

LA TRASCRIZIONE AUTOMATICA DI TESTI CON SISTEMI DI HANDWRITTEN TEXT RECOGNITION (HTR)

INTRODUZIONE ALLA PIATTAFORMA
TRANSKRIBUS (READ COOP)

Università degli Studi di Trento – Seminari di Digital Humanities
13 febbraio 2025



UNIVERSITÀ
di **VERONA**
Dipartimento
di **LINGUE**
E LETTERATURE STRANIERE



STEFANO BAZZACO
stefano.bazzaco@univr.it

struttura del seminario

PART1_ATR IN THEORY

1. introduction to ATR (digitization, brief history of OCR/HTR)
2. State-of-the-art of ATR softwares
3. Transkribus in theory (features, general workflow, results)
4. Transkribus: recent advances

PART2_ATR IN PRACTICE

Transkribus Lab

MATERIALI
SEMINARIO



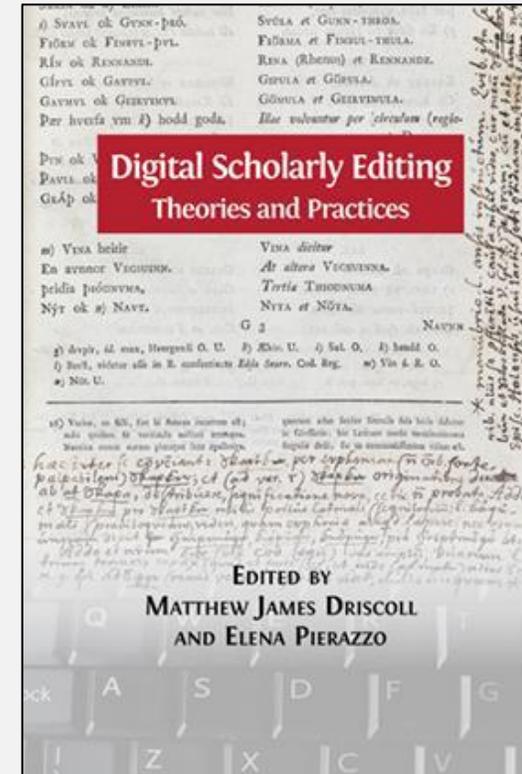
STEFANO BAZZACO
stefano.bazzaco@univr.it

Automatic Text Recognition and the Humanities (1)

M. Driscoll and E. Pierazzo

detected 7 different aspects in which computer science affects philological/historical work (*Digital Scholarly Editing. Theories and Practices*, 2016)

1. *Detection of primary sources*
2. *Digitalization (images)*
3. *Transcription of primary sources*
4. *Manipulation of large amount of data*
5. *Collation / Cladistic methods*
6. *Encoding and metadata standards*
7. *Social / Collaborative editing*

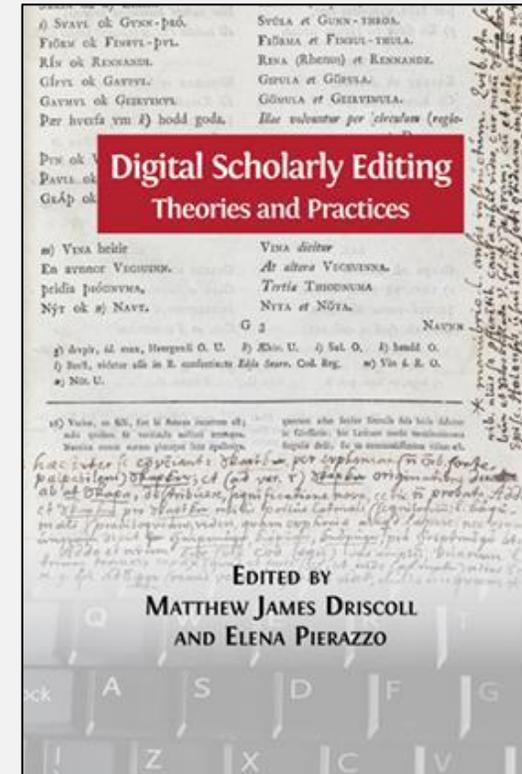


Automatic Text Recognition and the Humanities (1)

M. Driscoll and E. Pierazzo

detected 7 different aspects in which computer science affects philological/historical work (*Digital Scholarly Editing. Theories and Practices*, 2016)

1. *Detection of primary sources*
2. *Digitalization (images)*
3. *Transcription of primary sources*
4. *Manipulation of large amount of data*
5. *Collation / Cladistic methods*
6. *Encoding and metadata standards*
7. *Social / Collaborative editing*



last 40 years: > the supports of our collective memory have changed

changes the relationship between humanities scholars and their own object of study

DIGITIZATION (image format)

It consists of scanning library documents and converting them into a digital image.

- Evolve with technology
- It is based on standards and good practices
(e.g. high resolution, metadata, technical declaration)
- Promotes synergy between academics and library institutions

In the beginning there was **DIGITIZATION**

M. Terras (2010), *The rise of digitization. An overview*

Defined 3 different stages of development of this process:

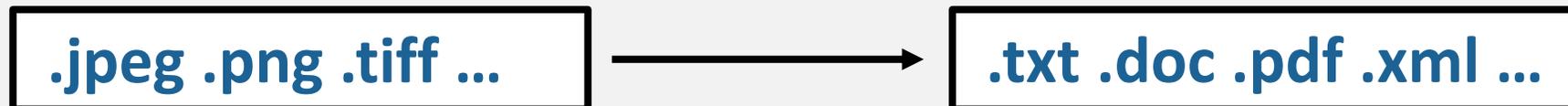
- **Early years:** in the 1980s conversion of printed source materials into digital files started to be widespread (small case, in-house projects of limited scope of interesting)
- **Decade of Digitization:** in the 1990s digitization efforts significantly increased (different forces: changes in public policies, networked technologies, institutional funds > large scale, ambitious projects)
- **Rise of Digitization programmes:** from 2000 digitization became a commonplace (advanced images technologies, large scale collections, centrally funded initiatives, emergence of commercial interests)

«The focus on the majority of early projects tends to be large scale, with large volumes of material being captured, *in the hope that **Optical Character Recognition technologies** would then turn the resulting images into electronic text.* Much of this research was optimistic, but the trial and error approach adopted by pioneering projects [...] helped to establish many useful guidelines for subsequent digitization attempts.» (Terras, 2010)

Digitization process and best practices in some way were closely linked to the development of OCR systems

The term OCR refers to all those instruments and practices that permits to transform a *digitized object* (scanned image) into an *electronic encoded text* (machine readable form / secuencia of bits), measurable and quantifiable by computers

It corresponds to the extraction of the text content of an image file and its conversion into an electronic text file (different formats):



In the very beginning there was... TEXT RECOGNITION

Automatic Text Recognition is a sub-field of **Automatic Recognition**
(other areas: *speech recognition, radio frequencies, magnetic bands, barcodes*)

Early stages in this field may be traced back to technologies involving telegraphy and creating reading devices for the blind

Overview [\[edit\]](#)

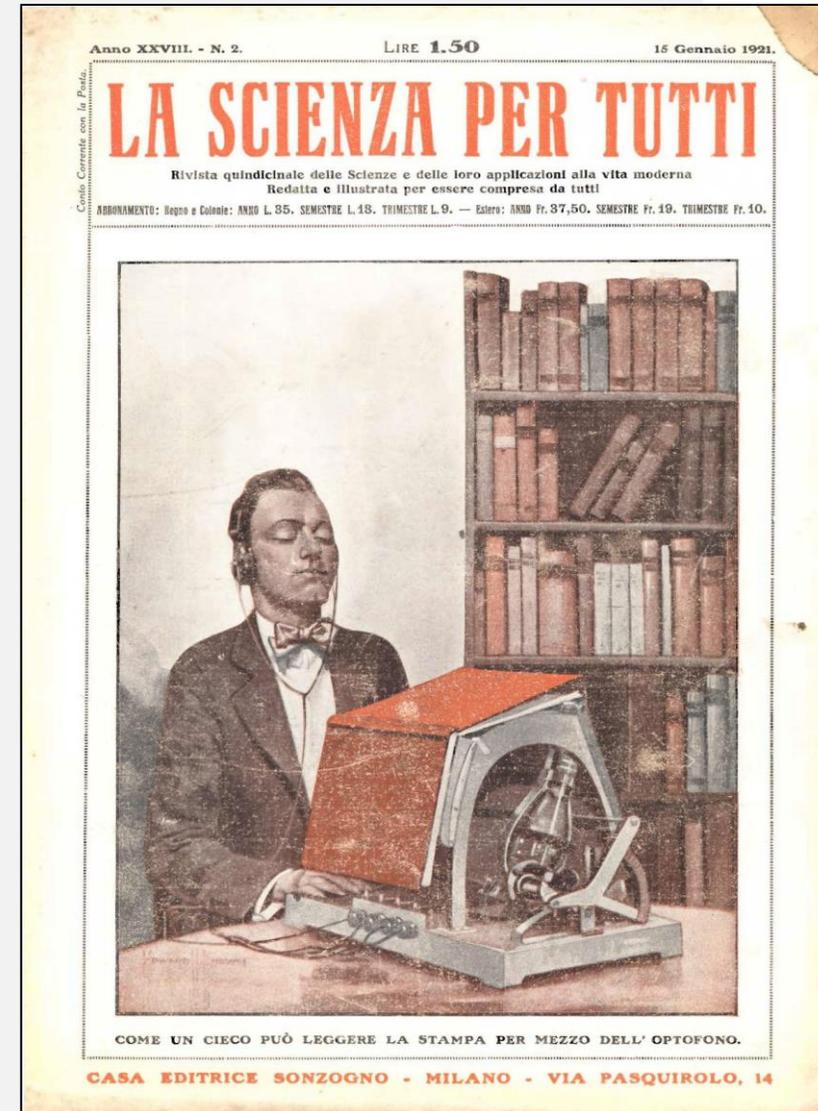
Time period	Summary
1870–1931	Earliest ideas of optical character recognition (OCR) are conceived. Fournier d'Albe's Optophone and Tauschek's Reading Machine are developed as devices to help the blind read. ^[1]
1931–1954	First OCR tools are invented and applied in industry, able to interpret Morse code and read text out loud. The Intelligent Machines Research Corporation is the first company created to sell such tools. ^[2]
1954–1974	The Optacon , the first portable OCR device, is developed. Similar devices are used to digitise Reader's Digest coupons and postal addresses. Special typefaces are designed to facilitate scanning. ^{[1][3][4]}
1974–	Scanners are used massively to read price tags and passports. ^[5] Companies such as Caere Corporation, ABBYY and Kurzweil Computer Products Inc, are ...

https://en.wikipedia.org/wiki/Timeline_of_optical_character_recognition

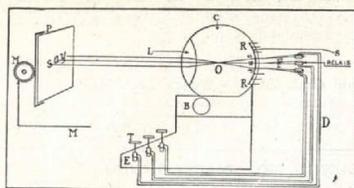
A brief history of Optical Character Recognition (OCR)

