INTELLIGENZA ARTIFICIALE E BENI CULTURALI
Text Recognition

Federico Boschetti

CNR-ILC

federico.boschetti@ilc.cnr.it

Festival PORTE APERTE AL CNR: #patrimonioculturale nelle transizioni verde e digitale

FIRENZE, 11 OTTOBRE 2023

INTRODUZIONE

Mi presento brevemente

Sono un ricercatore dell'Istituto di Linguistica Computazionale "A. Zampolli" e lavoro presso l'unità di ricerca di Venezia, dove ho l'opportunità di collaborare sia con il Laboratorio di Filologia Collaborativa e Cooperativa (**CoPhiLab**) del mio istituto, sia con il Venice Centre for Digital and Public Humanities (**VeDPH**) dell'Università Ca' Foscari Venezia

I miei interessi di ricerca si concentrano principalmente sui metodi di creazione e di mantenimento a lungo termine delle edizioni scientifiche digitali linguisticamente annotate

Un ponte fra beni culturali materiali e immateriali

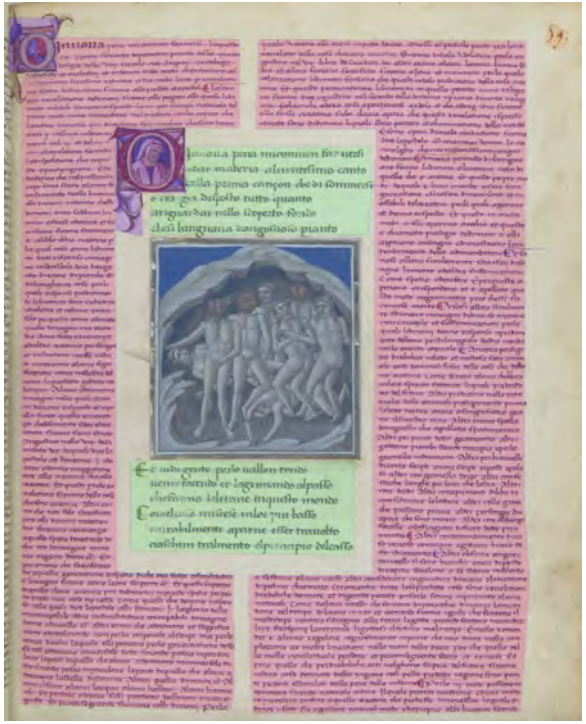
L’Handwritten Text Recognition (HTR) permette di passare dalla rappresentazione digitale di beni materiali come manoscritti e altri text-bearing object alla rappresentazione digitale di beni immateriali come i testi

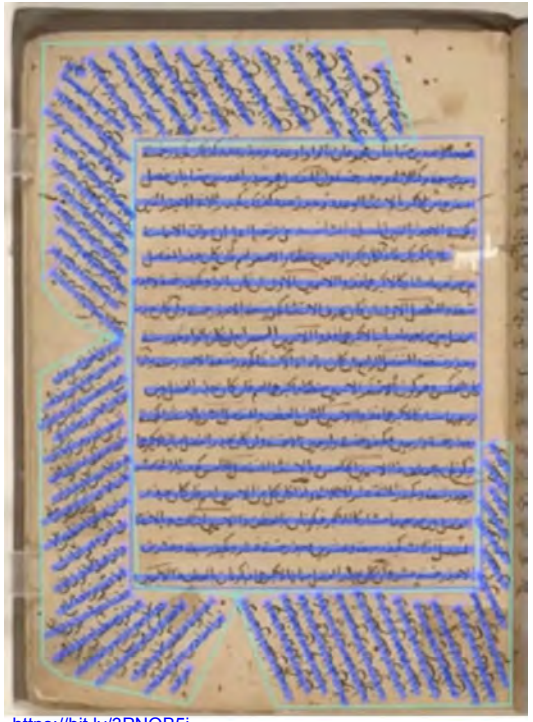
Edizioni scientifiche digitali

- orientate al **documento**
  - facsimilari
  - ultradiplomatiche
  - diplomatiche
- orientate al **testo**
  - critiche
  - genetiche

