# Clio

Dalla digitalizzazione alla conoscenza:
AI, metadata, validazione umana
e reinforcement con crowdsource.

**Matteo Osio** Data Engineer
**Lais Kantor** Designer
**Elisa Faggio** Project Manager culturale

## Dati Estratti e Tracciamento

Engineering del Prompt

Il fulcro del progetto è l'**ottimizzazione del prompt** fornito al modello AI.
Attraverso un lavoro iterativo, abbiamo guidato il modello verso l'estrazione di dati specifici e strutturati, partendo dalle schede di catalogo ed in conformità con le norme catalografiche Manus, aumentando significativamente l'**accuratezza nell'identificazione dei metadati.**

### Obiettivi

- Restringere il campo d'azione verso dati catalogabili
- Garantire conformità allo standard di catalogazione MANUS
- Ottenere il miglior grado di accuratezza dei dati possibile

### Strategie

- Creazione di categorie specifiche per ogni campo
- Esempi concreti e istruzioni dettagliate

### Risultato

- Estrazione consistente dei metadati
- Output direttamente integrabile nel database

```
        prompt = f"""Sei un paleografo esperto. Cataloga questo manoscritto secondo standard MANUS/ICCU.

DOCUMENTO: {doc_context}
{age_hint}
{location_hint}

ANALIZZA IL MANOSCRITTO E COMPILA TUTTI I CAMPI:

Se nessun {doc_context} è identificabile, cerca:
- Autore (se menzionato o deducibile da firme presenti)
- Titolo (se menzionato o deducibile da intestazioni)

1. DESCRIZIONE: Descrivi il contenuto del documento (cosa tratta, argomento principale, soggetti, luoghi geografici e periodi temporali). 3-4 frasi dettagliate.

2. LINGUA: Identifica la lingua principale (latino, italiano volgare, italiano moderno, altro).

3. SCRITTURA: Identifica il tipo paleografico tra: capitale elegante, capitale rustica, capitale corsiva, corsiva nuova, onciale, semi onciale, beneventana, gotica, cancelleresca,
mercantesca, umanistica, cancelleresca italica, bastarda italica.

4. SUPPORTO: cartaceo / membranaceo / misto / altro supporto

5. COMPOSIZIONE: unitario / composito / frammento / foglio sciolto

6. DIMENSIONE: Stima dimensioni in mm (altezza x larghezza)

7. LUOGO DI ORIGINE: Identifica provenienza geografica dal testo o deduci dallo stile.

8. DATAZIONE: Periodo in formato YYYY-YYYY o secolo. Se stimata aggiungi "(stimata)".

9. ELEMENTI STORICI: Segnature, timbri, etichette di archivi/biblioteche visibili.

10. INCIPIT (OBBLIGATORIO): Trascrivi ESATTAMENTE le prime 5-10 parole leggibili del testo principale. Se non leggibile scrivi "n/d".

11. EXPLICIT (OBBLIGATORIO): Trascrivi ESATTAMENTE le ultime 5-10 parole leggibili del testo principale. Se non leggibile scrivi "n/d".

12. DECORAZIONI: Descrivi iniziali ornate, miniature, disegni, uso di oro. Se assenti: "Nessuna decorazione".

13. CONSERVAZIONE: Valuta leggibilità (ottima/buona/parziale/difficile), danni visibili, stato supporto.

14. TIPOLOGIA: letterario/documentario/epistolare/giuridico/liturgico/amministrativo/scientifico

REGOLE OUTPUT:
- TITOLO e AUTORE non identificabili = "DA IDENTIFICARE" (in maiuscolo)
- TITOLO e AUTORE identificati = scrivi in MAIUSCOLO
- Campi non determinabili = "n/d"
- INCIPIT e EXPLICIT sono OBBLIGATORI: trascrivi il testo o scrivi "n/d"
- Il campo "features" è una STRINGA unica con valori separati da " | "

Rispondi SOLO con questo JSON (nessun commento, nessun testo extra):
{{
"description": "descrizione dettagliata del contenuto",
"tags": "Autore: NOME o DA IDENTIFICARE, Luogo origine: luogo o n/d, Datazione: periodo, Tipologia: tipo",
"features": "lingua: valore | scrittura: valore | supporto: valore | composizione: valore | dimensione: valore | luogo_dettagliato: valore |
datazione_dettagliata: valore | elementi_storici: valore o n/d | incipit: TESTO TRASCRITTO o n/d | explicit: TESTO TRASCRITTO o n/d | decorazioni: valore | conservazione: valore | note: valore"
}}"""
```