First steps:

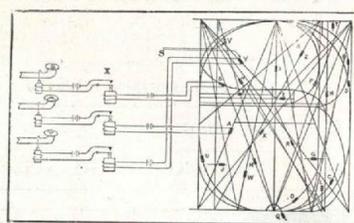
- 1870 Charles R. Carey invents the **retina scanner** (an image transmission system that used photocells)
- 1885 P. Nipkow invents the **Nipkow Disc** (an image scanning device; breakthrough for several devices, f.e. television)
- 1912 E. Fournier d'Albe conceived the **Optophone** (a scanner that produced different sounds if moved across the page)
- From the 1920s: **machine reading devices** in support of the blind



LA MACCHINA CHE LEGGE E CHE SCRIVE



Schema generale della macchina che legge e che scrive: P, pagina da copiare; M, movimento d'orologeria a scappamento elettrico che regola lo spostamento del foglio; O, punto d'incrocio dei raggi, che può essere spostato verso la lente per ingrandire l'immagine rovesciata delle lettere; C, camera oscura a sfera cava o ad elissoide, per aumentare il percorso dei raggi dopo l'incrocio e quindi l'immagine; R, retina; S, cellule di selenio; F, fili formanti circuito con le cellule congiungendole elettricamente ai relais rispettivi; D, circuiti prin-



cipali azionanti ognuno un'elettrocalamita (E) comandante un tasto (T); B, rotolo su cui è avvolta la carta da scrivere. — Schema del funzionamento dell'occhio elettromeccanico; S, cellule di selenio; X, relais che mantengono aperti i circuiti dei tasti, salvo lasciarli chiudere quando le rispettive cellule di selenio sono oscurate dall'immagine della lettera.

Dare « un occhio » alla macchina da scrivere e farle leggere e copiare uno scritto senza bisogno, diciamo così, della « dettatura delle dita »! — Ecco il problema proposto da un inventore; ecco non ancora la macchina ma la notizia della sua invenzione che perviene dall'America del Sud. Vediamo notizia, macchina e funzionamento; e cominciamo dall'occhio della macchina.

È, naturalmente, un occhio meccanico; occhio che « vede » lo stampato da trascrivere come, od all'incirca, il fonografo vede la pagina musicale incisa sul disco; un occhio elettrico, fondato su quel notissimo fatto che è la resistenza variabile del selenio alla luce e che ha generato tante interessanti ricerche di fotografia a distanza e di trasformazioni della luce in suono e viceversa.

Il principio sul quale tutto il congegno si basa è di una certa geniale semplicità: consiste nella constatazione che ogni lettera dell'alfabeto ha nella sua forma un punto caratteristico che non si confonde con nessun'altra lettera. Ciò, se si sovrappongono tutte le lettere una sull'altra, tracciandole sufficientemente grandi e fini, per la chiarezza, si potranno sempre trovare tanti punti quante sono le lettere incrociate. Il che si può constatare in uno degli schemi che qui figurano ad illustrazione di quanto veniamo esponendo.

L'inventore ha disposto sulla macchina una piattaforma orizzontale e su di essa un occhio, formato da una camera oscura sferica che anteriormente, nel centro, porta una lente convessa. Questa ha per effetto di raccogliere i raggi provenienti dallo scritto da copiare, che le sta dinanzi, e di rifletterne l'immagine capovolta in fondo alla camera. Come è noto, e come si vede in altro dei nostri schemi, tale capovolgimento è dovuto all'incrociarsi dei raggi: solo che il punto d'incrocio non si verifica nel centro della sfera, come si rappresenta per comodità di disegno e d'illustrazione, ma assai più vicino alla lente; e l'inventore, per rendere anche più sensibile la distanza dell'incrocio dal fondo, parla d'una camera a sezione elittica, con l'asse maggiore orizzontale. In tal modo, i raggi, deviando, producono un ingrandimento dell'immagine capovolta, e facilitano così al costruttore la fabbricazione d'una retina più grande, coi punti caratteristici più distanti l'uno dall'altro, e di più sicura sensibilità.

Difficoltà notevole è quella di mantenere l'immagine sempre della medesima grandezza, qualun-

que sia il carattere da copiare: ma vi si può riuscire o interponendo fra il leggio e l'occhio una o più lenti concave, o — ed è meglio — rendendo mobile la lente dell'occhio, per avvicinarla od allontanarla come occorre dal centro della sfera o dell'elissoide.

La retina è formata da una serie di fili metallici, meno complicati che nel nostro disegno, perché raffigurano soltanto le linee speciali di ciascuna lettera: su tali linee i punti caratteristici sono rappresentati da minuscole cellule di selenio, ad ognuna delle quali fanno capo i due fili d'una corrente. Nella nostra figura schematica, per maggior chiarezza, ogni cellula è inserita in un circuito speciale con batteria propria; ma nel fatto è più comodo porre tutte le cellule in derivazione da un unico circuito principale, equiparando le diverse distanze delle cellule con piccole resistenze supplementari nascoste nella tavoletta che sorregge l'occhio. Ognuno di questi circuiti derivati, che normalmente è chiuso e quindi percorso dalla corrente, va a finire, a breve distanza dalla cellula, in un relais, il quale, quando funziona, mantiene normalmente aperto un altro circuito in cui è inserita una elettrocalamita, posta proprio sotto al tasto della lettera corrispondente. La corrente che dovrà azionare è più forte di quella attraversante il selenio; anch'essa può provenire in derivazione da un'unica pila, tanto più che i tasti devono usarla, per abbassarsi, uno per volta.

Si supponga ora che dinanzi all'occhio si presentino uno stampato qualsiasi.

Nel campo della lente penetrano le immagini di parecchie lettere, sopra, sotto, a destra ed a sinistra del centro; ma essendo la retina limitata nel fondo dell'occhio, potrà rimanere impressa soltanto dalla lettera che si trova sull'orizzontale passante per il centro e per la retina medesima. Se sopra il leggio vi è, ad esempio, la parola inglese *say*, che significa « dire », soltanto la lettera *a* colpirà la parte sensibile dell'apparecchio. L'impressione consiste nel sovrapporsi dell'immagine sopra il punto caratteristico — e dello solo — ad essa corrispondente: ma siccome l'immagine è nera su bianco, così rappresenterà un'ombra in mezzo alla luce. La cellula di selenio, oscurata, aumenterà la sua resistenza indebolendo la corrente che la percorre: questa non avrà più la forza di far funzionare il relais e di mantenere

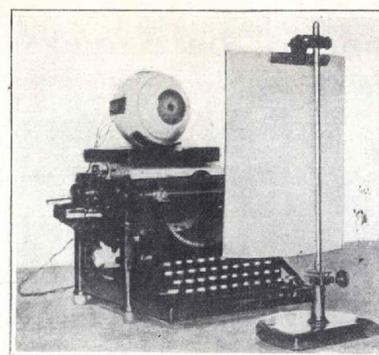
aperto il circuito del tasto, il quale si abbasserà per l'azione della elettrocalamita sottostante.

Se dopo aver fatto scrivere la lettera *a* nella parola *say*, facciamo scorrere orizzontalmente il leggio, verso sinistra o verso destra, passerà dinanzi al centro della lente la lettera *s* o la *y*; e così sfileranno tutte quelle di una riga. Alzando in seguito il leggio e facendolo retrocedere, mentre uno schermo riparerà la lente, incomincerà la sfilata della riga seguente, e così avanti di riga in riga fino a che la pagina sia terminata.

A tale uopo l'inventore ha immaginato anche un semplice apparecchio meccanico con una ruota dentata mossa da orologeria a scappamento di un dente ad ogni tasto — che comanda esso medesimo lo scappamento, per mezzo della stessa corrente che lo abbassa — per far passare regolarmente dinanzi all'occhio tutta la pagina da copiare.

L'apparecchio è stato costruito, per la prima volta, allo scopo di copiare lo scritto medesimo della macchina da scrivere. Ma questa particolarità rivela anche il difetto più grave dell'apparecchio stesso.

Abbiamo già parlato della influenza dovuta alla grandezza delle lettere: se troppo grandi, ciascuna di esse non sarà più contenuta nella retina, e il punto dell'immagine corrispondente alla cellula di selenio può spingersi fuori del campo; se troppo piccole, possono cadere contemporaneamente in parecchie sulla retina, impressionare due cellule ed azionare due tasti, col rischio di rovinare il meccanismo della stampa. Abbiamo indicato anche, è vero, il mezzo per ovviare all'inconveniente: ma ve ne è un altro molto più serio, quello del *fa* forma. Esso basta perché i punti caratteristici corrispondenti della retina e dell'immagine



La macchina che legge e che scrive in atto di funzionare.

retina trovare nell'una o nell'altra il punto caratteristico. Perché, mentre nella macchina da scrivere ogni lettera occupa il medesimo spazio, dalla *i* resa larghissima alla *W* resa strettissima; nella stampa comune, invece, accanto alle lettere che potremmo chiamare di larghezza normale (*a, b, c, d, e, f, g, h, k, n, o, q, r, s, u, v, x, y, z*) ve ne sono larghe appena la metà (*i, j, l, t*), altre una volta e mezza (*m, n*), e le maiuscole in genere salvo *J, J*, che sono della grandezza normale per le maiuscole; altre quasi due volte (*æ, æ, M, W* ed anche più *E, E*).

Quanto allo scritto a mano, non è nemmeno da pensare a riprodurlo in tal modo. Perciò l'utilità immediata dell'invenzione è discutibile. Pure, nessuno potrebbe negarle il pregio della genialità; e nessuno può escludere che, come già avvenne altre volte per novità che parvero folle, si riesca un giorno o l'altro a perfezionare « l'occhio elettromeccanico », magari staccandolo dalla macchina e ingrandendo la sua costruzione assieme alle immagini ed alla retina per complicare quest'ultima coi caratteri di testo, sino a renderlo pratico.

A. SCIENTI.

I CACCIATORI DI SOMMERSIBILI

La guerra dei sommergibili inaugurata dalla Germania ha provocato, come tutte le novità belliche, le contro-novità, destinate a neutralizzare od a controbattere. L'Inghilterra ha infatti provveduto a dare la caccia agli insidiosi battelli nemici con navi speciali, meno insidiose forse, ma più agili: sono piccoli « monitori », azionati da motori a scoppio, rapidissimi, leggermente corazzati, ricoperti anche nelle parti superiori, per sfidare impunemente le onde e passarle da parte a parte quasi sdegnassero di sormontarle. Mobilissimi e quindi capaci di scansare facilmente le torpedini; armati di un cannone sufficiente a produrre in un sommergibile un foro per cui affonda se è immerso o non può più immergersi se sornuota; molto più veloci e numerosi di essi, in modo che la flottiglia risente poco danno per l'eventuale perdita di una unità; i monitori — di essi due sono raffigurati nella nostra copertina a colori — hanno compiuto già una volta una vera strage di sommergibili tedeschi. E sembra che avessero già ricominciato a compierla, perché, anche prima che gli

Stati Uniti minacciassero di rompere le relazioni diplomatiche, l'asprezza della guerra tedesca agli innocenti era già diminuita di parecchio dopo l'annunciata ripresa.