LAYOUT ANALYSIS

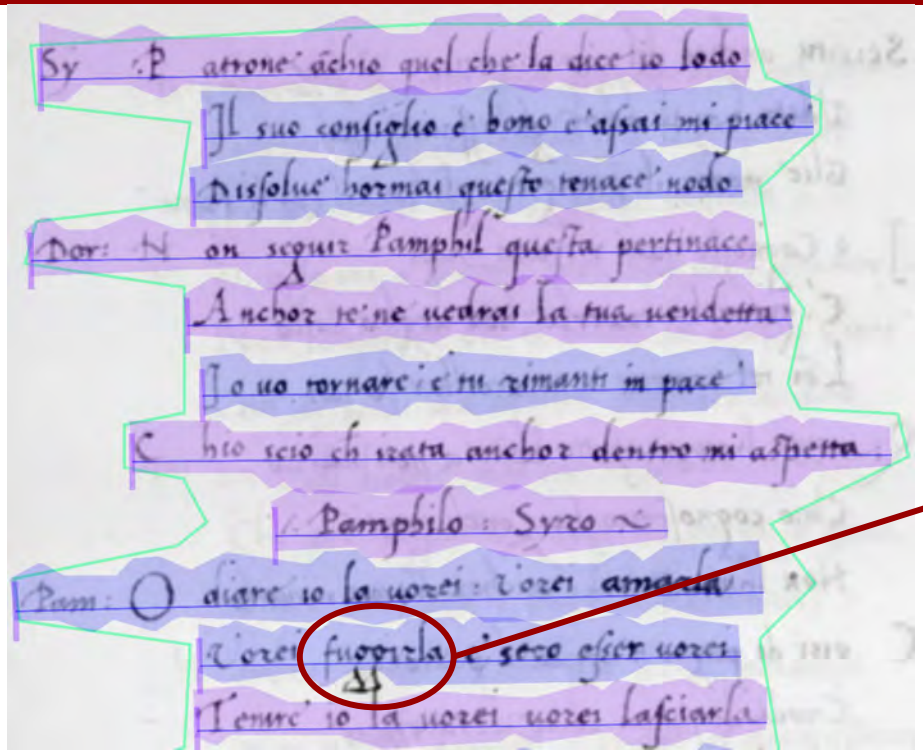
Layout complessi

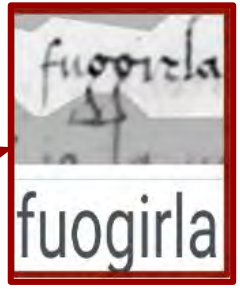




https://bit.ly/3PNOB5i

Segmentazione





Manoscritti epigrafici

Venice, Marciana National Library, ms. Lat. XIV, 200 (4336), f. 2r, nr. 13 = *CIG* 3316

Per ogni iscrizione Astori scrive:

- il luogo in cui l'aveva vista
- il testo epigrafico in maiuscolo

Per le iscrizioni greche aggiunge una traduzione in latino.

In molti casi ha realizzato anche un disegno dei monumenti che recano le iscrizioni

I disegni erano fatti su fogli separati, attaccati al manoscritto tramite ceralacca

Le caratteristiche del codice rendono chiaro che si tratta di un lavoro preparatorio

Semantica del layout

- Numerazione
- Informazione di contesto
- Disegno dell'iscrizione
- Testo di iscrizione latina
- Testo di iscrizione greca
- Traduzione dal greco al latino

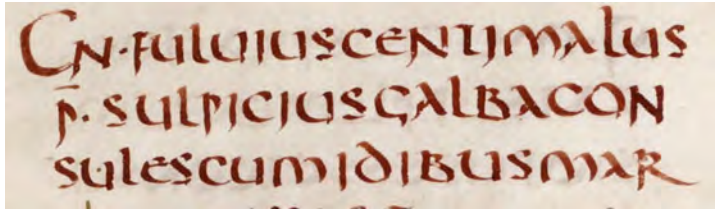
Aree sovrapposte

Nel caso di iscrizioni, cartigli e altre forme di scrittura su immagine, le aree d'interesse si sovrappongono

