeline di elaborazione e trasformazione 7
5. Esposizione e distribuzione 9
5.1. Nota sulle condivisioni dei dati 10
6. Possibili evoluzioni future 12

1.Obiettivi del Data Product

Nell'ambito del Lotto 4 del Dialogo Competitivo (PNRR – Missione 1, Componente 3, Investimento 1.1 "Strategie e piattaforme digitali per il patrimonio culturale"), il Data Product "Condition Report" - identificativo DPaaS DP00000117 - è stato sviluppato per il Museo delle Civiltà di Roma con l'obiettivo di supportare le attività di documentazione e di certificazione dello stato conservativo dei beni culturali e, allo stesso tempo, di valorizzare il patrimonio informativo già custodito dall'istituto.

L'obiettivo del Data Product è la realizzazione del condition report, strumento a supporto del prestito dei beni museali. A tal fine il Data Product integra i dati catalografici con le informazioni sul restauro e sulle analisi diagnostiche relative al bene in un dataset unificato e strutturato, arricchito mediante tecniche di intelligenza artificiale per l'estrazione di metadati semantici, rendendo il patrimonio accessibile e utilizzabile all'interno della piattaforma di gestione e documentazione.

Il Data Product genera valore trasformando un insieme di dati eterogenei e frammentati in un report informativo integrato, coerente e facilmente riutilizzabile. L'integrazione tra dati catalografici e documentazione archivistica consente di avere una visione completa e strutturata del bene, migliorando la qualità e l'efficacia delle attività di analisi e intervento. Dal punto di vista operativo questo si traduce in un supporto concreto alle attività di restauro e di prestito, permettendo ai professionisti di effettuare valutazioni, confronti e simulazioni; allo stesso tempo la standardizzazione e l'interoperabilità dei dati ne facilitano il riutilizzo in diversi contesti applicativi, riducendo duplicazioni e aumentando l'efficienza complessiva nella gestione del patrimonio.

Il valore generato dal Data Product si concretizza in tre pilastri fondamentali:

- **Efficienza nei flussi di lavoro**: semplifica la produzione della reportistica relativa ai prestiti, favorendo il tracciamento dello stato conservativo del bene;
- **Preservazione e valorizzazione digitale**: permette di tracciare nel tempo lo stato conservativo del bene e, grazie ai contenuti generati, rende il patrimonio più accessibile e comprensibile anche a un pubblico non specialista;
- **Supporto decisionale scientifico**: offre agli esperti di dominio - archivisti, conservatori, restauratori - uno strumento flessibile per valutare le azioni da intraprendere per la protezione del bene.

All'interno del catalogo della piattaforma DPaaS il Data Product è censito e opera nel perimetro di due macro-categorie: "Diagnostica, Restauro & Monitoraggio", che

raccoglie gli strumenti applicati alla conservazione preventiva e al restauro digitale, e "Gestione e arricchimento dei dati", relativa alla data augmentation e alla generazione di metadati descrittivi e visivi evoluti per la ricerca avanzata.

Il Data Product si rivolge alle seguenti figure professionali e profili di accesso operanti nell'ambito della piattaforma DPaaS del Ministero della Cultura:

- **Data Producer**, responsabile dell'inserimento della documentazione archivistica, curatoriale e conservativa nel Data Product, della configurazione iniziale dell'applicazione e dell'esecuzione e del monitoraggio;
- **Ricercatori, restauratori e tecnici**, utenti esperti di dominio che utilizzano la piattaforma di documentazione e gestione del bene museale e che sono abilitati alla creazione dei condition report tramite l'applicazione di gestione;
- **Data Consumer e utenti business**, professionisti, ricercatori ed enti terzi autorizzati che accedono al catalogo DPaaS per consultare, riutilizzare o integrare nei propri flussi di lavoro i dataset e i metadati generati;
- **Visitatori**, utenti finali che accedono alla piattaforma per fruire dei contenuti, con accesso limitato all'interfaccia di fruizione;
- **Product Manager e amministratori di piattaforma**, responsabili della supervisione e del monitoraggio delle risorse computazionali allocate, della gestione delle configurazioni di sistema e degli accessi all'applicazione remota.