Naturalmente, sarebbe ingenuo considerare che una ripresa numero tre non possa seguire a quella numero due, magari dando luogo ad una riapertura del processo fatto dall'America alla Germania; ma è probabile che ogni tentativo, quanto più volte sarà ripetuto, tanto più rimarrà sterile per le maggiori contromisure che con l'andar del tempo si saranno prese. Ad esempio, nel nostro numero del 1° febbraio annunciammo che 40 cacciatori di sommergibili erano stati fabbricati in America per l'Inghilterra, e molti altri nei cantieri inglesi: ma d'allora in poi, il loro numero è gradatamente cresciuto e crescerà ancora, perché nessuno si fida delle promesse tedesche. Il problema consiste nell'affondare un sottomarino per ogni nave silurata: la partita sarebbe allora vinta — e potrebbe essere già stata vinta — perché i primi sono molto meno numerosi che le seconde.

We can talk about OCR systems referring to their actual meaning only from the 1940s

1950s – Machine Reading techniques development and necessity to control huge amounts of textual data > first commercial OCR systems starts to be developed

1960s – **First Generation of OCR hardwares**: prototypes recognized at most 10 different typefaces (fonts) / first standardized graphic system for commercial use

OCR A | OCR B

1970s – **Second Generation of OCR hardwares**: recognition extended to other printed typefaces and even to some handwritten small texts (f.e. postal codes) / Kurzweil developed the first omni-font recognition system

1980s – Production and distribution of **OCR software packages** (reduction of hardware costs, spread of personal computers > spread of OCR applications)

OPTICAL CHARACTER RECOGNITION nowadays

From 2000: along with the development of european programmes of digitalization, OCR softwares experiment their biggest enhancement

Objective: control the Big Data derived from digitalization and transform it in something computable, searchable and sharable

Consequences:

- natural contrast between e-text and facsimile image
- OCR softwares became part of the interests of private companies
- most recent OCR platforms opened recognition to **complex scripts**
(Arabic and Asian texts, Historical printed texts, Handwritten texts)

AUTOMATIC TEXT RECOGNITION and DH PROJECTS

What can humanities scholars ask to OCR softwares?

Possible applications in Digital Humanities fields:

- ❖ **Creation of Digital (Scholarly) Editions**
- ❖ **Extraction of metadata / lemmas (population of corpora/databases)**
- ❖ **Expansion of Text Mining projects**

(cuantitative text studies: *algorithmic query, Stylometry, AA, Sentiment Analysis...*)

OCR softwares can enhance all these fields providing an automated transcription of the digitized text (they supply manual, time consuming transcription)

from OCR to HTR

humanists noted some problems concerning OCR reliability:
different reasons (bad scans quality, errorfull transcriptions, ...)

> **BIAS towards OCR softwares** / sharpen the distinction between «*clean transcription*» and «*dirty OCR*».

Smith-Cordell (2017), *A research agenda for historical and multilingual OCR*

Contemporary Texts (from 1930) considered as a solved problem	Historical Texts / Multilingual Texts still a growing field, only recent improvements in deep learning assure good results
---	--

Handwritten Text Recognition (HTR) has become a solution for scholars
to transcribe Historical/Multilingual Texts with good results
(it is based on recurrent neural networks processes)

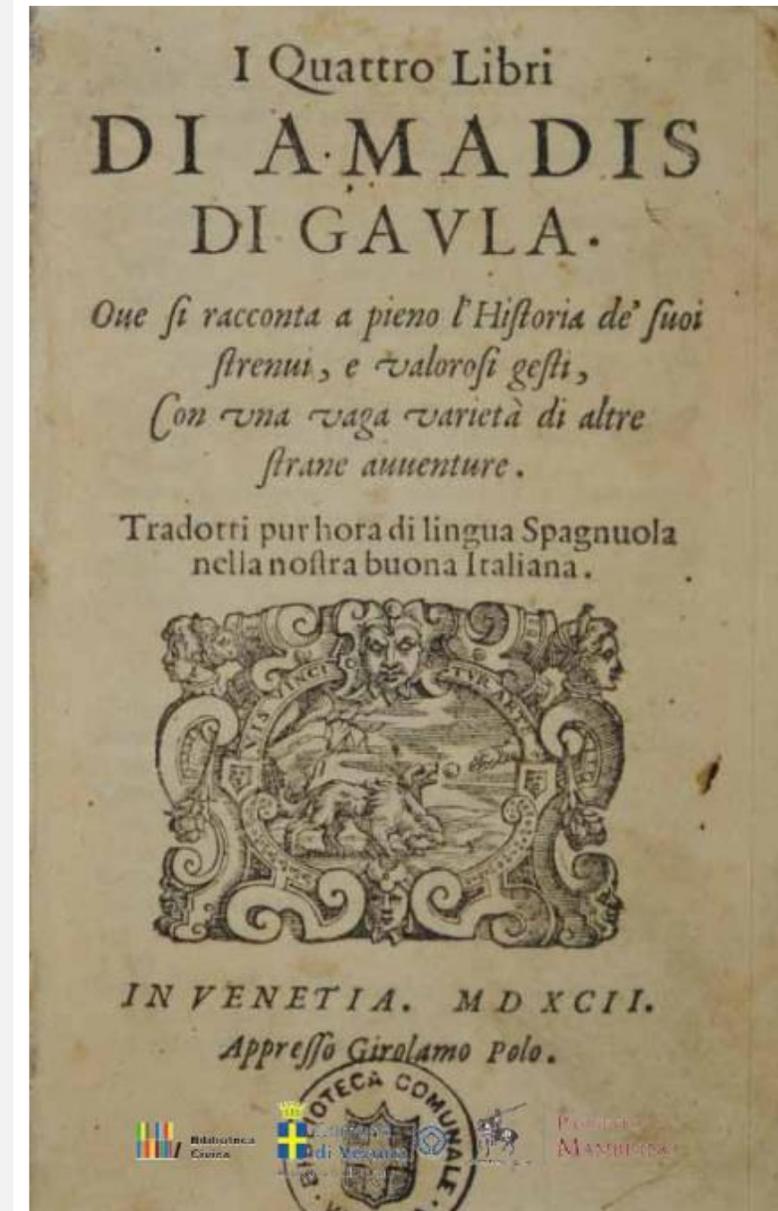


Progetto Mambrino (founded in 2008 by Anna Bognolo and Stefano Neri) studies Spanish and Italian Romances of Chivalry

Objectives: book census, libraries digitalization programmes, bibliographical databases, **creation of a Digital Library of the corpus**

Characteristics of the corpus:

- **Printed books, Venice 1530-1580**
- **writing: italics (*Manuzio's italics*)**
- **Format: octavo («pocket books»)**
- **Extension: 900-1000 pgs.**

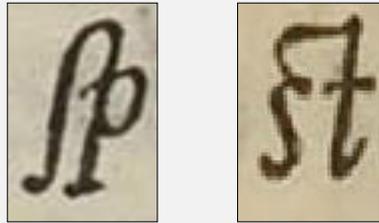


ATR and historical printed sources (Venice, XVIth C)

Specific issues of historical writings

f.e. in Italics scripts:

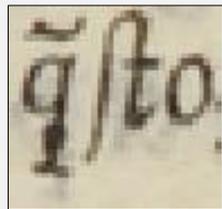
LIGATURES



TIRONIAN NOTE

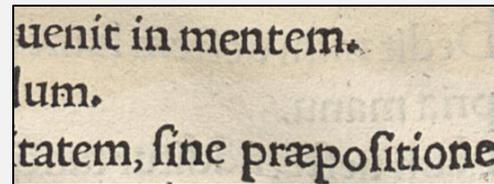


ABBREVIATIONS

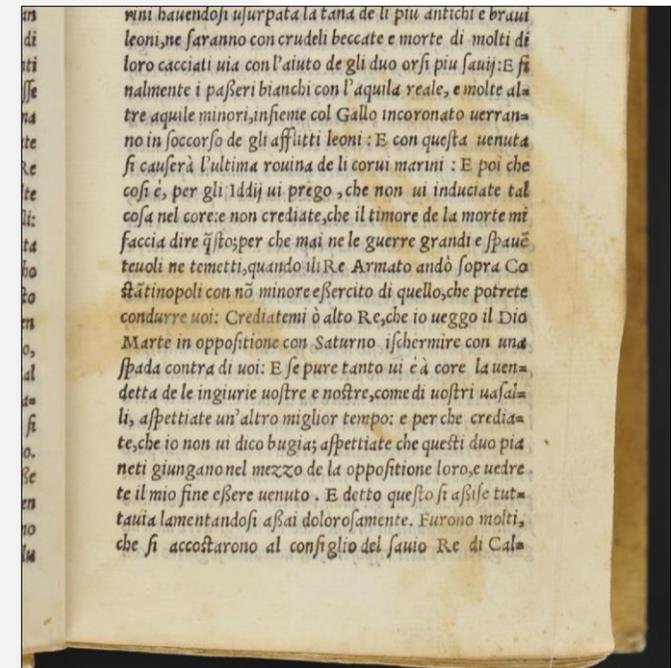


State of preservation of the sources

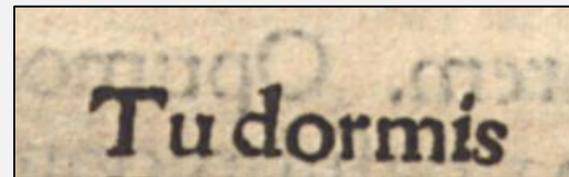
WARPED PAGES



STAINS



INK TRANSFER



CHARACTER RECOGNITION SOFTWARES

OCR / HTR softwares can be distinguished by the following features:

PROPRIETARY SOFTWARES

OPEN ACCESS SOFTWARES

SINGLE CHARACTERS RECOGNITION

LINES RECOGNITION

Keep in mind that:

- > Every software has its weaknesses and none of them ensures a 100% correct transcription for Historical / Handwritten texts
- > Only some recent softwares can provide a reliable transcription as they can be trained with the creation of a **Golden Standard Transcription** (Ground Truth)

Most famous OCR softwares

Tesseract	ABBYY Fine Reader	OCR4All
open access	proprietary	open access
Windows, Linux, MacOS	Windows, Linux, MacOS	Linux VM for Windows and MacOS
single characters recognition	single characters recognition	Recognition: characters in context
training: per glyphs	training: per glyphs	Golden Standard training neural networks and LSTM
output: .txt .doc .pdf .xml .html	output: .txt .doc .pdf .xml .html	output: .txt .xml