# Dati del Modello AI

## I tre tipi di metadati

### Descrizione del contenuto

- **Descrizione e sintesi** del testo del manoscritto
- **Identificazione:** luoghi, persone e argomento trattato
- **Tipologia:** giuridico, letterario, religioso, amministrativo
- **Natura della descrizione:** descrizione da copia digitale

### Caratteristiche generali del documento (Tags)

- **Autore:** nome dell'autore - se identificabile
- **Luogo di origine:** città/regione di produzione
- **Datazione:** periodo storico - es. XIV secolo

### Caratteristiche dettagliate del documento

- **Scrittura:** tipologia - gotica, umanistica, mercantesca
- **Materiale del supporto**: cartaceo, membranaceo
- **Incipit/explicit:** prime e ultime parole
- **Decorazioni:** elementi decorativi
- **Conservazione:** stato del documento

+Bonus

### Tracciamento del processo

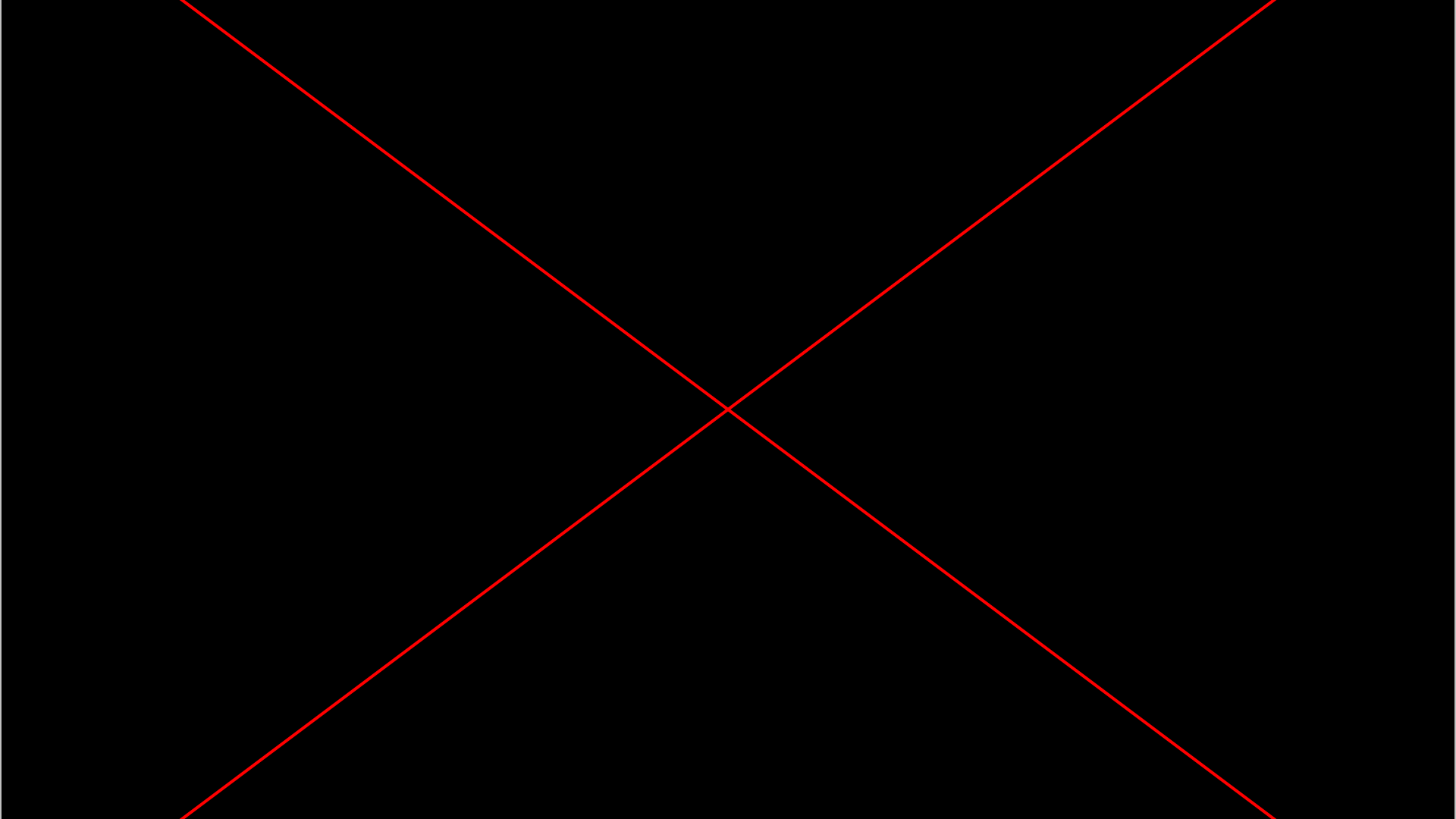
- Utente che ha caricato il manoscritto
- Utente che ha effettuato l'ultima modifica
- Timestamp di caricamento e ultima modifica

