Trascrivere automaticamente i registri catastali sardi del primo Novecento con *Transkribus*: risultati di un caso di studio

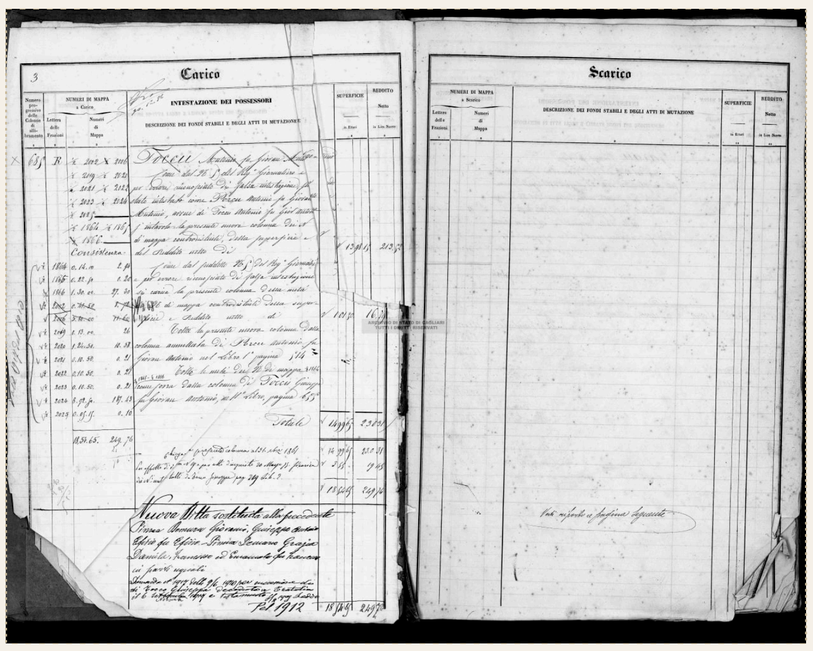
Andrea Pergola · *Cecilia Tasca*
Università degli Studi di Cagliari – Centro Interdipartimentale per l'Umanistica Digitale (DH.UniCA)
andrea.pergola@unica.it | tasca.cecilia@tiscali.it

Introduzione

I sistemi di Handwritten Text Recognition (HTR) sono sempre più impiegati in progetti di analisi e valorizzazione dei patrimoni documentari. Transkribus, sviluppata nell'ambito dei progetti Transcriptorium e READ-COOP, si è affermata come piattaforma di riferimento per la trascrizione automatica di documentazione storica manoscritta, trasformando immagini digitalizzate in risorse interrogabili. Nel 2025 l'ICAR ha adottato Transkribus come standard nazionale per i 102 Archivi di Stato italiani.

Corpus e Obiettivi

La ricerca è stata condotta sui registri catastali del fondo UTE (Ufficio Tecnico Erariale) dell'Archivio di Stato di Cagliari: unità manoscritte di diversi Comuni della Sardegna, caratterizzate da forte eterogeneità di formato, qualità di scrittura e stato di conservazione (pieghe, bruciature, disallineamenti, contrasto). I contenuti sono organizzati in strutture tabellari con dati numerici e testuali su appezzamenti, possessori, localizzazione, colture e stime economiche.



21	1.120	3
unità archivistiche per il training	file immagine elaborati	modelli personalizzati

Metodologia e Workflow

Il flusso di lavoro è stato articolato in più fasi per rispondere alle caratteristiche specifiche del corpus: registri catastali manoscritti, fortemente eterogenei per qualità materiale, organizzazione grafica e fisionomia scrittoria.

1. **Acquisizione digitale**
   Registri in formato TIFF; conversione massiva in PDF e PNG tramite script Python.
2. **Organizzazione per Comune**
   Collezioni distinte per provenienza, per gestire la variabilità interna e aumentare la capacità di generalizzazione dei modelli.
3. **Modello Tabellare**
   Riconoscimento della struttura delle tabelle (righe, colonne, celle). Basato su 'Amsterdam Tables'. Addestramento su 50 → 100 immagini di 10 Comuni.
4. **Modello Layout**
   Individuazione delle linee testuali e delle baseline. Basato su 'Universal Lines'. Riduce i falsi positivi nelle colonne numeriche.
5. **Modello di Trascrizione**
   HTR: il testo manoscritto viene interpretato come flusso orizzontale unitario, superando i limiti dell'OCR. Addestramento progressivo: 10 → 25 → 50 → 80 immagini.
6. **Post-processing**
   Esportazione in Excel, aggregazione, normalizzazione di toponimi, abbreviazioni e valori numerici per ottenere un dataset interrogabile.