HTR softwares for Historical / Handwritten texts

Transkribus (READ COOP)

Open Access

Expert client: Windows, Linux (VM), MacOS
Web app from 2023

Servers at Innsbruck Univ. and community support / costs expected for extended projects

Golden Standard training
deep learning neural networks

output: TXT / DOC / PDF / PAGE XML / ALTO XML / XML-(TEI?)

eScriptorium

Open Access

Windows, Linux, MacOS

External server support or institutional installation (needs computing power)

Golden Standard training
deep learning neural networks

output: TXT / PAGE XML / ALTO XML

eScriptorium

<https://gitlab.inria.fr/scripta/escriptorium>

Developed by Scripta Project

Based on Kraken HTR engine > needs huge computational power (server hosting)

The screenshot displays the eScriptorium web interface. At the top, there is a navigation bar with the 'es' logo, 'Home', and 'Contact' links. A search bar is present with the text 'Search in dataset' and a magnifying glass icon. On the right side of the navigation bar, there are links for 'My Projects', 'My Models', and 'Hello Stefano'. Below the navigation bar, there is a breadcrumb trail: 'Description' > 'Ontology' > 'Images' > 'Edit' > 'Models' > 'Reports' > 'dataset'. The main content area is a large dashed blue box with the text 'Drop images here or click to upload.' Below this box, there is a row of action buttons: 'Select all', 'Unselect all', 'Selected 0/176', 'Import', 'Export', 'Train', 'Binarize', 'Segment', 'Transcribe', and 'Align'. At the bottom, there are 8 thumbnails of document pages, numbered 1 to 8. Each thumbnail shows a snippet of text from a manuscript, with some text highlighted in red. The text is in a historical script, likely Italian or Latin, and appears to be a historical document or letter.

The screenshot displays the eScriptorium web application interface. At the top, there is a navigation bar with the eScriptorium logo, links for Home and Contact, a search bar with the text "Search in dataset", and user information "My Projects My Models Hello Stefano". Below this, a secondary navigation bar shows tabs for Description, Ontology, Images, Edit, Models, and Reports, along with a "dataset" label and a "Zip Import" button. The main content area is split into two panes. The left pane shows a scanned page of a manuscript with the title "PRIMA PARTE DEL TERZODECIMO LIBRO DI AMADIS DI GAVLA." and a decorative flourish. The right pane shows a text editor with the transcribed text of the same page, including the title and the beginning of the first paragraph. The interface includes various toolbars for navigation, editing, and viewing.

eScriptorium Home Contact AI Search in dataset My Projects My Models Hello Stefano

Description Ontology Images Edit Models Reports dataset
Element 1 - 0001_0016_13.1_15.tif - (2441x3517) - 244.26 KB

Zip Import

PRIMA PARTE DEL
TERZODECIMO LIBRO
DI AMADIS DI GAVLA.
*Nelqual si contiene le stupende, & marauigliose
prodezze del Prencipe Sferamundi figliuolo
del ualoroso don Rogello, & di Ama-
dis d'Altra, & altri sforza-
ti cauallieri.*
Che il Prencipe Sferamundi insieme con Ama-

1 2
2 PRIMA PARTE DEL
3 TERZODECIMO LIBRO
4 DI AMADIS DI GAVLA.
5 Nel qual fi contiene le ftupende, & marauigliofo
6 prodezze del Prencipe Sferamundi figliuolo
7 del ualoroso don Rogello, & di Ama-
8 dis d'A&tra, & altri sforza-
9 ti cauallieri.
10 Che il Prencipe Sferamundi infieme con Ama-
11 dis d'Altra, fu portato dalla naue del gran Serpē
12 te dalla sfera col mezzo della donzella che la gui-

Transkribus (READ Coop)

<http://transkribus.eu>

Developed by DEA group (Digitalisierung & Elektronische Archivierung) of Innsbruck University (with other 11 institutions) and funded by Horizon2020 Programme: READ Project (Retrieval and Enrichment of Archival Documents)

Released in 2015 as part of the **tranScriptorium** project



The screenshot shows the website's header with the "tranScriptorium" logo in a cursive font and social media icons for Twitter, Facebook, and Email. A dark blue navigation bar contains the following menu items: HOME », CONSORTIUM », WORK PACKAGES, DELIVERABLES, DEMONSTRATIONS, DATASETS, PUBLICATIONS, and CONTACT INFO. The main content area displays a handwritten document snippet with three lines of text: "the whole produce and profit of their labour, and the", "same to retain subject only to the allowances to be", and "made to such prisoners in manner herein after". The first two lines are highlighted with a green border, and the third line is highlighted with a blue border. Below the document, there is a text input field containing the text "made to such prisoners in manner herein after" and a "Demonstrations" section with the text "Some project demonstrations are available at the Demonstrations section." The background of the bottom section shows more of the handwritten document.

HTR / OCR

Transkribus and eScriptorium are **Handwritten Text Recognition (HTR) softwares**

They can be adapted to printed Historical Texts (conceived as much regular HW texts)

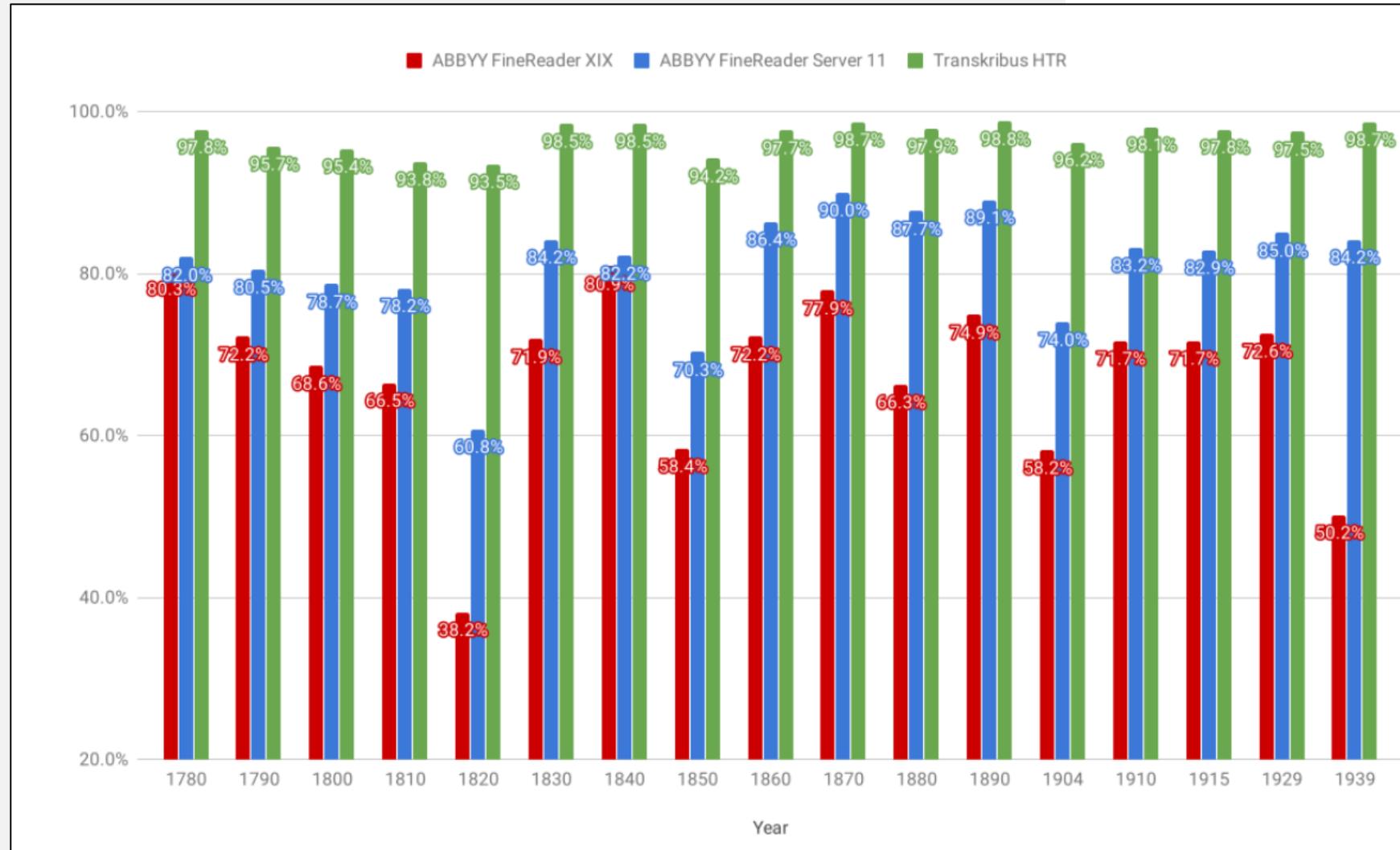
OCR	HTR
Focus on single characters	Focus on both single characters, words and context (sentence-based with an n-gram)
Preferably bi-tonal images (bynarized)	Prefer clear background, but full color and greyscale images assures good results (density detection)
Language/character variability can cause trouble	Can handle language/character variability
Trained-fixed tools	Individual or general models

A. Romein (2020), «Entangled Histories: OCR + HTR = ATR: Automatic Text Recognition»
<https://lab.kb.nl/about-us/blog/entangled-histories-ocr-htr-atr-automatic-text-recognition>

HTR / OCR: a comparison

Scholars of the University of Utrecht studied the effectiveness of HTR, on medium resolution pdf images of a German historical newspaper

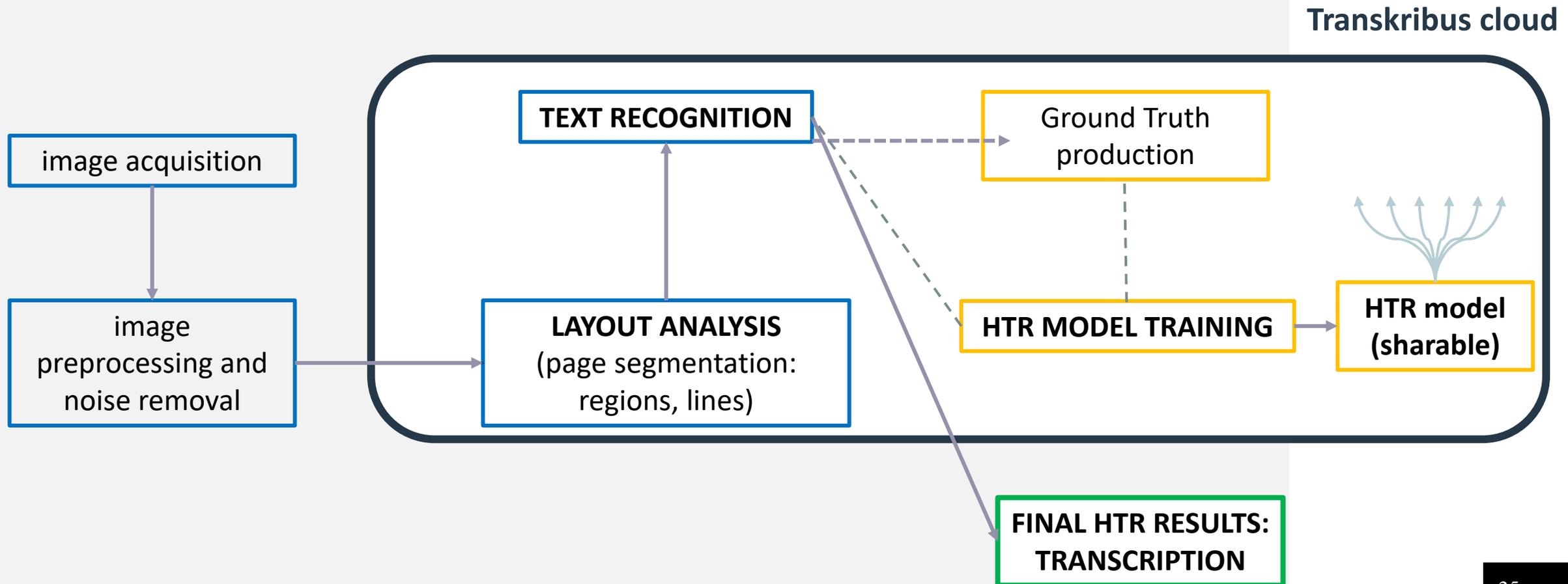
(comparing it with ABBYY FineReader OCR)