Al termine dei processi di elaborazione il Data Product mette a disposizione la documentazione elaborata sotto forma di testi strutturati e categorizzati nei formati JSON e PDF, insieme ai metadati semantici estratti dai materiali sorgente.

I principali casi d'uso individuati sono:

- **Realizzazione del condition report a supporto del prestito**, con produzione semi-automatica della documentazione necessaria ad autorizzare la movimentazione di un bene museale;
- **Gestione delle versioni e delle epoche del condition report**, con la possibilità per l'operatore di gestire più versioni del report sullo stesso bene e di confrontare stato di conservazione e interventi nel tempo;
- **Integrazione della documentazione eterogenea**, con unificazione in un dataset strutturato dei dati catalografici, della documentazione archivistica e conservativa e degli esiti delle analisi diagnostiche;

- **Riutilizzo dei dataset da parte di enti terzi**, attraverso la pubblicazione sul catalogo DPaaS e l'accesso programmatico all'API REST del Data Product.

2. Origine dei dati ed eventuali elaborazioni

Il Data Product viene realizzato a partire dalla documentazione già custodita dal Museo delle Civiltà e dai dati gestiti nella banca dati dell'istituto. I dati sorgente comprendono la documentazione archivistica, curatoriale e conservativa relativa al bene, i dati catalografici con le informazioni sul restauro e sulle analisi diagnostiche, i condition report storici già associati ai beni e il materiale fotografico.

La diversità delle fonti costituisce un valore aggiunto per il Data Product, poiché permette di raccogliere e mettere in relazione informazioni complementari sul medesimo bene culturale. È proprio l'integrazione tra dati catalografici e documentazione archivistica e conservativa a restituire una visione completa e strutturata del bene, altrimenti dispersa tra archivi, relazioni e campagne di documentazione differenti.

I dati in ingresso sono disponibili nei formati JPEG e PNG per le immagini - con acquisizioni anche in formato TIFF - e nei formati PDF e TXT per la documentazione testuale, mentre i dati e i metadati sorgente sono resi disponibili in formato strutturato. I materiali vengono acquisiti all'interno del Data Product mediante caricamento sullo storage a oggetti S3, che costituisce il punto di ingresso dei dati nella pipeline e dal quale vengono resi disponibili alle successive attività di elaborazione.

Il caricamento della documentazione e la configurazione iniziale dell'applicazione sono di competenza del Data Producer. Prima dell'avvio delle elaborazioni i materiali sono sottoposti a una verifica del formato, primo dei controlli di qualità previsti sul dato.

A partire dai dati caricati sullo storage S3 hanno quindi inizio le attività svolte all'interno del DP Lab, che comprendono le operazioni di trasformazione e arricchimento previste dalla pipeline del Data Product.

Il Dataset di partenza è costituito da:

- circa 250 schede oggetto
- circa 20 indagini diagnostiche
- 10 Mostre
- 2400 elementi di archivio fotografico
- 50 GB di archivio documentale + immagini

3. Servizi abilitanti DPaaS

Nell'ambito delle attività di elaborazione e arricchimento dei dati il Data Product utilizza i servizi CPA (Content Processing Advanced) disponibili nel catalogo della piattaforma DPaaS. Questi mettono a disposizione le funzionalità necessarie per eseguire le elaborazioni basate sull'intelligenza artificiale e vengono utilizzati in funzione della tipologia di dato da trattare e dell'operazione da effettuare.

Dal punto di vista dell'integrazione con gli altri componenti, i servizi CPA intervengono nella fase di trasformazione dei dati. I materiali sorgente vengono inizialmente caricati sullo storage S3 e, a partire da questo, la pipeline preleva i contenuti necessari e li sottopone alle elaborazioni previste attraverso il servizio CPA più adatto; i risultati prodotti vengono successivamente riscritti su S3, andando a costituire il dataset elaborato.

Il servizio CPA utilizzato per l'arricchimento del Data Product è (o può essere):

- **Qwen**, modello linguistico multimodale impiegato per la Retrieval-Augmented Generation. Il servizio opera sui contenuti recuperati dal corpus documentale del museo e genera i contenuti testuali che compongono il condition report, sostenendo il retrieval semantico e il sistema di question answering previsti dalla pipeline.