## Opportunità nella Catalogazione

La forza del nostro approccio risiede nella flessibilità del prompt: **modificando le istruzioni, lo stesso modello AI può catalogare molteplici collezioni,** applicando la stessa infrastruttura tecnologica a documenti, raccolte librarie e archivi eterogenei.

### Possibili applicazioni

- **Giornali, carteggi, lettere:** estrazione automatica di autore, periodo, titolo e contenuto
- **Mappe geografiche:** riconoscimento di toponimi, scale, legende cartografiche e annotazioni territoriali
- **Fotografie storiche:** estrazione automatica di periodo, soggetti ritratti e tecniche fotografiche utilizzate
- **Documenti in altre lingue:** estrazione automatica di dati da documenti in latino, greco, arabo etc.



## Il Modello Attuale

Qwen2.5-VL-3B

Attualmente utilizziamo Qwen2.5-VL-3B, un vision-language model compatto da 3 miliardi di parametri, ottimizzato per l'analisi di immagini e testo. **Il modello è in grado di comprendere documenti storici e estrarre informazioni strutturate**, ma opera su CPU con risorse limitate.

### Caratteristiche Attuali

- **3 miliardi di parametri:** Modello compatto ma capace
- **Inferenza su CPU:** Tempo di elaborazione ~1-2 minuti per immagine (single user)
- **Multimodale:** Comprende sia testo che immagini (persino video potrebbe!)
- **Open source:** Qwen2.5-VL (Alibaba Cloud)

### Limitazioni tecnologiche

- Velocità di elaborazione limitata dalla CPU
- Elaborazione sequenziale (un'immagine alla volta)
- Capacità di output ridotta (max 400 token)
- Processata solo un'immagine e solo un utente attivo inoltre la trascrizione non è stata generata, ma sostituita con una breve descrizione - prediletto l'utilizzo di token per completare la catalogazione

## Prospettive di Miglioramento

Qwen2.5-VL-3B

Tutto questo potenziale è limitato dai constraint tecnologici attuali.
Con un po' di **addestramento aggiuntivo** (fine-tuning o reinforcement learning) fatto sui dati corretti, in aggiunta a una macchina più potente con **GPU dedicata**, sarebbe possibile **ottimizzare drasticamente le capacità del modello** nell'eseguire questi task specifici di catalogazione paleografica.

### Ottimizzazioni Possibili

- **Fine-tuning:** Addestramento su dataset di manoscritti italiani correttamente catalogati
- **Reinforcement learning** con feedback di esperti paleografi
- **GPU acceleration:** Riduzione del tempo da minuti a secondi
- **Modelli più grandi:** Upgrade a 7B o 72B parametri per maggiore accuratezza

### Risultati attesi

**15-30x più veloce** con GPU (da 1-2 min a 5-10 sec)

**Maggiore accuratezza** nell'estrazione dei metadati

**Batch processing** catalogazione simultanea di multiple immagini

## Human-in-the-Loop tramite Crowdsourcing

Un sistema di **crowdsourcing CAPTCHA** dove i cittadini, in modo gamificato, contribuiscono attivamente all'addestramento del modello creando un vero human-in-the-loop per la preservazione del patrimonio culturale.

