31 luglio 2026

Finanziato dall'Unione europea NextGenerationEU

MINISTERO DELLA CULTURA

DIGITAL LIBRARY
Istituto centrale per la digitalizzazione del patrimonio culturale

Ecomic ecosistema digitale per la cultura

DPaaS Piattaforma data product as a service

PAPYRUS

Knowledge base semantica per esplorazione delle tecniche di decifrazione dei papiri ercolanesi

Nel 79 d.C., l'eruzione del Vesuvio seppellì la Villa dei Papiri di Ercolano, preservando una straordinaria biblioteca di rotoli carbonizzati. Resi fragili e illeggibili dal calore, questi reperti sono da oltre due secoli al centro di tentativi di apertura e decifrazione

Interrogare gli studi sui papiri in linguaggio naturale

Studi e documenti descrivono tecniche di decifrazione, imaging e analisi dei papiri carbonizzati della Villa dei Papiri.

La ricerca per parole chiave non sempre intercetta concetti espressi con termini diversi.

Il Data Product crea un indice semantico della documentazione e restituisce i passaggi più pertinenti a una domanda.

Può alimentare chatbot, assistenti RAG e strumenti di ricerca specialistica.

Il documento viene trasformato in frammenti ricercabili



Documento

Il file originale costituisce l'unità di partenza della base di conoscenza. Il sistema acquisisce documenti in formato TXT o DOCX dallo storage S3 e ne estrae il testo insieme ai principali metadati descrittivi: titolo, autore, provenienza, tipologia documentale e data di acquisizione. A ogni documento viene assegnato un identificativo univoco, che consente di mantenere il collegamento tra le informazioni restituite dalla ricerca e la relativa fonte. Prima dell'elaborazione, il sistema verifica inoltre che il documento non sia già stato acquisito, evitando duplicazioni all'interno del corpus.



Chunk semantici

Per ottimizzare la ricercabilità del documento, il testo viene suddiviso in frammenti più piccoli, chiamati chunk. Ogni chunk contiene circa 800 caratteri e mantiene una sovrapposizione di 120 caratteri con quello successivo, così da non interrompere bruscamente frasi, concetti o passaggi argomentativi. La suddivisione viene effettuata seguendo la struttura del testo, privilegiando paragrafi, frasi e separatori naturali. Ogni frammento conserva il riferimento al documento di origine e la propria posizione al suo interno. La ricerca può quindi restituire direttamente il passaggio pertinente, senza obbligare l'utente a consultare l'intero documento.



Vettori + ranking

Ogni chunk viene convertito in un vettore semantico a 768 dimensioni attraverso il modello di embedding nomic-embed-text. Il vettore rappresenta numericamente il significato del frammento e viene archiviato in PostgreSQL con l'estensione pgvector. Quando l'utente formula una domanda, anche la query viene trasformata in un vettore. Il sistema confronta quindi la domanda con i frammenti indicizzati, utilizzando la similarità coseno e un indice HNSW per individuare rapidamente quelli semanticamente più vicini. I risultati vengono ordinati attraverso un punteggio di pertinenza compreso tra 0 e 1: più il valore è elevato, maggiore è la corrispondenza semantica con la domanda. Per ogni risultato vengono restituiti il testo del chunk, il documento di provenienza, il punteggio e i metadati associati. La ricerca può inoltre essere filtrata per fonte o tipologia documentale.

Dalla fonte alla risposta contestualizzata

1

INPUT

01 · Ingestion
Nuovi file TXT o DOCX vengono acquisiti dallo storage e controllati.

2

ANALISI

02 · Indicizzazione
Testo, chunk ed embedding sono archiviati nel database vettoriale.

3

OUTPUT

03 · Ricerca
L'API confronta la domanda con i frammenti e restituisce i più pertinenti.

Un motore senza interfaccia, pronto per molteplici applicazioni

Il Data Product è un microservizio privo di una propria interfaccia grafica. È progettato come una capacità tecnologica riutilizzabile, accessibile tramite API REST e integrabile in chatbot, assistenti virtuali, sistemi RAG, motori di ricerca e applicazioni culturali. La separazione tra motore e interfaccia consente a ciascun ente di progettare modalità di fruizione diverse, mantenendo invariata la base documentale e il sistema di ricerca.