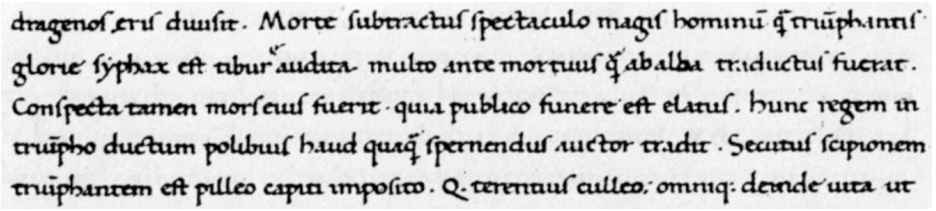
RICONOSCIMENTO DEL TESTO

Esempi di scrittura

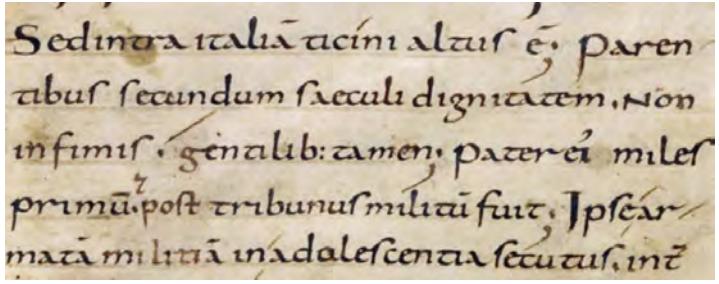
Onciale



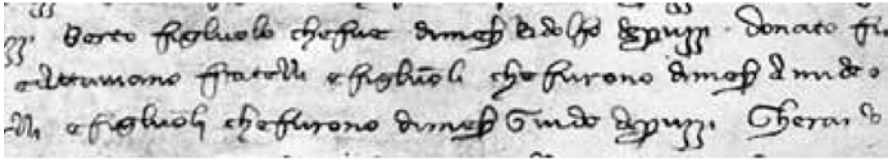
Umanistica



Carolina



Mercantesca



Scelta del modello

```
federico@hen:~/Desktop$ kraken list
Retrieving model list ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 83% 20/24 0:00:04 0:01:00
10.5281/zenodo.8425684 (pytorch) - Experimental printed syriac model
10.5281/zenodo.8193498 (pytorch) - Transcription model for Lucien Peraire's handwriting (French, 2
0th century)
10.5281/zenodo.7933463 (pytorch) - HTR model for German manuscripts trained from several datasets
10.5281/zenodo.7933402 (pytorch) - Fraktur model trained from enhanced Austrian Newspapers dataset
10.5281/zenodo.7631619 (pytorch) - Model trained on all available data, from 8th to 16th century,
from GalliCorpora, CREMMA Medieval and CREMMA Medieval Lat, as well as data from Eutyches, Carolin
e Minuscule, DecameronFR. Transcription guidelines: https://hal.archives-ouvertes.fr/hal-03697382
10.5281/zenodo.7410529 (pytorch) - Gallicorpora+ ancient prints (Litterature)
10.5281/zenodo.7051646 (pytorch) - Printed Urdu Base Model Trained on the OpenITI Corpus
10.5281/zenodo.7051644 (pytorch) - Printed Persian Base Model Trained on the OpenITI Corpus
10.5281/zenodo.7050342 (pytorch) - Printed Ottoman Base Model Trained on the OpenITI Corpus
10.5281/zenodo.7050296 (pytorch) - Printed Arabic Base Model Trained on the OpenITI Corpus
10.5281/zenodo.7050270 (pytorch) - Printed Arabic-Script Base Model Trained on the OpenITI Corpus
10.5281/zenodo.6669508 (pytorch) - Cremma-Medieval Old French Model (Litterature)
10.5281/zenodo.6657809 (pytorch) - Model train on openly licensed data from HTR-United. All French
 manuscript data from the 17th century to the 21st were used (72k lines).
10.5281/zenodo.6542744 (pytorch) - LECTAUREP Contemporary French Model (Administration)
10.5281/zenodo.5468665 (pytorch) - Medieval Hebrew manuscripts in Sephardi bookhand version 1.0
10.5281/zenodo.5468573 (pytorch) - Medieval Hebrew manuscripts in Italian bookhand version 1.0
10.5281/zenodo.5468478 (pytorch) - Medieval Hebrew manuscripts in Ashkenazi bookhand
10.5281/zenodo.5468286 (pytorch) - Medieval Hebrew manuscripts version 1.0
10.5281/zenodo.2577813 (pytorch) - A generalized model for English printed text
```

Valutazione dell'accuratezza

L'HTR è valutato in base all'**accuratezza**, una misura espressa dalla seguente formula:

matches / (matches+mismatches+adds+dels)

coerentemente con la seguente formula generale:

TP+TN / (TP+TN+FP+FN)

i matches rappresentano l'accordo tra l'HTR e la ground truth

Fine-tuning

Il fine-tuning è fondamentale per aumentare l'accuratezza anche di molti punti sui propri documenti

Consiste nella creazione di un modello ad hoc a partire da un modello generico

Generalmente, possono essere necessarie da 5 a 40 pagine di trascrizione per creare un nuovo modello, molto più accurato sui propri documenti rispetti a un modello generico

Limiti e prospettive dell'HTR

Un sistema di HTR, così come un sistema di OCR, esce dalla pura sperimentazione per diventare economicamente vantaggioso solo quando il tempo impiegato a correggere manualmente gli errori è effettivamente inferiore al tempo impiegato a produrre una trascrizione interamente manuale

L'HTR come sistema di classificazione di segni alfabetici dovrebbe articolarsi al punto di poter distinguere automaticamente allografi e idiografemi. Integrando questi requisiti della paleografia digitale, dovrebbe essere capace di identificare aree geografiche di provenienza dei manoscritti e mani diverse sullo stesso manoscritto