Link: https://deaf-fair-17174760.figma.site/

### Vantaggi

- Training data gratuito e scalabile
- Identificazione automatica dei casi difficili
- Coinvolgimento civico attivo

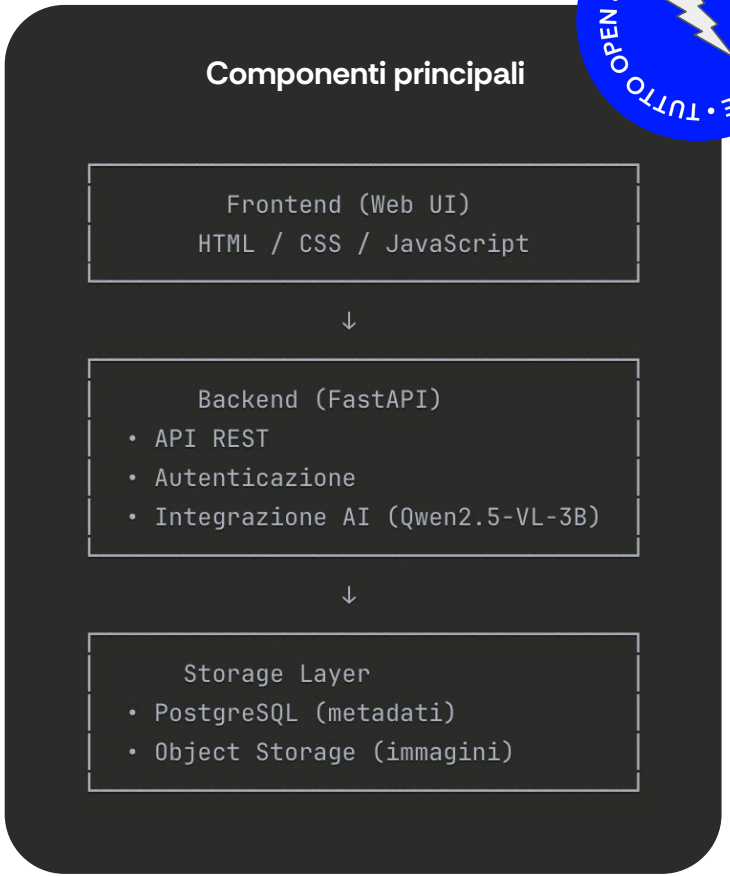
### Potenziali blocker

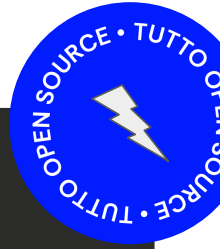
- Cold-start problem del modello iniziale
- Difficoltà di contesto dei caratteri paleografici
- Quality control delle risposte

## Architettura del Sistema

Il sistema è costruito su un'architettura moderna e modulare che separa chiaramente le responsabilità:

- **frontend** web intuitivo per l'interazione utente
- **backend** per la logica applicativa e l'integrazione con il modello AI
- **layer di storage** dual-level che gestisce sia i metadati strutturati che le immagini dei manoscritti.



## Architettura del Sistema

### Frontend
Interfaccia Web Intuitiva

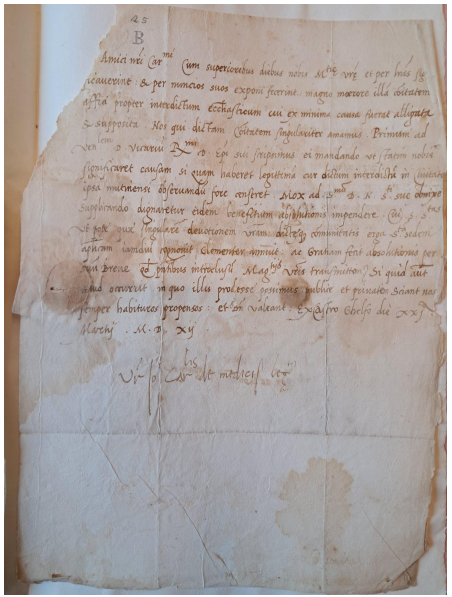
- HTML/CSS/JavaScript
- Upload drag-and-drop per le immagini
- Tabella interattiva con sorting e ricerca
- Form di editing per correzioni manuali
- Visualizzazione metadati estratti dall'AI
- Responsive design per tablet e desktop