<https://dev.clariah.nl/files/dh2019/boa/0694.html>

Transkribus workflow

- Transkribus **works in cloud** > training models are stored in the Transkribus cloud
- > transcriptions are exported in the local machine



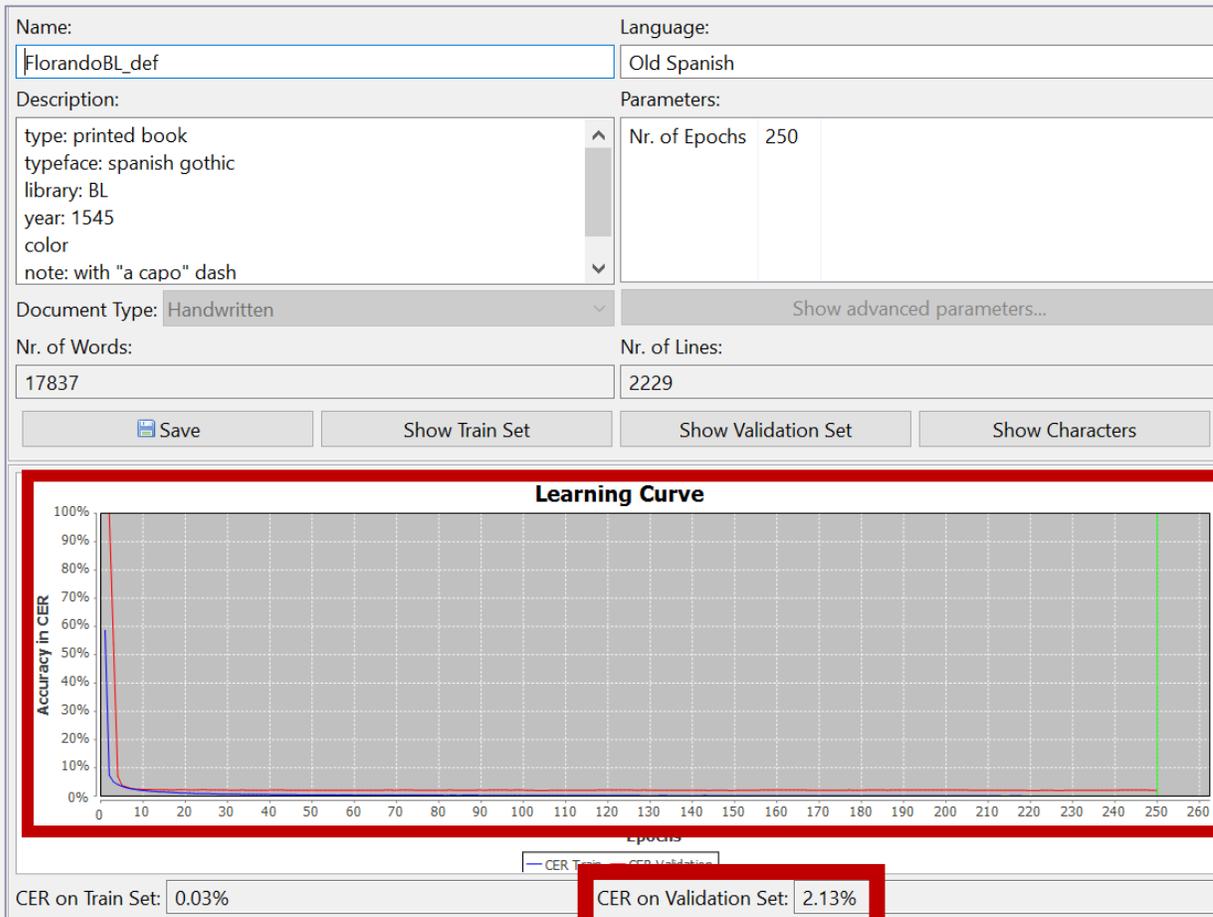
HOW TO CREATE A HTR MODEL

1. **Image import:** upload image preprocessed files to the cloud
2. **Layout Analysis:** page segmentation
3. **Ground Truth production:** manual transcription (1500-2000 words)
4. **Model Training:** dataset selection: Train Set (90%) - Validation Set (10%)
5. **Creation of a text recognition model**

TRAINING PROCESS AND MODEL EVALUATION

Training results are expressed by a percentage index called **CER (Character Error Rate)**

> CER resembles the *edit distance* between recognized text and ground truth text
(additions, suppressions, changings)



The learning curve certifies the **adequacy** of training materials – if they are not enough the 2 curves will pull away

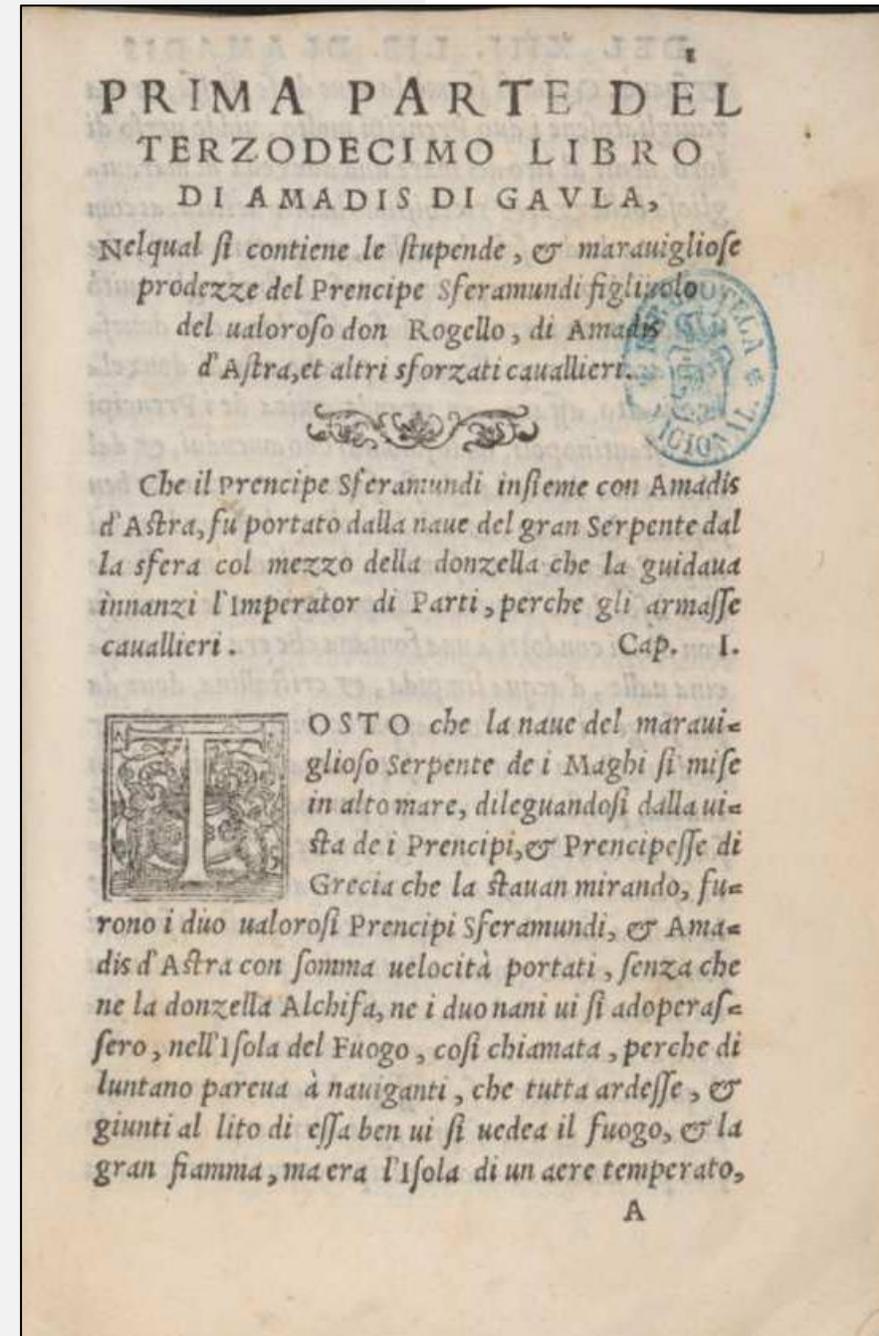
Model **efficiency** is calculated by CER on Validation Set percentage

Individual models:

Some results with printed texts

Italics (Venice, XVIc)

book	source text	CER (Validation set)
A4 – Aggiunta al Quarto Libro di Amadis di Gaula. 1563	Santiago de Compostela, Biblioteca Universitaria, 13996	0.49%
12 – Don Silves de la Selva. 1551	Verona, Biblioteca Civica, Cinq. 350-16	0.90%
13.1 – Sferamundi. Prima parte. 1558	Madrid, Biblioteca Nacional de España, 5-4978	0.81%
13.2 – Sferamundi. Seconda parte. 1560	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 19)	0.96%
13.3 – Sferamundi. Terza parte. 1563	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 20)	1.31%
13.4 – Sferamundi. Quarta parte. 1563	München, Bayerische Staatsbibliothek, P.o.hisp. 105 k-4.	0.64%
13.5 – Sferamundi. Quinta parte. 1565	Wien, Österreichische Nationalbibliothek, 40.J.16 (Vol. 22)	1.58%