FastAPI – accesso, ricerca e catalogo

Il livello API, sviluppato con FastAPI e Uvicorn, mette a disposizione le funzionalità del Data Product attraverso richieste HTTP standard e risposte in formato JSON. L'endpoint /search riceve le domande in linguaggio naturale e restituisce i frammenti più pertinenti; /documents/list permette di consultare il catalogo dei documenti indicizzati; /health consente di verificare lo stato del servizio. La documentazione interattiva Swagger permette inoltre di esplorare e testare le API direttamente dal browser.



Embedding e database vettoriale – il cuore della ricerca semantica

Il testo dei documenti viene suddiviso in chunk e trasformato in vettori semantici a 768 dimensioni tramite il modello nomic-embed-text, reso disponibile attraverso KubeAI. Documenti, frammenti e vettori vengono conservati in PostgreSQL con l'estensione pgvector. L'indice HNSW consente di confrontare rapidamente il significato della domanda con quello dei contenuti indicizzati, ordinando i risultati secondo la similarità semantica. La ricerca può essere ulteriormente circoscritta attraverso filtri per fonte o tipologia documentale.



Applicazioni esterne – chatbot, RAG e strumenti di ricerca

Il Data Product restituisce evidenze documentali strutturate: testo del frammento, documento di provenienza, metadati e punteggio di pertinenza. Questi risultati possono essere utilizzati direttamente in uno strumento di ricerca oppure forniti come contesto a un modello linguistico. Il Data Product non genera autonomamente la risposta finale, ma offre alle applicazioni una base informativa verificabile e collegata alle fonti.

Quattro scenari applicativi

01/Ricerca scientifica

Trovare rapidamente passaggi su tecniche e metodologie specifiche.

02/ Chatbot museale

Fornire contesto documentale a risposte rivolte ai visitatori.

03/Nuovi studi

Indicizzare automaticamente documenti aggiunti allo storage.

04/Ricerca filtrata

Limitare i risultati per fonte o tipologia documentale.

Riusabilità

Una capacità trasversale per rendere interrogabili patrimoni documentali specialistici

Il Data Product può essere applicato a studi, cataloghi, pubblicazioni, relazioni e fondi documentali di musei, archivi, biblioteche e centri di ricerca. Il motore organizza il corpus e recupera, attraverso domande in linguaggio naturale, i passaggi più pertinenti mantenendo il collegamento con le fonti originali.

Condizioni per il riuso: L'ente deve selezionare documenti testuali in formato TXT o DOCX, definire metadati, fonti, categorie e regole di accesso. Il Data Product recupera evidenze documentali verificabili; l'applicazione esterna gestisce l'interfaccia e l'eventuale generazione della risposta.

Valorizzare il patrimonio

- Rendere accessibili raccolte documentali estese o specialistiche.
- Ridurre i tempi di ricerca e analisi bibliografica.
- Riunire fonti precedentemente distribuite.
- Recuperare informazioni per significato, non solo per parole chiave.

Creare servizi per gli utenti

- Motori di ricerca semantica per studiosi e curatori.
- Chatbot museali basati sulle fonti dell'ente.
- Assistenti alla ricerca per operatori e ricercatori.
- Portali di approfondimento per studenti e visitatori.

Integrare nuove applicazioni

- API REST e risposte JSON.
- Filtri per fonte e tipologia documentale.
- Ranking dei frammenti per pertinenza.
- Integrazione in chatbot, portali e sistemi RAG.

I tre obiettivi

Il Data Product trasforma un corpus specialistico in una base di conoscenza capace di collegare le domande degli utenti alle fonti scientifiche più pertinenti.

CONSERVARE

Organizza studi, pubblicazioni e documenti mantenendo il collegamento con le fonti originali.

COMPRENDERE

Supera la ricerca per parole chiave e individua i contenuti rilevanti sulla base del loro significato.

RIUTILIZZARE

Alimenta motori di ricerca, chatbot, assistenti RAG e nuovi servizi digitali per studiosi, operatori e pubblico.

Un patrimonio è pienamente accessibile solo quando può essere interrogato, compreso e riutilizzato.