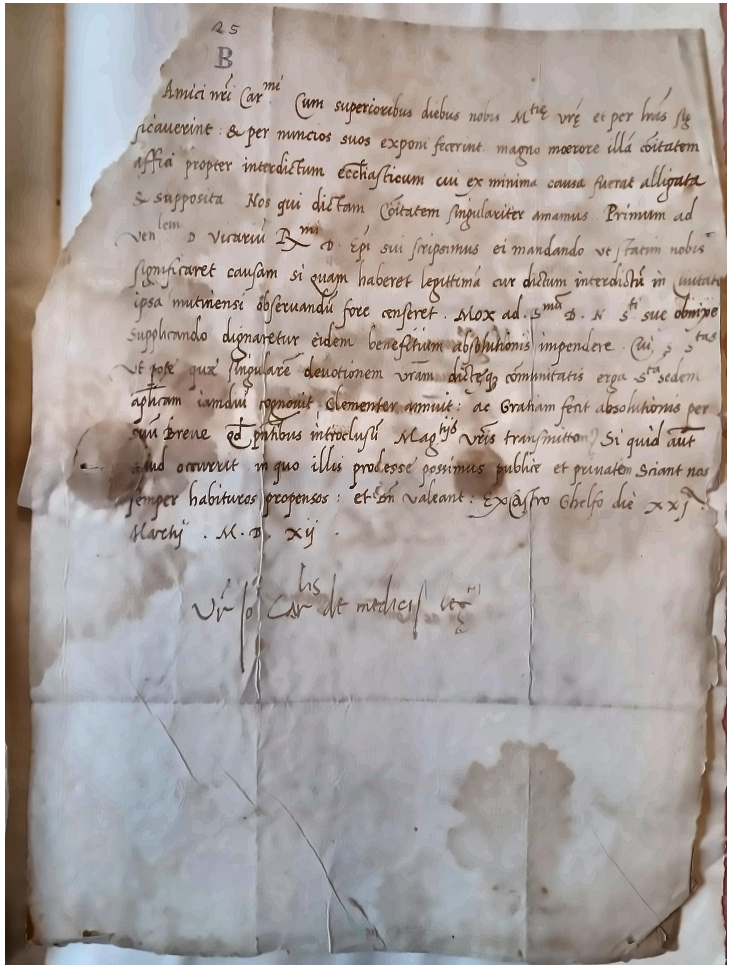

### Backend
FastAPI + AI Integration

- FastAPI framework (Python)
- RESTful API per CRUD operations
- Autenticazione
- Integrazione Qwen2.5-VL-3B per catalogazione
- Image preprocessing pipeline (deskew, enhance)
- Gestione upload e validazione file

**Image pre-processing**

Ottimizzazione dell'immagine per il riconoscimento facilitato





## Storage Layer

Il layer di storage è progettato per gestire sia dati strutturati che file multimediali. Attualmente utilizziamo PostgreSQL per i metadati e un file system locale per le immagini, ma l'**architettura è pronta per scalare** verso soluzioni più robuste e performanti.

**Storage attuale**

PostgreSQL

- Metadati strutturati (autore, data, luogo, features)
- User tracking (uploaded_by, edited_by, timestamps)
- Relazioni e query complesse
- Plugin addizionali (pgvector)

File System Locale

- Immagini originali e preprocessate
- Organizzazione per file_id
- Soluzione semplice per MVP

**Sviluppi futuri**

LanceDB (Vector Database)

- Storage di embeddings vettoriali
- Ricerca semantica sui manoscritti
- Similarity search per immagini simili

BLOB Storage (Binary Large Objects)

- Migrazione da file system a object storage
- Gestione efficiente di grandi volumi di immagini

S3 Object storage

- Per possibili integrazioni con tool già esistenti

## Sviluppi Futuri: Integrazioni AI

L’architettura è progettata per integrare modelli AI specializzati che arricchiscono progressivamente il sistema.

Un processo schedulato settimanale può generare embeddings vettoriali dei manoscritti già catalogati, aprendo la strada a retrieval semantico avanzato e knowledge graph per navigare le connessioni tra documenti storici.