Risultati

CER 40–50%

Modello Tabella (Table Model v2)

Miglioramento netto rispetto al modello base 'Amsterdam Tables', che spesso non rilevava righe e aggiungeva colonne fittizie. Il modello personalizzato identifica la struttura tabellare in modo coerente nella maggior parte dei casi, riducendo le correzioni manuali. Il margine di errore elevato è causato dalla variabilità delle condizioni materiali e digitali.

CER ~9,5%

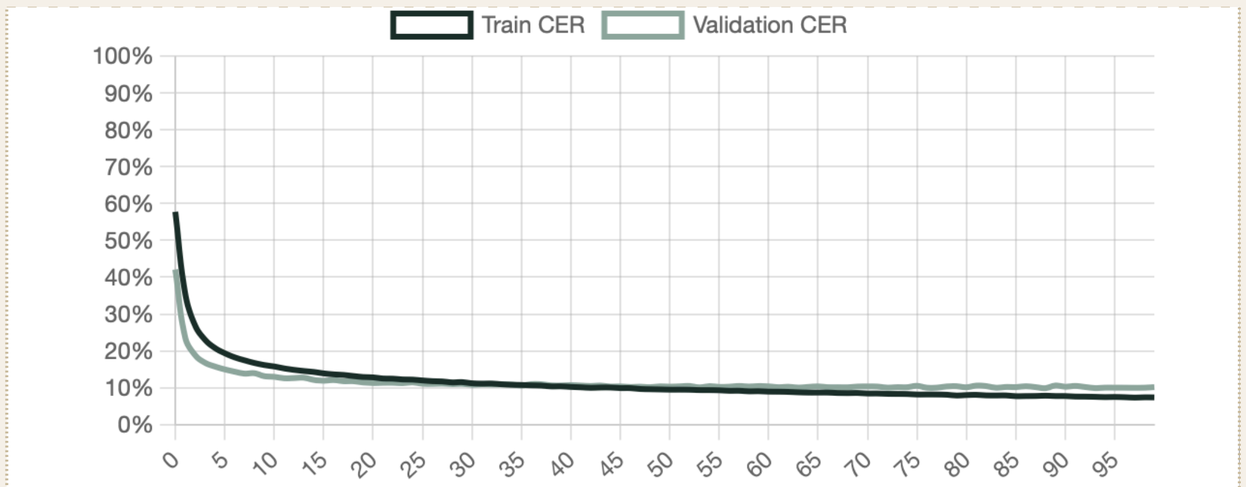
Modello Layout (Baseline model v1)

Basato su 'Universal Lines', ha raggiunto un livello complessivamente buono nel riconoscimento di campi testuali e numerici. Ha eliminato i falsi positivi e migliorato la copertura delle colonne numeriche, con errori residui limitati al mancato riconoscimento di doppie virgolette ripetitive.

CER 8–10%

Modello Trascrizione (v5)

Sviluppato da modelli base con CER teorico 2,38%–6,7%, che alla prima applicazione mostravano CER del 50–60% su questo corpus. Dopo addestramento progressivo, trascrive correttamente intestazioni, titoli di colonna e gran parte dei contenuti. Consente ricerche per parola chiave e analisi quantitative su larga scala.



Conclusioni

La sperimentazione dimostra che anche con un numero limitato di unità è possibile costruire un flusso di lavoro semi-automatico in cui l'HTR riduce consistentemente il carico di lavoro umano, pur non eliminando la necessità di controllo e normalizzazione. Il valore principale dei modelli personalizzati non è la sostituzione del trascrittore umano, ma la trasformazione dei registri in dataset interrogabili e riusabili. La metodologia — dalla definizione del ground truth all'addestramento, fino alla normalizzazione — è riapplicabile ad altre tipologie di registri amministrativi storici ad alta complessità.

Ringraziamenti

Il gruppo di ricerca (coord. Giampaolo Salice e Cecilia Tasca) comprende Bianca Fadda, Mariangela Rapetti, Eleonora Todde, Andrea Pergola, Beatrice Schivo, Stella Barbarossa, Giommaria Carboni e Davide Pisanu. Collaborazione tecnica: Linkalab S.r.L., Cagliari. Finanziamento: Assessorato della Pubblica Istruzione, Beni Culturali, Spettacolo, Sport e Informazione – Regione Autonoma della Sardegna (L.R. 1/2023).

Riferimenti bibliografici (selezione)

Muehlberger et al. (2019). *Journal of Documentation*, 75(5), 954–976.
Nockels et al. (2022). *Archival Science*, 22(3), 367–392.
Romein et al. (2025). *Int. Journal of Digital Humanities*, 7(1), 115–134.
Bazzaco, S. (2024). *DigItalia, 19(1), 63–86.*
Schwarz-Ricci (2022). *Umanistica Digitale,* 13, 171–181.

QR code all'articolo