Spanish Gothic script (XV-XVIc)

book	source text	CER (Validation set)
Leandro el Bel Toledo, Ferrer, 1563	Madrid, Biblioteca Nacional de España, R/9030	1.43%
Florando de Ingalaterra Lisboa, Gallarde, 1545	London, British Library, C62 H14	1.13%
Silves de la Selva Sevilla, De Robertis, 1549	Madrid, Biblioteca Nacional de España, R/865	1.58%
Siete Partidas Sevilla, cuatro alemanes, diciembre 1491	(Proyecto 7partidas digital)	0.77%

to el principe Isaladiano cō otros caualleros se fueron onde el estava / y d̄lar m̄dolo no le hallarō herida ninguna: saluo que del gran cansācio estava ansī fuera de su acuerdo: y trayendo agua se la echaron en la cara y tomo en si de lo que todos ouieron gran plazer. Al ora s̄bed q̄ qual quiera cauallero que se combatia cō las ymages aun que parecia tener muchas heridas en la batalla mas acabada quedaua tan sano como lo era de ātes que la epeçasse. El principe y los caualleros lleuando a Adātileo de la plaça se fuerō a los miradores / aun q̄ el rey quisiera q̄ lo lleuarā al palacio y le hazer algū remedio si lo vuisse menester mas el le d̄ixo que no s̄t̄ia en si mas de vn poquito cāfācio de lo mas estava muy bueno a d̄ios gracias. y viendo esto el rey mando hazer seña que se començassen las justas avn que era ya muy tarde: las quales yo aqui no cuento por turarē muy poco: mas ētre todos aquel día lleuo lo mejor Bunnifor de rosto que q̄ muy buen cauallero se mostro: y tras el Oliman de flandes y por ser ya noche uando el Rey tocar las trompetas y seçando las justas decendiendose de los miradores se fueron a los palacios: aun q̄ Adātileo yua muy triste por auer faltado en la auentura ante su señoza: pareciendolle q̄ por ello lo despreciaria teniendole en poca cuenta por ser uicido / mas no era así que la infanta Mercilana parciendole que por su amor auia prouado la auentura lo amo y crecio mas d̄de adelante llegando todos a palacio por ser ya horas las melas fueron puestas y se sentaron acenar d̄de fueron seruidos muy altamente de muchos y muy diuersos manjares. Acabada la çena vuo muchas danças en todo el palacio: en q̄ Adātileo danço cō su señoza cō tanta gracia y defenbultura que de los mirar grā plazer todos recebiā. Por ser ya muy tarde se fuerō a reposar / lleuado Adātileo en voluntad de salir otro dia a las justas y hazer tales co-

las q̄de todos mayormēte d̄ su señoza fue se preciado: y q̄ así emedarā el pasado.

Capitulo. iiii. Como se hizieron las justas: y Adātileo lleuo la honra de las yllas: viendo dos caualleros estraños q̄ a ellas vinieron y quien eran estos caualleros.

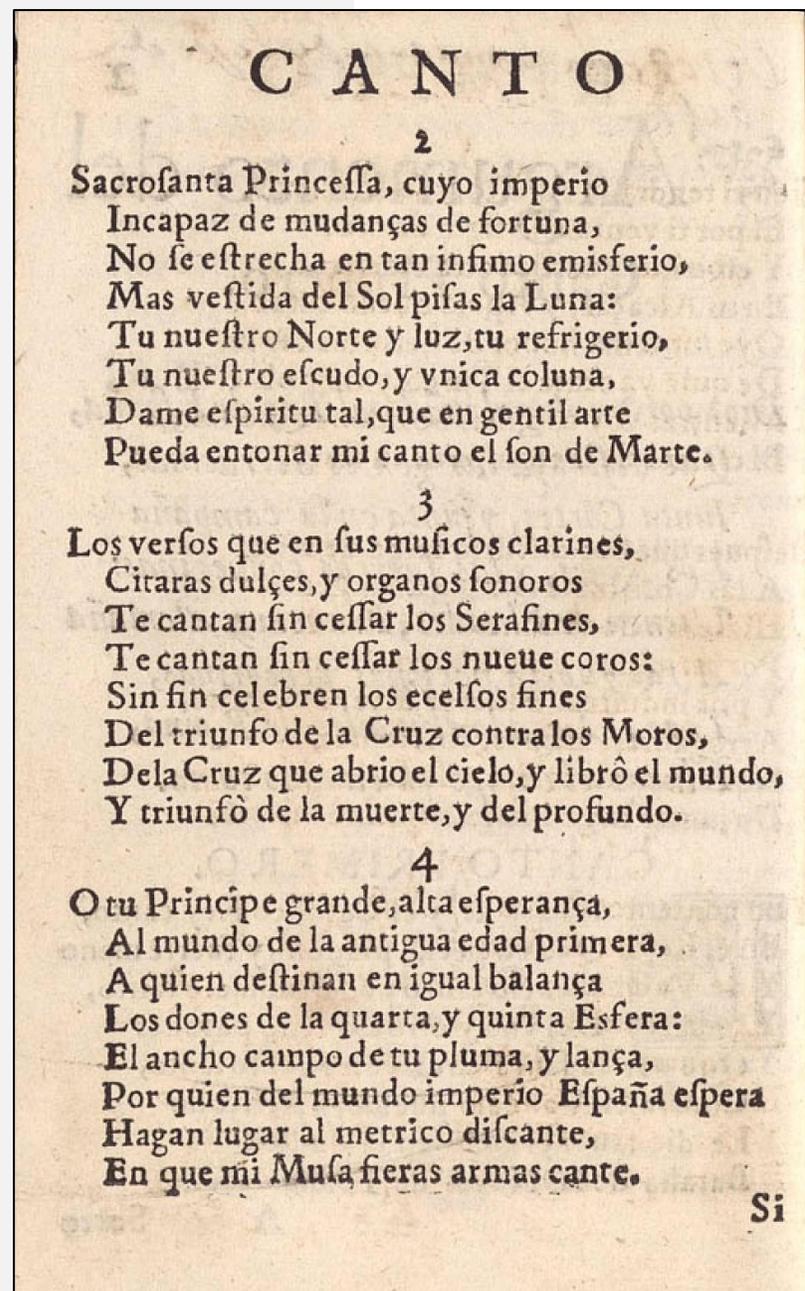
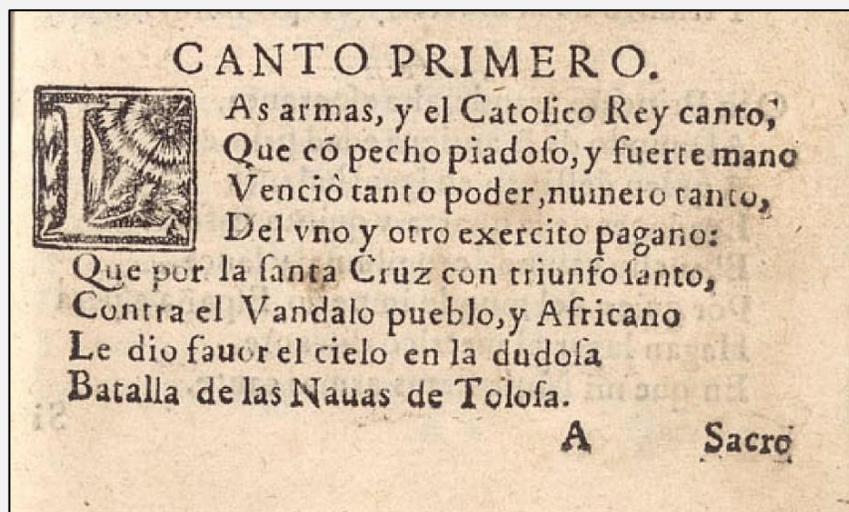


Como la mañana fue venida se leuanto el rey con todos sus altos hombres y se fue ala yglesia y despues d̄ ser oydo missa se tomo a jantar y auiendo comido se fueron a los miradores pa ver las justas: mas primero que se començasse mando el Rey a todos los caualleros q̄ ende estauan q̄ se pronassen con las ymages: mas de tāto s̄bed q̄ i. q̄ guardaua el escudo no se q̄so cōbatir cō ninguno d̄ los noueles: cōbatios con otros muchos caualleros: mas no vino tal q̄ media ora se pudiesse mātener cōtra ellas: y así se prouarō otros muchos con la ymage d̄ dios d̄ amor mas todos fuerō uicidos. Como el Rey esto vido mando q̄ se epeçassen las justas y luego se pusterō a vna pte por mātenedores Adustiel d̄ rosto q̄ y Bunnifor su hermano: y d̄ la otra Duradel d̄ cleue y Olimā d̄ flandes: y Adustiel d̄ rosto q̄ y Duradel d̄ cleue se vinierō vno cōtra otro heriēdo los cauallos d̄ las espuelas y ē medio d̄ la tela se d̄ierō tā rezios ēcuētros q̄ ābos vinierō al suelo. Bunnifor y Olimā cada vno d̄ seoso d̄ vgar a su cōpañero se vinierō

Al iiii

Spanish Round script (XVI-XVIIc)

book	source text	CER (Validation set)
Libro de los Siete Sabios de Roma Barcelona, Andreu, 1678	Madrid, Biblioteca Nacional de España, R/530	1.30%
Farol Indiano y Guía de curas de indios México, 1713	BX1757 P4 / Fondo Reservado UNAM-IIH	1.75%
Libro del Orlando Determinado Lérida, Prats, 1578	(proyecto ArDiTeHis)	1.11%



New insights: General models / Modelli misti

- General models are based on a defined number of pages of **different works**
- Transcriptions have to respect the same transcription criteria > **consistency**

Work 1 – n pages transcribed

Work 2 – n pages transcribed

Work 3 – n pages transcribed

Work 4 – n pages transcribed

Work 5 – n pages transcribed

....



