UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Centro Interdipartimentale
di Ricerca sulle Digital Humanities
DHMoRe

READ-ME

La lettura dei registri e dei documenti manoscritti storici richiede spesso tempi lunghi e rende difficile trasformare le informazioni d'archivio in dati confrontabili. La tecnologia sviluppata automatizza questo processo combinando intelligenza artificiale per il riconoscimento della scrittura (HTR), individuazione delle aree testuali e controllo da parte dell'operatore. Le informazioni estratte vengono organizzate in tabelle e file strutturati, pronti per analisi quantitative e statistiche.
Lo strumento accelera così l'elaborazione di grandi serie documentarie e amplia le possibilità di ricerca sulla storia sociale ed economica.

"Ottimizzazione di processi di lettura automatica delle fonti"

Laboratorio	DHMoRe
Area di specializzazione	Cultura e creatività, Digitale
Referenti	Matteo Al Kalak, Lorenzo Baraldi
Keyword	HTR, OCR, Storia economica, Fonti seriali, IA



Fig. 1: READ-ME

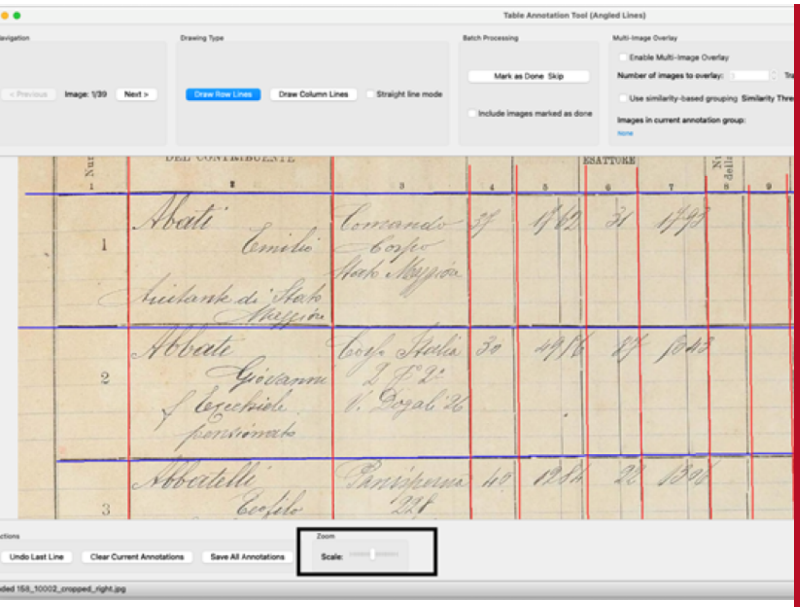


Fig. 2: Prodotto

Descrizione

La tecnologia consiste in uno strumento software per l'estrazione automatizzata di dati strutturati da registri, censimenti e altre fonti documentarie manoscritte organizzate in tabelle o in registrazioni ricorrenti. È progettata per convertire le immagini digitali di documenti storici in insiemi di dati elaborabili, riducendo il lavoro manuale necessario per trascrivere e organizzare le informazioni. Il sistema integra diversi componenti: algoritmi di elaborazione delle immagini per individuare
righe e colonne; modelli di rilevamento delle aree testuali, comprese le singole celle; tecnologie HTR (Handwritten Text Recognition) per il riconoscimento della scrittura manoscritta; procedure di verifica e correzione con intervento umano (human-in-the-loop). I risultati vengono raccolti in dati tabellari ed esportati in formato CSV, utilizzabile con fogli elettronici, database e strumenti di analisi statistica.
La tecnologia è stata sviluppata e sperimentata nel contesto del progetto READ-ME, con particolare attenzione alle fonti utili alla ricostruzione quantitativa della storia sociale ed economica italiana tra Ottocento e Novecento. L'architettura modulare e la possibilità di addestrare i modelli su grafie e strutture documentarie differenti ne rendono potenziale l'applicazione anche ad altre serie archivistiche caratterizzate da registrazioni organizzate in campi e colonne.

Aspetti innovativi

Le tecnologie di riconoscimento automatico dei testi, quali OCR per i documenti a stampa e HTR per quelli manoscritti, sono ampiamente utilizzate per rendere accessibili e ricercabili le fonti digitalizzate. La soluzione qui sviluppata persegue un obiettivo complementare: trasformare le informazioni contenute nei documenti in dati strutturati, utilizzabili per analisi quantitative e statistiche. L'innovazione risiede nell'integrazione tra riconoscimento della scrittura, identificazione automatica delle strutture tabellari e validazione assistita dall'operatore. Tale approccio consente di elaborare documenti nei quali il solo riconoscimento testuale non è sufficiente, poiché l'estrazione dei dati richiede l'individuazione delle celle e delle relazioni tra le informazioni registrate. Il sistema è progettato per elaborare interi complessi documentari e adattarsi a differenti grafie mediante annotazione e addestramento dei modelli. La supervisione umana garantisce il controllo qualitativo dei risultati. La piattaforma rende così le informazioni interrogabili e comparabili su larga scala, aprendo nuove prospettive di ricerca in ambito storico, demografico, economico e sociale.

Applicazioni

Studio delle trasformazioni demografiche, economiche e sociali in prospettiva storica.