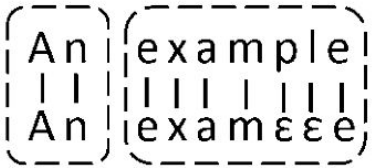
POST-PROCESSING

Allineamento con GT o con riconoscimenti automatici

https://link.springer.com/article/10.1007/s10032-020-00359-9

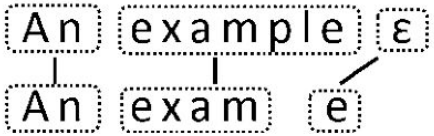
GT: An example
OCR: An exame

Char alignment first:



char errors: 2
word errors: 1

Direct word alignment:



word errors: 2

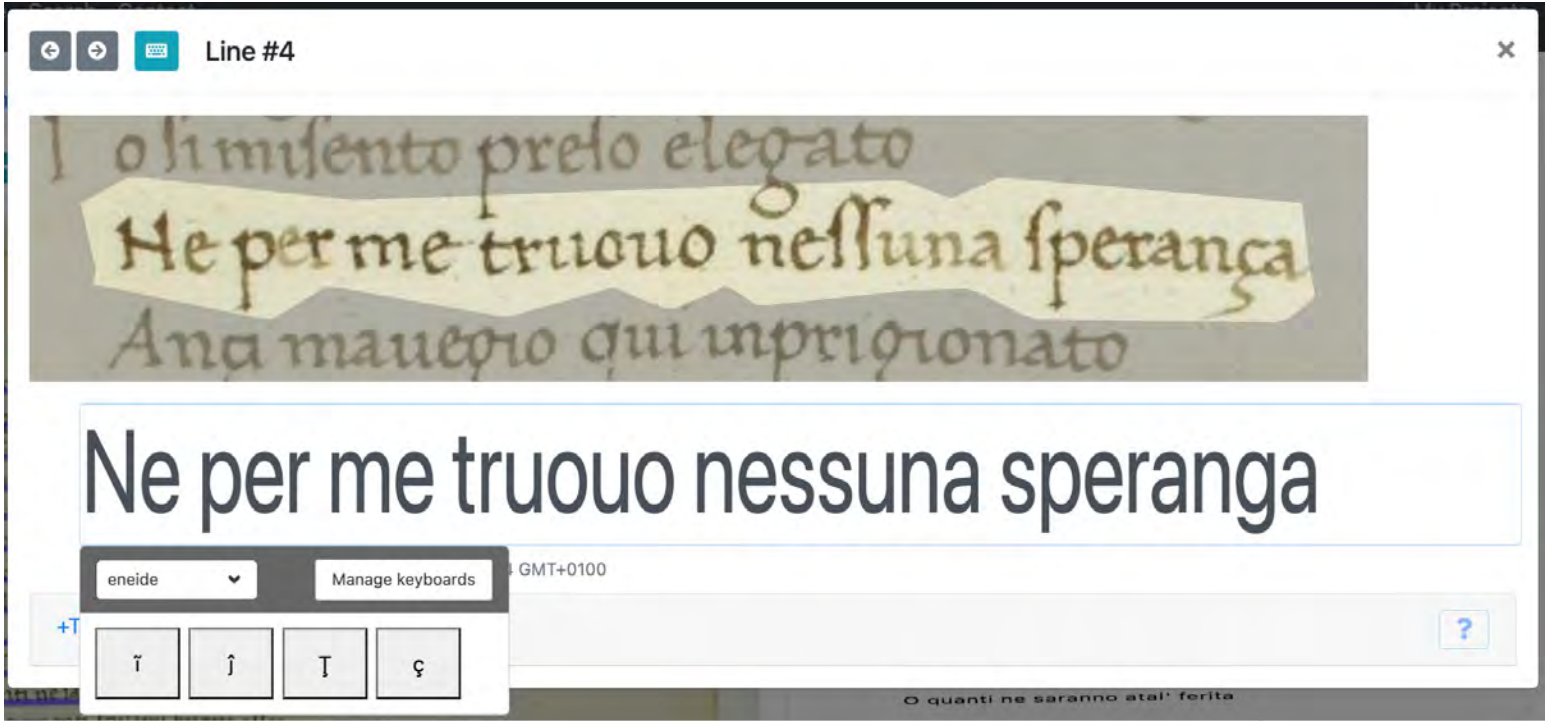
https://bioboot.github.io/bimm143_W20/class-material/nw

Allineamento di edizioni concorrenti

L’allineamento dei risultati dell’HTR con edizioni digitali già disponibili dello stesso testo può essere molto utile. Ma per evitare **contaminazioni**, è fondamentale distinguere gli errori di HTR dalle varianti reali. L'accordo tra la base di collazione (BC) e l'output di HTR rafforza il riconoscimento automatizzato.
Il disaccordo invece è dovuto:

- a una vera **variante**
  - correttamente riconosciuta da HTR ← da evidenziare come una possibile variante (parole vere molto diverse dalla CB, ad es. "biondi capelli" vs "crini dorati")
  - riconosciuta dall’HTR con errori ← da evidenziare come una possibile variante con errori (non-parole o pseudoparole molto diverse dalla BC, ad es. "biordi capclli" vs "crini dorati")
- a un **artefatto** generato da HTR
  - che può essere autocorretto con un alto grado di confidenza ← da autocorreggere con la/le parola/e nella BC ma da evidenziare per un controllo manuale (non-parole o pseudoparole molto vicine alle parole nella BC, ad es. "cnini poiati" vs "crini dorati")
  - che necessita di intervento umano ← da evidenziare come un possibile errore (parole vere molto vicine alle parole nella BC, ad es. "canini orati" o "crini indorati" vs "crini dorati")