HTR GENERAL MODEL

- multiple scripts / no single work prevails
- new similar documents can be interpreted (nearness to trained scripts)

SPANISH GOTHIC 15TH-16TH CENTURY



Spanish Gothic 15th-16th Century

By: Stefano Bazzaco

Spanish 15th, 16th Gothic Script 0.92 (CER) VIEW MODEL

<https://readcoop.eu/model/spanish-gothic-15th-16th-century/>

DOI: <https://doi.org/10.5281/zenodo.4888926>

Authors:

Federica Zoppi
Giada Blasut
Nuria Aranda García
Ángela Torralba Ruberte
Ana-Milagros Jiménez Ruiz
Pedro Monteiro
José Manuel Fradejas
Eduardo Camero Santos
Laura Lecina Nogués
Almudena Izquierdo Andreu

SPANISH REDONDA 16TH-17TH CENTURY



En defension de vna auentura estraña: Quanto el menester graue lo requie
Del qual por oy tratar no de termino, (O sea, ò no, tu subdito y vassallo),
Pues ya los quatro figuen su camino. Que en campo tres dias pueda sustentallo.
En corte en fin llegados, no quisieron Con esto tãbien pide, que aunque dado
es reves entrar con la donzella: Fuesse el dia primero al que parezca

ID: 33399

SPANISHREDONDA_XVI-XVII_EXTENDED (V1.0.0) - PRINT

Spanish Redonda (Round Script) 16th-17th Century

By: Stefano Bazzaco

Spanish 16th, 17th Round Script 1.07 (CER) [VIEW MODEL](#)

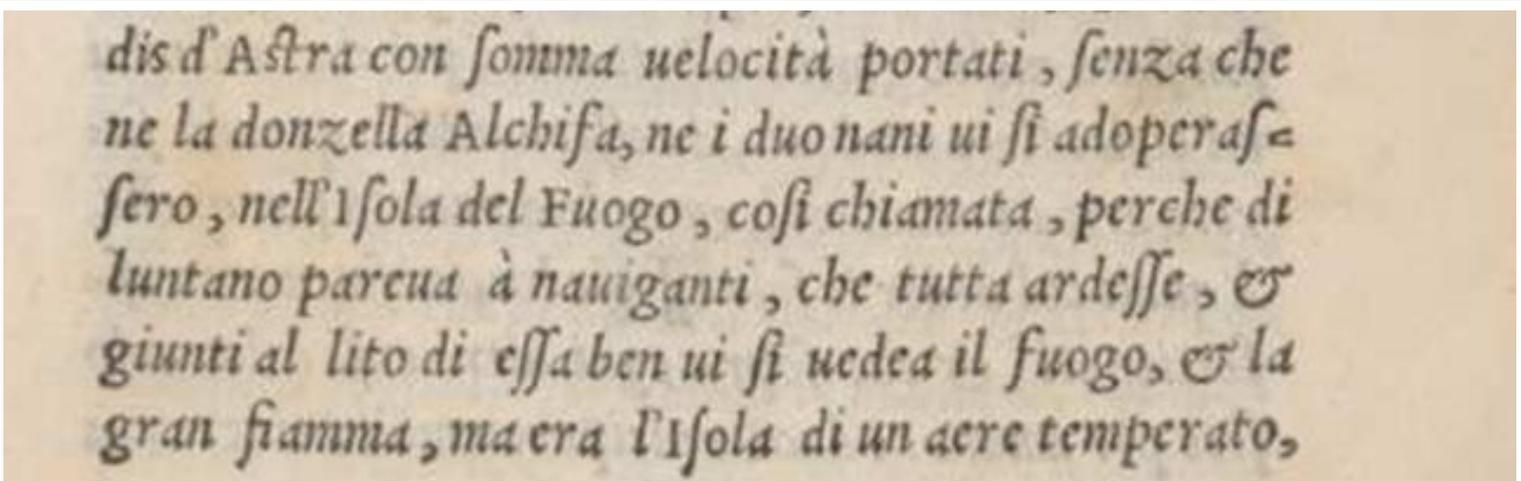
Authors:

Gaetano Lalomia
Daniela Santonocito
Manuel Garrobo Peral
Mónica Martín Molares
Carlota Cristina Fernández Travieso
Giulia Tomasi
Alessia Fichera
Soledad Castaño Santos
Almudena Izquierdo Andreu

<https://readcoop.eu/model/spanish-redonda-round-script-16th-17th-century/>

DOI: <https://doi.org/10.5281/zenodo.4889217>

Italics_VeniceXVIc



dis d'Astra con somma uelocità portati, senza che
ne la donzella Alchifa, ne i duo nani ui si adoperas-
fero, nell'Isola del Fuogo, così chiamata, perche di
luntano pareua à nauiganti, che tutta ardesse, &
giunti al lito di essa ben ui si uedeua il fuogo, & la
gran fiamma, ma era l'Isola di un aere temperato,

Language: Italian

CER: 0.70%

Authors:

Stefano Bazzaco (coord.)

Giulia Lucchesi

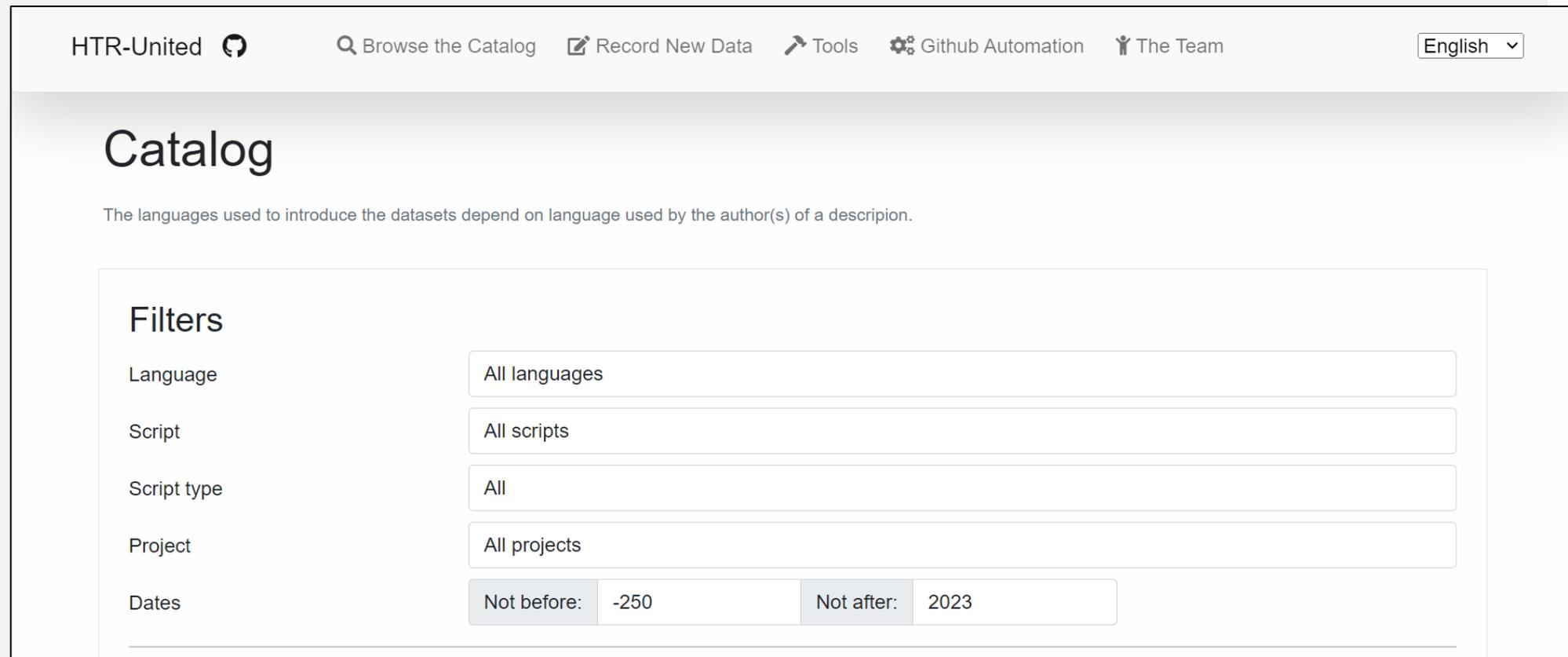
<https://zenodo.org/records/10674282>

DOI: 10.5281/zenodo.10674282

General models and the scientific community

Preservation of dataset:

- Export of data in standard formats XML / ALTO (Analyzed Layout and Text Object)
- Distribution among HTR colab projects: f.e. [HTR United](#) (Chagué et al. 2020)



The screenshot shows the HTR-United website's 'Catalog' page. The header includes the site name 'HTR-United' with a logo, navigation links for 'Browse the Catalog', 'Record New Data', 'Tools', 'Github Automation', and 'The Team', and a language dropdown set to 'English'. The main heading is 'Catalog', followed by a note: 'The languages used to introduce the datasets depend on language used by the author(s) of a description.' Below this is a 'Filters' section with the following options:

Filter Category	Selected Value
Language	All languages
Script	All scripts
Script type	All
Project	All projects
Dates	Not before: -250 Not after: 2023

Transkribus: recent advances (1)