RETE ALTA TECNOLOGIA
EMILIA-ROMAGNA
HIGH TECHNOLOGY NETWORK

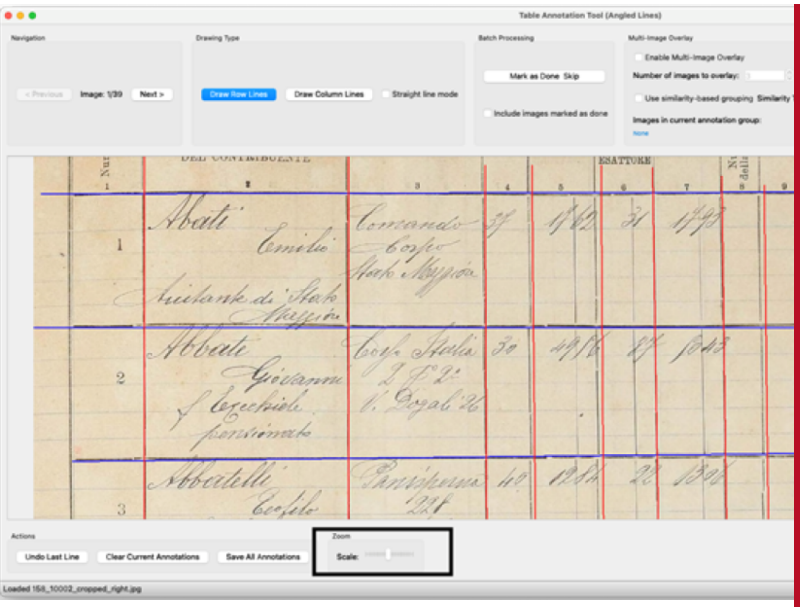


Fig. 3: Applicazione

Esempio di applicazione

Serie dei registri manoscritti dei contribuenti della tassa di famiglia/focatico per il Comune di Roma tra Otto e Novecento

L'applicazione è stata sviluppata a partire dal Focatico di Roma, una fonte manoscritta tabellare, ed estesa tramite una ricognizione di fonti archivistiche italiane per addestrare modelli di riconoscimento automatico. Il processo si articola in una pipeline sequenziale: algoritmi di elaborazione immagini individuano righe e colonne dei registri, un modello di text detection riconosce le aree testuali nelle celle e un sistema HTR (Handwritten Text Recognition) trascrive automaticamente il testo. Entrambi i modelli sono stati addestrati tramite l'annotazione delle fonti storiche.
Il sistema è integrato in un'interfaccia grafica che permette di elaborare interi volumi con un approccio human-in-the-loop, consentendo all'operatore di verificare e correggere le trascrizioni. I dati finali vengono esportati in formato CSV per successive analisi statistiche e gestionali. Il tool include un tutorial dedicato ed è stato sottoposto a test di usabilità. La sperimentazione ha confermato l'efficacia della pipeline nella conversione di registri manoscritti in dati strutturati, mentre la misurazione precisa di velocità e accuratezza richiederà ulteriori test quantitativi.

Partner coinvolti	Partner del progetto READ-ME
Tempi di realizzazione	6
Livello di maturità tecnologica	TRL 8 - sistema completo e validato
Valorizzazione applicazione	Ricerca di partner aziendali per ulteriori applicazioni

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Centro Interdipartimentale
di Ricerca sulle Digital Humanities
DHMoRe

DHMoRe

Centro interdipartimentale di ricerca sulle Digital Humanities dell'Università di Modena e Reggio Emilia



Sito web https://www.dhmore.unimore.it/

Direttore Matteo Al Kalak

Il Centro interdipartimentale di ricerca sulle Digital Humanities dell'Università di Modena e Reggio Emilia (DHMoRe) è promosso dal Dipartimento di Studi linguistici e culturali e dal Dipartimento di Ingegneria "Enzo Ferrari" con l'adesione dei Dipartimenti di Giurisprudenza, di Economia "Marco Biagi", di Educazione e Scienze Umane e di Comunicazione ed Economia.
Ha sede amministrativa presso il Dipartimento di Studi linguistici e culturali e sede operativa nel complesso dell'ex Ospedale di Sant'Agostino, all'interno di Ago Modena Fabbriche Culturali.
Attualmente conta oltre cinquanta aderenti fra docenti e ricercatori dell'Ateneo.
Tra gli obiettivi del Centro, si citano:
• incrementare e coordinare studi e ricerche interdisciplinari nel campo delle Digital Humanities, e in particolare nella creazione, sviluppo e gestione di piattaforme di raccolta del patrimonio storico in forma digitalizzata, nello sviluppo di soluzioni per la digitalizzazione e la comprensione automatica del manoscritto, e lo sviluppo di nuovi modelli economici per la valorizzazione e la fruizione dei contenuti culturali.
• proporsi come interlocutore di enti pubblici e privati ed enti di ricerca per la realizzazione di progetti a livello nazionale e internazionale;
• favorire l'attivazione di impresa culturale e creativa, anche attraverso logiche di piattaforma;
• dialogare con i corsi di laurea e i dottorati di ricerca di Unimore che affrontano temi in linea con gli interessi del Centro.