Correzione manuale

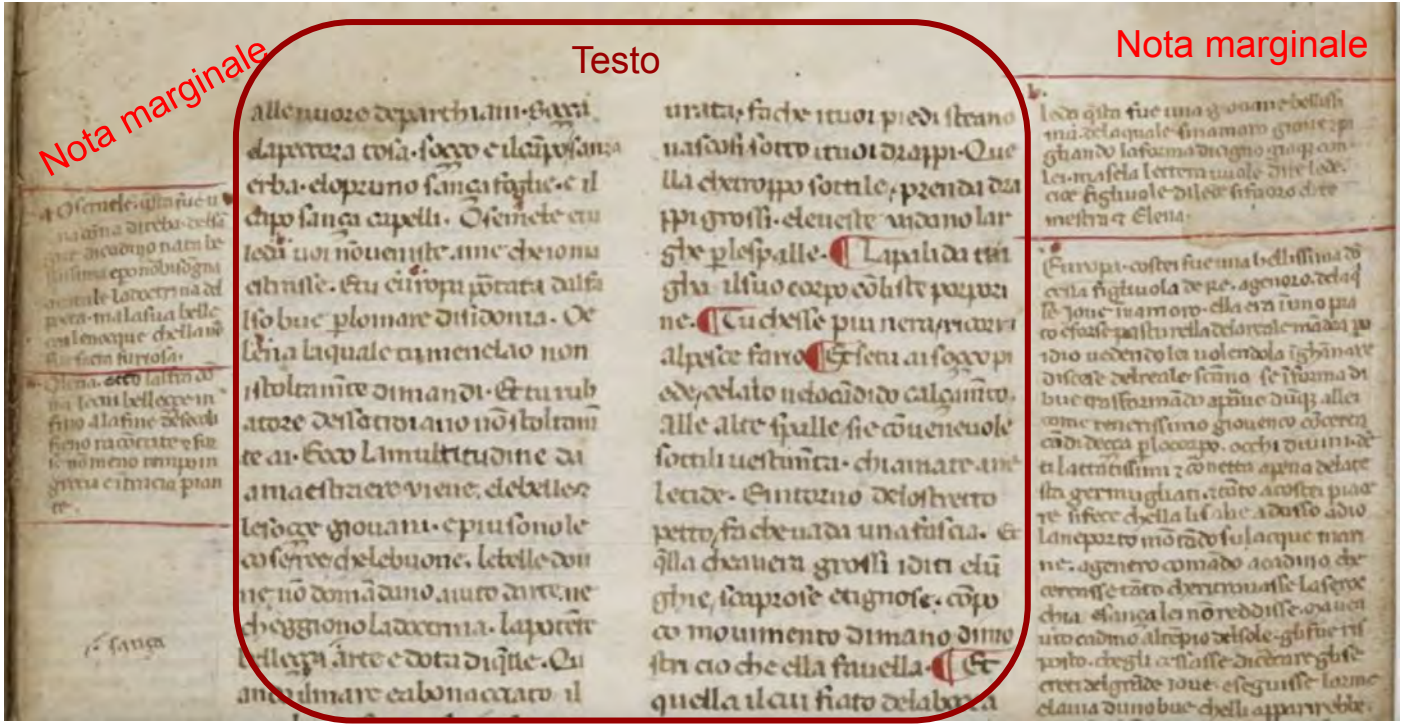


Codifica del testo

Una volta corretto manualmente il testo, un file XML-ALTO può facilmente essere convertito in XML-TEI tramite fogli di trasformazione XSLT. L'edizione (ultra)diplomatica sarà rappresentata all'interno dell'elemento **<sourceDoc>...</sourceDoc>**, che preserza nell'elemento **<zone>...</zone>** le coordinate delle aree di testo, e la corrispondente edizione interpretativa sarà invece predisposta all'interno dell'elemento **<text>...</text>** per ulteriori elaborazioni automatiche o manuali.