**Pipeline di Embedding Automatico**
Weekly Job sul Database

- Processo batch su manoscritti catalogati
- ColQwen-Omni (**vidore/colqwen-omni-v0.1**) per embedding multimodal
- Storage in LanceDB per ricerca vettoriale
- Aggiornamento incrementale dei nuovi documenti

## Evoluzione del Sistema

# AI Agents

- Assistenti intelligenti per ricerca complessa
- Query in linguaggio naturale: "Trova manoscritti fiorentini del XIV secolo con decorazioni miniate"
- Reasoning multi-step su metadati e contenuti

*Funzionano tramite Retrieval-Augmented Generation (RAG) potenziata da open source SLM o LLM.*

# Knowledge Graph

- Relazioni semantiche tra manoscritti (autori, luoghi, temi)
- Navigazione visuale delle connessioni storiche
- Discovery automatico di pattern e collegamenti

**Grazie alla Cosine Similarity**
Non cerchi più per parole chiave,
ma per concetti e stile visivo

## Scalibilità

Il progetto è stato pensato per essere scalabile e sostenibile nel lungo periodo. L'architettura modulare permette di iniziare con infrastruttura minima e crescere in base alle esigenze, bilanciando costi operativi e performance attraverso scelte strategiche tra cloud e on-premise.

Il deployment è gestito tramite Docker e docker-compose per garantire portabilità e facilità di installazione.

### Requisiti hardware

- **MVP Attuale (CPU):** Hardware esistente - €0
- **Phase 2 (GPU):** €240-500/mese cloud o €2,000-2,500 on-premise
- **Phase 3 (Fine-Tuning):** €1-3/ora on-demand o €4,000-5,000 server dedicato

**Scalabilità lineare**
Ogni istanza aggiuntiva aumenta proporzionalmente capacità di accessi simultanei e potere di calcolo, permettendo crescita graduale in base alla domanda effettiva.

### Stima annuale

Stime indicative senza validazione approfondita delle risorse.

- **Cloud:** €3,000-4,000/anno (per uso sporadico)
- **On-Premise:** €2,500/anno primo anno, poi €1,500/anno - richiede skills per mantenere il server (per uso continuo ed intensivo)

**Da valutare** in base alle necessità reali considerando variabili come costi fluttuanti delle macchine e opportunità quali **Cloud EU** o **infrastrutture pubbliche esistenti**.

## Sostenibilità

**Deployment Docker**
Portabilità e facilità di installazione

**Open Data**
Cataloghi accessibili via API REST/CORS per integrazione con servizi esistenti

**Approccio ibrido**
cloud/on-premise

**Software open source**
€0 licenze

**Crowdsourcing per training data**
Coinvolgimento dei cittadini

# Code base

Link: https://codeberg.org/osiom/hack-the-data-culture

```
hack05@hack05:~/clio$ curl -X POST "http://localhost:8000/api/catalog" \ \
  -F "file=@data_test/004_Giovanni_de__Medici._Lettera_B.jpg" \
  -F "title=Lettera di Giovanni de Medici" \
  -F "author=Giovanni de Medici" \
  -F "date=XV secolo" \
  -F "region=Firenze"

{"success":true,"file_id":"20251211_111522_004_Giovanni_de__Medici._Lettera_B.jpg","filename":"004_Giovanni_de__Medici._Lettera_B.jpg","storage_path":"uploads/20251211_111522_004_Giovanni_de__Medici._Lettera_B.jpg","catalog":{"description":"Il documento è una lettera autografa di Giovanni de Medici, scritta in latino, risalente al XV secolo. La lettera si occupa di affari di stato, specificamente riguarda l'approvazione di un documento di proprietà di Giovanni de Medici. Il documento è stato trovato a Firenze.","tags":"Autore: GIOVANNI DE MEDICI, Luogo origine: Firenze, Datazione: XV secolo, Tipologia: letterario/documentario","features":"lingua: latino | scrittura: capitale elegante | supporto: cartaceo | composizione: unitario | dimensione: n/d | luogo_dettagliato: Firenze, Italia | datazione_dettagliata: XV secolo | elementi_storici: n/d | incipit: n/d | explicit: n/d | decorazioni: n/d | conservazione: buona | notes: n/d"},"metadata":{"title":"Lettera di Giovanni de Medici","author":"Giovanni de Medici","date":"XV secolo","region":"Firenze"}}hack05@hack05:~/clio$
```