Malevi - Vera

Ecomic: Hack the Data Culture

Metadati pubblici

12/05/2026

Descrizione breve:

Piattaforma per digitalizzare e curare manoscritti: carichi le pagine (sia una alla volta sia da ZIP con immagini, XML e opzionalmente trascrizioni), le salvi in cloud e le colleghi a un manoscritto. Vera è l'interfaccia per operatori e revisori; il backend Admin espone le API e collega database (PostgreSQL), file (S3), accessi (FusionAuth) e Ollama per trascrivere, estrarre metadati codicologici, epoca/stile/provenienza, tag. Il flusso è human-in-the-loop: l'AI propone, la persona verifica e corregge prima di archiviare. In prospettiva: code e contesto aggregato per gestire volumi molto grandi di pagine senza superare i limiti del modello.

Descrizione Full:

Il progetto è una piattaforma digitale per gestire manoscritti storici (o analoghi): non solo conservare le immagini delle pagine, ma accompagnarle con trascrizioni, metadati descrittivi e strumenti di ricerca e archiviazione. Ha due "facce" che lavorano insieme: da un lato il sito applicativo **Vera** (interfaccia per operatori e revisori), dall'altro il servizio **API** (motore che salva i dati, parla con il database, con lo spazio file e con i modelli di intelligenza artificiale).

Cosa fa l'operatore?
Chi usa **Vera** può creare un manoscritto (un "contenitore" logico con titolo, autore, testi descrittivi) e aggiungere le pagine come file immagine. Le pagine possono arrivare una alla volta oppure in blocco da un archivio ZIP: in quel caso il sistema si aspetta un file XML (spesso in formato TEI) per capire nome e altre informazioni, una cartelle delle Immagini e, se c'è, una cartelle delle Trascrizioni già pronte in testo. Ogni pagina diventa un documento nel sistema, con il suo ordine, la sua immagine salvata in cloud (Amazon S3) e, quando disponibile, il testo trascritto.

Una parte importante del lavoro è assistita da un modello linguistico collegato tramite Ollama: ad esempio si può chiedere di trascrivere automaticamente una pagina guardando l'immagine, oppure di leggere più pagine insieme per estrarre una descrizione fisica e codicologica del manoscritto (materiale, rigatura, legatura, annotazioni, conservazione, ecc.) o per proporre epoca, stile e provenienza. Esistono anche funzioni per provare queste estrazioni su file caricati al volo, senza salvarle subito nel database, e per ottenere tag riassuntivi a partire dai metadati già raccolti.

Il ruolo della persona (human-in-the-loop)
Il manuale insiste sul fatto che **Vera** è pensata per un flusso umano + macchina: il modello propone, ma l'operatore controlla, modifica e conferma. Ogni pagina può essere segnata come trascrizione verificata; il manoscritto può essere marcato come verificato e, quando tutto è a posto, archiviato. L'interfaccia mostra dashboard con numeri riassuntivi (quanto resta da convalidare, tasso di verifica, ecc.) e una coda di lavoro per riprendere i lavori lasciati a metà. I colori e i pulsanti distinguono azioni "di routine" da quelle che richiamano l'intelligenza artificiale.

Come funziona?
Dietro le quinte ci sono un database PostgreSQL descritto con Prisma, accesso protetto con FusionAuth (login con token), file su AWS S3, documentazione delle API con Swagger, e altri servizi di supporto come Weaviate (per ricerche semantiche sul contenuto indicizzato) e Redis con code (ad esempio per l'archiviazione in background). Le estrazioni testuali e visive passano da Ollama con un modello configurabile e un tempo massimo di attesa regolabile.

Dove vuole andare il progetto (idea di evoluzione)
Il manuale descrive anche una direzione futura: oggi analizzare molte pagine in una sola volta ha limiti pratici per il modello; in prospettiva si immagina una coda di lavori che si può riempire all'infinito, che elabora pagina per pagina (o piccoli gruppi) e aggiorna un "contesto del manoscritto" unico — un riassunto sempre più ricco di ciò che si è capito dall'intero volume — così le analisi globali restano gestibili anche su manoscritti enormi, con controlli su costi, errori e ripetizioni.

Use case principali
Analisi evolutiva della lingua italiana.
Addestramento di modelli NLP su testi storici.
Ricerca archivistica, storica e filologica.
Applicazioni di digital humanities.
Creazione di percorsi narrativi digitali basati su fonti manoscritte.

Target
Ricercatori, università, sviluppatori AI, enti culturali, biblioteche, archivi, editoria digitale e soggetti pubblici o privati interessati alla valorizzazione del patrimonio documentale italiano.