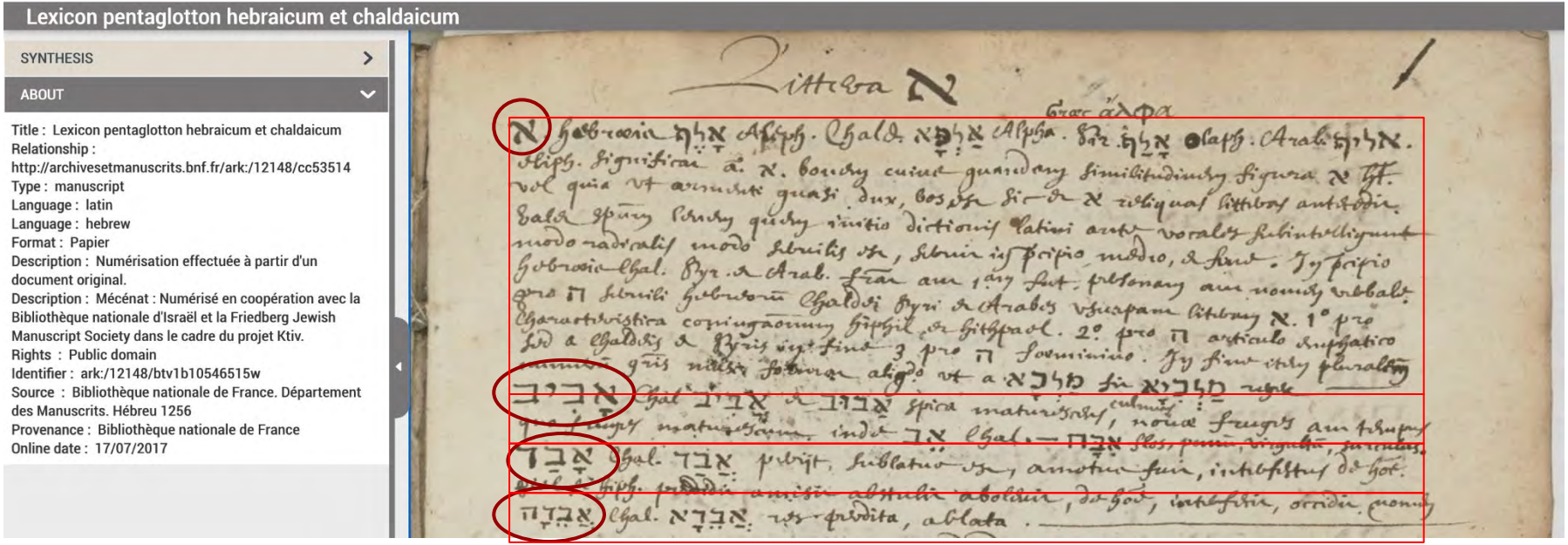
SEMILAVORATI

Testi e paratesti



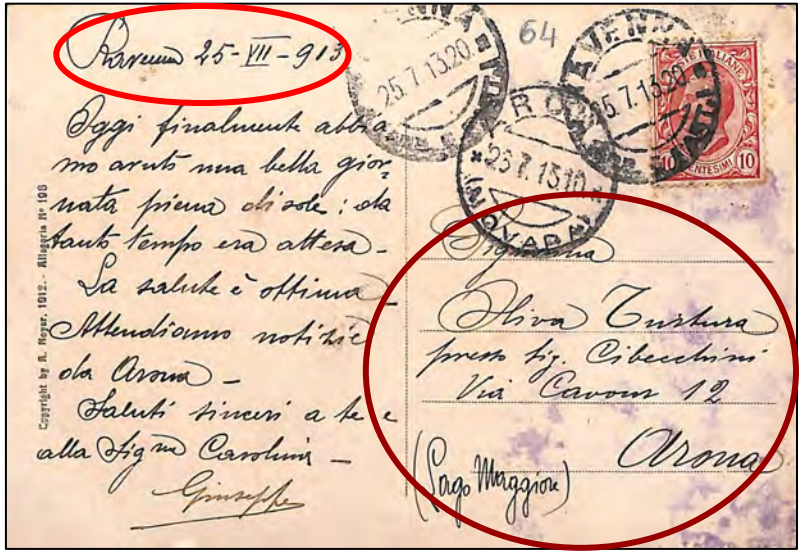
Opere lessicografiche

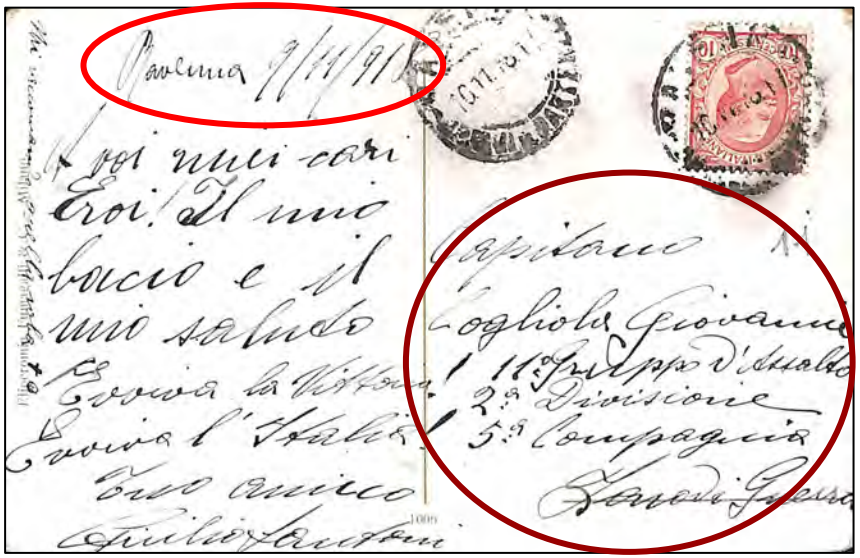
Per le opere lessicografiche, l'attenzione può concentrarsi sul lemma per indicizzare i box delle voci lessicali