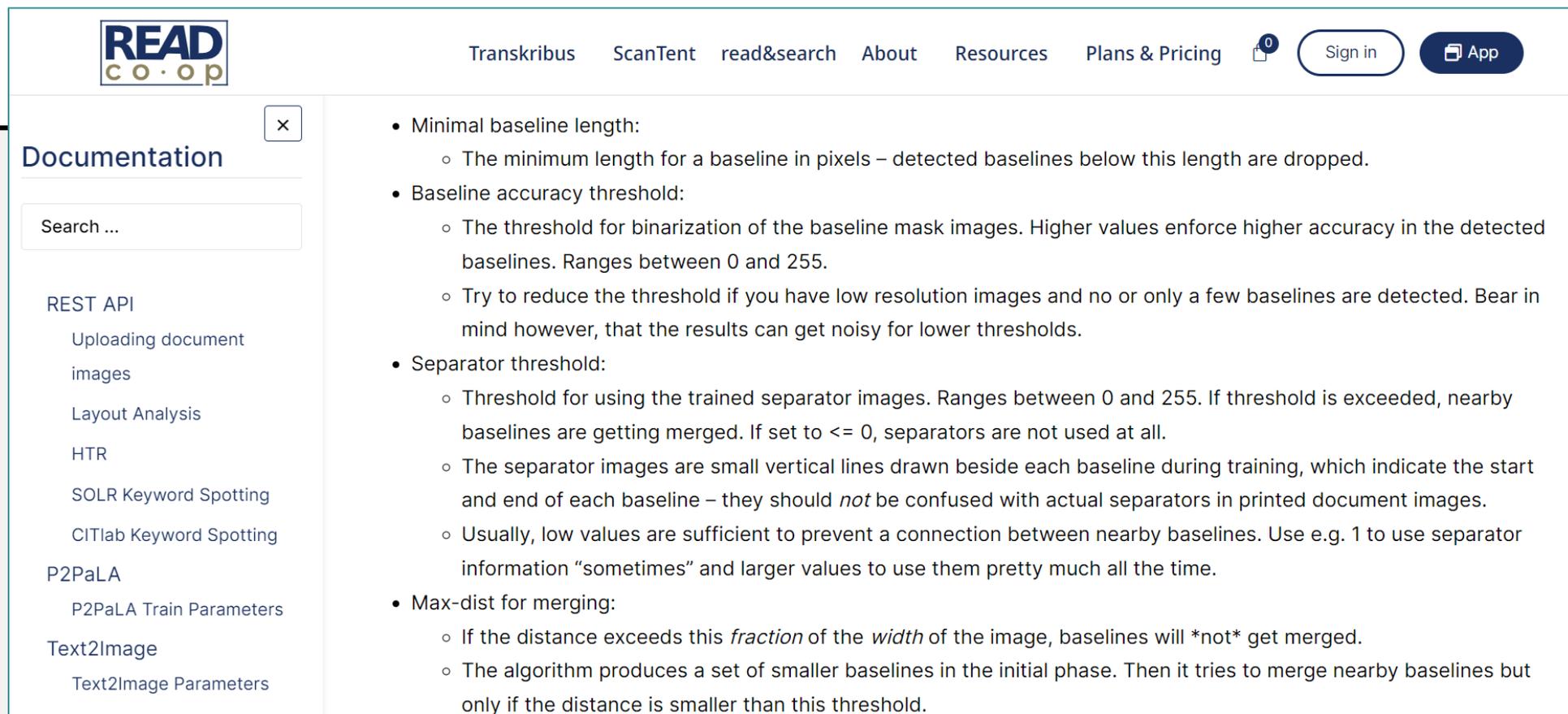
Transkribus APP

The screenshot displays the Transkribus web application interface. At the top left is the Transkribus logo, featuring a stylized bird. To the right of the logo are navigation links for 'Desk', 'Models', 'Sites', and 'Jobs', along with a user profile icon. Below this is a dark blue navigation bar with 'Home', 'Collections', and 'Tags' tabs. A search bar is positioned on the right side of this bar, containing the text 'Search text across all collections' and a 'Global Text Search' dropdown menu. The main content area begins with a personalized welcome message: 'Welcome to Transkribus, Stefano!'. To the right of this message are two buttons: 'Upload Files' and 'Quick Text Recognition'. Below the welcome message is a section titled 'Recent documents' with a sub-link 'Continue where I left >'. On the right side of this section is a 'See history >' link. Three document thumbnails are displayed in a row. The first thumbnail shows a historical illustration with the text 'EL MUSEO UNIVERSAL.' at the bottom. The second thumbnail shows the title page of a book: 'LIBRO PRIMERO, de la Diana de Iorge de Montemayor.' with a large decorative initial 'B' and the text 'Axaxa de las montañas de Leó el olvidado Sireno, a quien Amor, la fortuna, el tiempo, tratan de manera, q del'. The third thumbnail shows the masthead of a newspaper: 'LA ACTUALIDAD. PERIÓDICO POLÍTICO' with sub-sections for 'Liberal', 'Tárraco', and 'Progres'.

Transkribus: recent advances (2)

LA configuration

from CITLab to Transkribus LA <https://readcoop.eu/transkribus/docu/layout-analysis-help/>



The screenshot shows the Transkribus web application interface. At the top, there is a navigation bar with the READ COOP logo on the left and links for Transkribus, ScanTent, read&search, About, Resources, and Plans & Pricing on the right. There are also buttons for 'Sign in' and 'App'. Below the navigation bar, a 'Documentation' sidebar is visible on the left, containing a search bar and a list of documentation topics: REST API (with sub-items for Uploading document images, Layout Analysis, HTR, SOLR Keyword Spotting, and CITLab Keyword Spotting), P2PaLA (with sub-item for P2PaLA Train Parameters), and Text2Image (with sub-item for Text2Image Parameters). The main content area on the right displays the 'LA configuration' page, which lists several parameters for layout analysis:

- Minimal baseline length:
 - The minimum length for a baseline in pixels – detected baselines below this length are dropped.
- Baseline accuracy threshold:
 - The threshold for binarization of the baseline mask images. Higher values enforce higher accuracy in the detected baselines. Ranges between 0 and 255.
 - Try to reduce the threshold if you have low resolution images and no or only a few baselines are detected. Bear in mind however, that the results can get noisy for lower thresholds.
- Separator threshold:
 - Threshold for using the trained separator images. Ranges between 0 and 255. If threshold is exceeded, nearby baselines are getting merged. If set to ≤ 0 , separators are not used at all.
 - The separator images are small vertical lines drawn beside each baseline during training, which indicate the start and end of each baseline – they should *not* be confused with actual separators in printed document images.
 - Usually, low values are sufficient to prevent a connection between nearby baselines. Use e.g. 1 to use separator information “sometimes” and larger values to use them pretty much all the time.
- Max-dist for merging:
 - If the distance exceeds this *fraction* of the *width* of the image, baselines will **not** get merged.
 - The algorithm produces a set of smaller baselines in the initial phase. Then it tries to merge nearby baselines but only if the distance is smaller than this threshold.

Transkribus: recent advances (3)

Field Models: Complex LA models

<https://help.transkribus.org/field-models>

included marginalia, tables, etc.

Field Recognition Model

< Back

Feedback

Training Data Tag Selection Validation Data Model Setup Summary & Start

Select	Title
X	marginalia

Next >

1 documents Selected

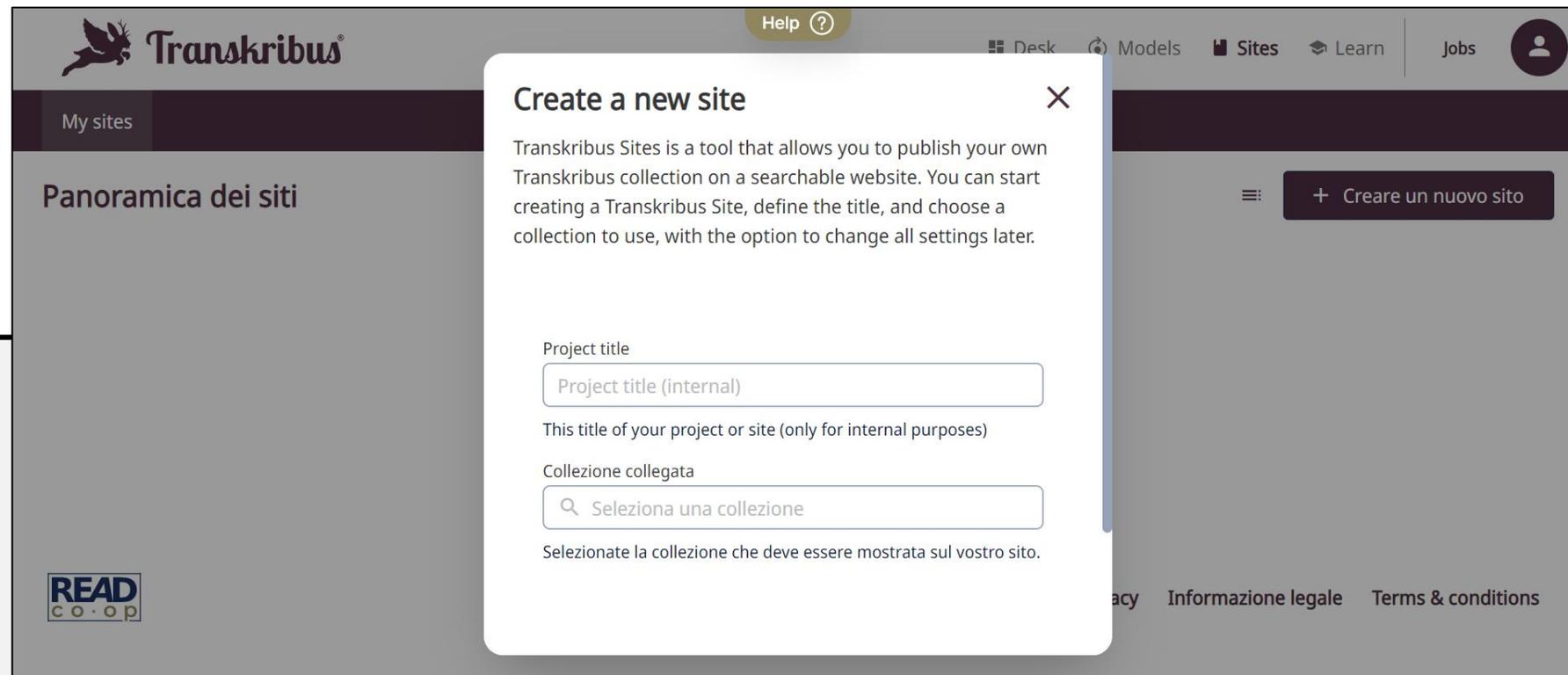
Transkribus: recent advances (4)

Transkribus Site + information extraction

New platform to easy publish processed documents

User can perform information extraction

<https://www.transkribus.org/sites#features>



The screenshot displays the Transkribus web interface. At the top, the Transkribus logo is on the left, and navigation links for 'Desk', 'Models', 'Sites', 'Learn', and 'Jobs' are on the right. A 'My sites' tab is active. A modal dialog titled 'Create a new site' is open in the center. The dialog contains the following text: 'Transkribus Sites is a tool that allows you to publish your own Transkribus collection on a searchable website. You can start creating a Transkribus Site, define the title, and choose a collection to use, with the option to change all settings later.' Below this text are two input fields: 'Project title' with the placeholder 'Project title (internal)' and a subtext 'This title of your project or site (only for internal purposes)'; and 'Collezione collegata' with a search icon and the placeholder 'Seleziona una collezione'. A subtext below the second field reads 'Selezionate la collezione che deve essere mostrata sul vostro sito.' In the background, a button labeled '+ Creare un nuovo sito' is visible. The bottom of the page features the 'READ CO-OP' logo and links for 'Privacy', 'Informazione legale', and 'Terms & conditions'.

Transkribus: sustainability and collaborative perspective

- **Solid technology:** Transkribus is based on Machine Learning technology > recognition grows with the number of processed documents by the whole users' community
- **Collaborative platform:** based on a Growing User Network, which provides:
 - new extended HTR recognition models
 - new Layout/Field models
- **Related infrastructures and tools:**
 - Scan tent + DocScan app* (portable digitalization tools)
 - Transkribus App* (web browser version)
 - Transkribus Learn* (to train young transcribers)

Bibliografia di riferimento

- Bazzaco, S. (2024). «La trascrizione automatica di documenti a stampa antichi. Appunti per un modello di riconoscimento della tipografia in corsivo», *DigItalia*, vol. 19, n. 1. <https://doi.org/10.36181/digitalia-00094>
- Hodel, T., D. Schoch, C. Schneider e J. Purcell (2021). «General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example». *Journal of Open Humanities Data*, vol. 7 (July). <https://doi.org/10.5334/johd.46>
- Mühlberger, G., et al. (2019). «Transforming scholarship in the archives through Handwritten Text Recognition. Transkribus as a case study». *Journal of Documentation - Emerald Publishing*, 75/5, 954-976.
- Schwarz-Ricci, V. I. (2022). «Handwritten Text Recognition per registri notarili (secc. XV-XVI): una sperimentazione». *Umanistica Digitale*, n. 13, pp. 171-181. <https://doi.org/10.6092/issn.2532-8816/14926>
- Spina, S. (2022). «Historical Network Analysis and HTR tools for a digital methodological historical approach to the Biscari Archive of Catania». *Umanistica Digitale*, n. 14, pp. 163–181. <https://doi.org/10.6092/issn.2532-8816/15159>
- Terras, M. (2022). «Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription». En Jaillant, S. (ed.) *Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections*. Verlag - Bielefeld University Press, 179-204.



UNIVERSITÀ
di **VERONA**
Dipartimento
di **LINGUE
E LETTERATURE STRANIERE**



STEFANO BAZZACO
stefano.bazzaco@univr.it