Il risultato dell'elaborazione contribuisce alla costruzione e all'arricchimento del dataset, memorizzato sullo storage S3 e costituito dai testi strutturati e categorizzati e dai metadati semantici prodotti nel corso del processo. Il dataset così ottenuto può essere successivamente gestito ed esposto attraverso l'API REST del Data Product.

4. Pipeline di elaborazione e trasformazione

La pipeline si avvia a partire dai dati di input: un pacchetto di documentazione relativo ai beni museali, composto da dati catalografici con le informazioni sul restauro e sulle analisi diagnostiche, documentazione archivistica e conservativa, condition report storici e immagini.

L'esecuzione è ad avvio manuale da parte di un operatore. In caso di errore è prevista la ripresa della sola fase di pipeline interessata verso il relativo servizio CPA, evitando di rielaborare gli output già prodotti e validati.

Il processo è articolato nelle quattro fasi descritte di seguito.

1. Import

Acquisizione dei dati catalografici con le informazioni sul restauro e sulle analisi diagnostiche, della documentazione archivistica, curatoriale e conservativa e del materiale fotografico, a partire da dati esterni o da un DP preesistente. I dati sorgente vengono caricati sullo storage S3, che costituisce il punto di ingresso dei materiali nella pipeline.

2. Ingestion

I dati sorgente vengono integrati all'interno della piattaforma di gestione dei beni museali.

3. Generazione dei contenuti

I dati elaborati con l'ausilio dei servizi CPA producono dati e metadati pronti per essere pubblicati all'interno del DPLab. I contenuti sono prodotti tramite il servizio CPA Qwen, impiegato in modalità Retrieval-Augmented Generation: il modello opera sui contenuti recuperati dal corpus documentale e genera le descrizioni testuali che compongono il condition report.

4. Pubblicazione

Gli output prodotti vengono pubblicati dal data producer sullo storage S3 come dataset elaborato e resi disponibili tramite l'API REST di GolangAPI. I metadati sono persistiti su PostgreSQL, a supporto della tracciabilità e del controllo di qualità sull'intero processo.

5. Esposizione e distribuzione

Il Data Product è classificato come Public ed è fruibile come API REST o, opzionalmente, come download diretto. È distribuito con licenza Creative Commons CC BY-NC-SA 4.0 e censito nel catalogo della piattaforma DPaaS, dove è accessibile ai professionisti, ai ricercatori e agli enti terzi autorizzati che intendano consultare, riutilizzare o integrare nei propri flussi di lavoro i dataset e i metadati generati.

Gli output delle elaborazioni sono conservati sullo storage a oggetti S3 dedicato al dataset elaborato, mentre il web service GolangAPI ne supporta l'esposizione programmatica attraverso l'API REST. Il dataset comprende la documentazione elaborata sotto forma di testi strutturati e categorizzati nei formati PDF e JSON e i metadati semantici collegati.

Fruizione tramite API

L'accesso al Data Product avviene tramite chiamate REST API. Questa modalità di fruizione è rivolta ad applicazioni e sistemi client che necessitano di consultare e utilizzare in maniera strutturata le informazioni disponibili, consentendone l'integrazione all'interno di servizi e flussi applicativi esterni.

Tutti gli endpoint sono protetti da una dipendenza globale di verifica della chiave API, con la sola eccezione di GET /health, che espone lo stato del servizio. L'API è esposta esclusivamente su canali cifrati (TLS). I gruppi di risorse e le operazioni esposte sono riportati nella tabella seguente.