Documenti d'archivio, cartoline…

In base agli obiettivi, alla mole di dati e ai finanziamenti dei progetti, può essere opportuno concentrare la correzione solo su aspetti specifici del testo, come le entità nominate o i dati del destinatario

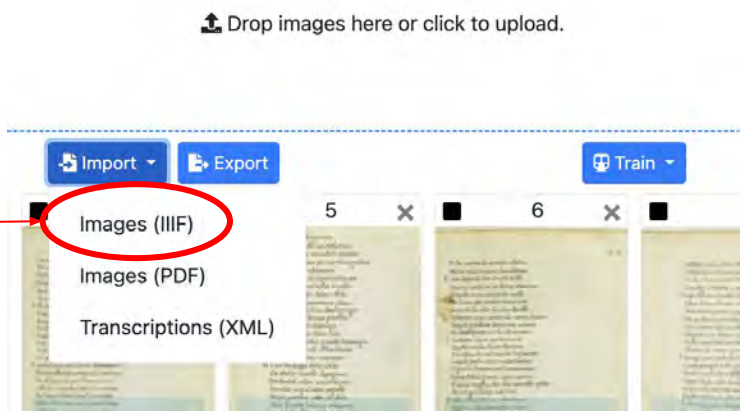




RUOLO DELLE INFRASTRUTTURE

IIIF

È possibile importare le immagini delle pagine de manoscritti tramite il protocollo IIIF



Linee guida dettagliate

Per garantire l'interoperabilità dei training sets prodotti, è necessario attenersi a linee guida

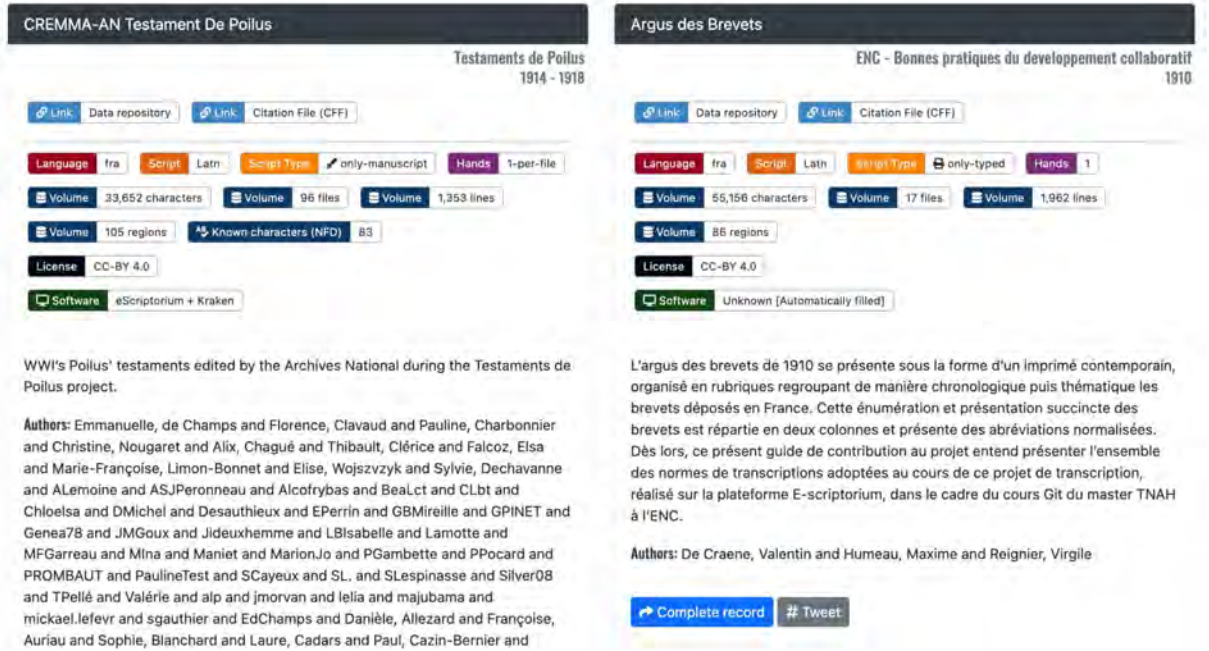
- autorevoli
- dettagliate
- modulari

Si vedano ad esempio le linee guida del progetto CREMMA dell'École des Chartes, Paris: https://hal.science/hal-03697382