Risorsa	Metodo	Descrizione	Autenticazione
/health	GET	Stato del servizio	Non richiesta
/cr/objects	GET	Lista paginata degli oggetti catalogati	Chiave API
/cr/objects/{ID OBJECT}	GET	Metadati di un singolo oggetto a catalogo	Chiave API
/cr/objects/{ID OBJECT}/reports	GET	Lista paginata dei condition report di un oggetto a catalogo	Chiave API
/cr/reports/{ID REPORT}	GET	Metadati di un singolo condition report	Chiave API

Tabella 3 Risorse esposte dall'API REST del Data Product

Struttura del dataset

Il servizio si appoggia a un database relazionale PostgreSQL. Dal punto di vista logico le informazioni sono organizzate attraverso entità che consentono di descrivere e mettere in relazione le diverse risorse del Data Product:

- **Reperto**: Oggetti con identificativo, codici inventariali, tipologia, cronologia, materiali, provenienza e stato di conservazione.
- **Exhibiti**ons/Mo**vimentazioni**, ovvero Metadati delle mostre, prestiti e movimentazioni collegate ai reperti e che hanno determinato la generazione di condition report;
- **Image**, gli asset immagine con i relativi metadati e le relazioni verso i beni;
- **Condition report**, ovvero versioni ed epoche del report associate al bene, confrontabili nel tempo;

Le relazioni tra queste entità permettono di collegare a ciascun reperto le descrizioni prodotte nel tempo, le immagini e la documentazione che ne testimonia lo stato, oltre agli interventi di restauro e agli esiti delle analisi diagnostiche che lo hanno riguardato.

Limitazioni

I limiti di utilizzo del Data Product sono principalmente riconducibili alla qualità e alla completezza della documentazione disponibile a monte. Il report generato può restituire soltanto ciò che la documentazione esistente attesta: per i beni la cui storia conservativa è documentata in modo frammentario l'output risulterà necessariamente parziale.

Va inoltre considerato che eventuali contenuti generati dal modello linguistico costituiscono materiale di lavoro sottoposto a revisione (man-in-the-loop) e non un documento definitivo: la produzione del condition report e la responsabilità del giudizio conservativo resta in capo agli esperti di dominio, che creano e approvano il report tramite l'applicazione di gestione.

5.1. Nota sulle condivisioni dei dati

La sicurezza degli accessi è presidiata su tre livelli: le credenziali dello storage S3 sono riservate e dotate dei soli privilegi minimi necessari sui bucket sorgente ed elaborato; la stringa di connessione al database è mantenuta riservata e l'utenza applicativa dispone esclusivamente dei privilegi necessari al proprio funzionamento; l'API è esposta unicamente su canali cifrati e protetta dalla verifica della chiave API. I segreti - chiave

API, credenziali S3 e stringa di connessione al database sono gestiti tramite il sistema di gestione dei segreti dell'ambiente.

6. Possibili evoluzioni future

Il Data Product è stato progettato in modo da consentire successive evoluzioni e integrazioni, sia rispetto alle tipologie di dati e contenuti gestiti sia rispetto alle tecnologie utilizzate nelle diverse fasi di elaborazione. La separazione tra pipeline di elaborazione, storage e layer di esposizione consente l'estensione a nuovi beni e a ulteriori sorgenti documentali senza modifiche strutturali al sistema.

Le evoluzioni - introduzione di nuovi servizi CPA, nuovi endpoint dell'API, cambi di formato - seguono il ciclo test, validazione e produzione concordato con il team DPaaS. Ogni rilascio del Data Product è identificato da versione e data, e le modifiche allo schema dei dati e all'API sono tracciate e comunicate al team che ospita ed esercisce il Data Product.

Una prima direttrice di sviluppo riguarda l'ampliamento del corpus trattato e l'estensione del modello ad altre istituzioni culturali: la struttura del condition report e la logica della pipeline non sono specifiche del Museo delle Civiltà e possono essere adottate da musei, soprintendenze e centri di restauro che affrontano il medesimo problema documentale, con il solo adattamento delle sorgenti e dei vocabolari di dominio.

Una seconda direttrice riguarda l'evoluzione delle modalità di esecuzione, oggi ad avvio manuale, verso forme di elaborazione programmata, e l'estensione dei controlli di qualità e del monitoraggio già previsti sulla disponibilità del servizio, sugli errori dei servizi CPA, sui tentativi di accesso non autenticati e sugli errori di accesso a storage e database.

Le specifiche direttrici di evoluzione saranno definite sulla base delle esigenze che emergeranno durante l'utilizzo del Data Product e degli eventuali sviluppi futuri della piattaforma e dei servizi a esso collegati.