HTR United

https://htr-united.github.io

HTR-United è un catalogo di metadati relativi ai training sets necessari a creare modelli di trascrizione o di segmentazione. I metadati sono forniti dai creatori stessi di questi set di dati



H2IOSC


CLARIN-IT


DARIAH-IT


H2IOSC
Humanities and cultural Heritage Italian Open Science Cloud


E-RIHS.it


OPERAS

K-Centres

I Knowledge Centres di CLARIN, tramite i loro **help desks**, mettono in contatto ricercatori, professionisti, studiosi e studenti con gli esperti di specifici settori

DiPText-KC

CLARIN Knowledge Centre for Digital and Public Textual Scholarship

DiPText-KC offers expertise on methods, data, instruments and technologies relevant in the field of Philological and Literary Studies, History, Art History and Cultural Heritage.

Its actions aim at:

- sharing information with scholars and students about the state of the art in digital scholarly editing and text annotation through domain-specific languages;
- supporting scholars and students in the creation and publication of digital scholarly editions and resources;
- organizing training activities (for instance webinars, workshops and summer schools).

Tour de CLARIN - IMPACT-CKC K-Centre

Submitted by Jakob Lenardič on 4 November 2019

Blog post written by Isabel Martínez-Sempere, edited by Darja Fišer and Jakob Lenardič

The IMPACT Centre of Competence in digitisation, founded in 2012, is a non-profit organisation that aims at making digitisation of historical texts better, faster and cheaper. Depending on the language, the period covered by historical texts is different as language change is not homogeneous. With regard to languages, IMPACT is mainly focused on European languages, but always open to widen the scope worldwide as our members' expertise comprises non-western languages as well.

IMPACT is based in Spain and hosted by Fundación Biblioteca Virtual Miguel de Cervantes. From an organisational point of view, IMPACT is governed by an Executive Board composed by representatives of its premium member institutions. On a daily basis, IMPACT is managed by a General Director, Francis Ballesteros (Spain), a Scientific and Technological Director, Tomasz Parkola (Poland), the Executive Board Chair, Frieda Steurs (the Netherlands), and a Manager, Isabel Martínez (Spain). A description on the people behind IMPACT is available at here.

In 2018, IMPACT was recognised as a CLARIN Knowledge Centre with the name IMPACT CLARIN K-Centre in Digitisation (IMPACT-CKC). This recognition is aligned with the Centre's objectives, i.e. supporting humanities researchers, cultural heritage professionals and computer scientists in their daily activities. In this context IMPACT offers the following:

IMPACT Activities

IMPACT-CKC Helpdesk

Through this helpdesk, IMPACT provides first-line assistance to researchers on digitisation techniques, tools, materials, etc. Researchers are welcome to seek advice on digitisation and related fields.

HTR E DIDATTICA

Coinvolgere insegnanti

È importante progettare l'attività di digitalizzazione **insieme** agli insegnanti per

- valutare la coerenza con il **programma scolastico**
- individuare **legami con il territorio**
- stabilire **obiettivi compatibili** con i tempi della programmazione scolastica e le competenze degli studenti

Coinvolgere studenti e studentesse

Gli studenti spesso trovano l'attività di digitalizzazione coinvolgente perché

- si sentono parte di un'attività di **ricerca**
- apprendono i rudimenti della **paleografia**
- devono decifrare le **abbreviazioni** (genuinità)
- affrontano testi **inediti** (novità)

J. Tortellius, *Orthographia graeca*, Venezia, 1471

Adattare e semplificare

Bisogna tener conto sia del contesto di ricerca sia del scolastico e adeguare le tecnologie alle contingenze

- disponibilità di una piattaforma di proof-reading come eScriptorium
  - correzione della segmentazione
  - correzione del testo linea per linea
  - uso del tastierino per i caratteri speciali
- indisponibilità di una piattaforma di proof-reading
  - predisposizione di documenti di testo con la numerazione delle linee
  - uso di convenzioni per le abbreviazioni

Riusabilità

Per fare in modo che i materiali prodotti durante l'attività didattica siano riusabili per scopi di ricerca, è necessario che vengano usate le medesime linee guida sia in progetti di ricerca sia in progetti didattici

Gli adeguamenti tecnologici non devono compromettere l'uniformità delle trascrizioni

Validazione

Le trascrizioni prodotte durante l'attività didattica necessitano di almeno due livelli di validazione

- gli insegnanti
    - correggono gli errori di trascrizione, prestando attenzione all'apprendimento delle varietà linguistiche previste dal programma scolastico
    - adottano criteri di valutazione adeguati al contesto scolastico
- i ricercatori
    - controllano la conformità alle linee guida proposte
    - verificano l'accuratezza delle trascrizioni