



UNIVERSITÀ
di VERONA

Dipartimento
di LINGUE
E LETTERATURE STRANIERE
Scuola di dottorato
in SCIENZE UMANISTICHE



Text Recognition: OCR/HTR softwares for Humanities



PROGETTO
MAMBRINO

A B C D H

Stefano Bazzaco

stefano.bazzaco.1@gmail.com

EnExDi2022
University of Poitiers – 10/05/2022

READ
co · op

Transkribus

Seminar structure

- introduction to OCR/HTR softwares
(digitalization, brief history of OCR/HTR, state-of-the-art)
- Digitalization and image preprocessing
- OCR/HTR softwares (presentation and evaluation)
- *Transkribus in theory*
(basic features, general workflow, results, costs, advanced features)

In the beginning there was... DIGITALIZATION

Last 40 years > **the formats of our collective memory are facing great changes**

Digitalization process changed in a significant way the relation between scholars of Humanities fields and their specific object of study.

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

Stated 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 Digitalization:** in the 1990s digitalization efforts significantly increased (different forces: changes in public policies, networked technologies, institutional funds > large scale, ambitious projects)
- **Rise of Digitalization programmes:** from 2000 digitalization 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)

Digitalization 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 *digitalized object* (scanned image) into an *electronic encoded text* (machine readable form / sequence 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 and extensions):



In the very beginning there was... TEXT RECOGNITION

Optical Character Recognition is a sub-field of **AI 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

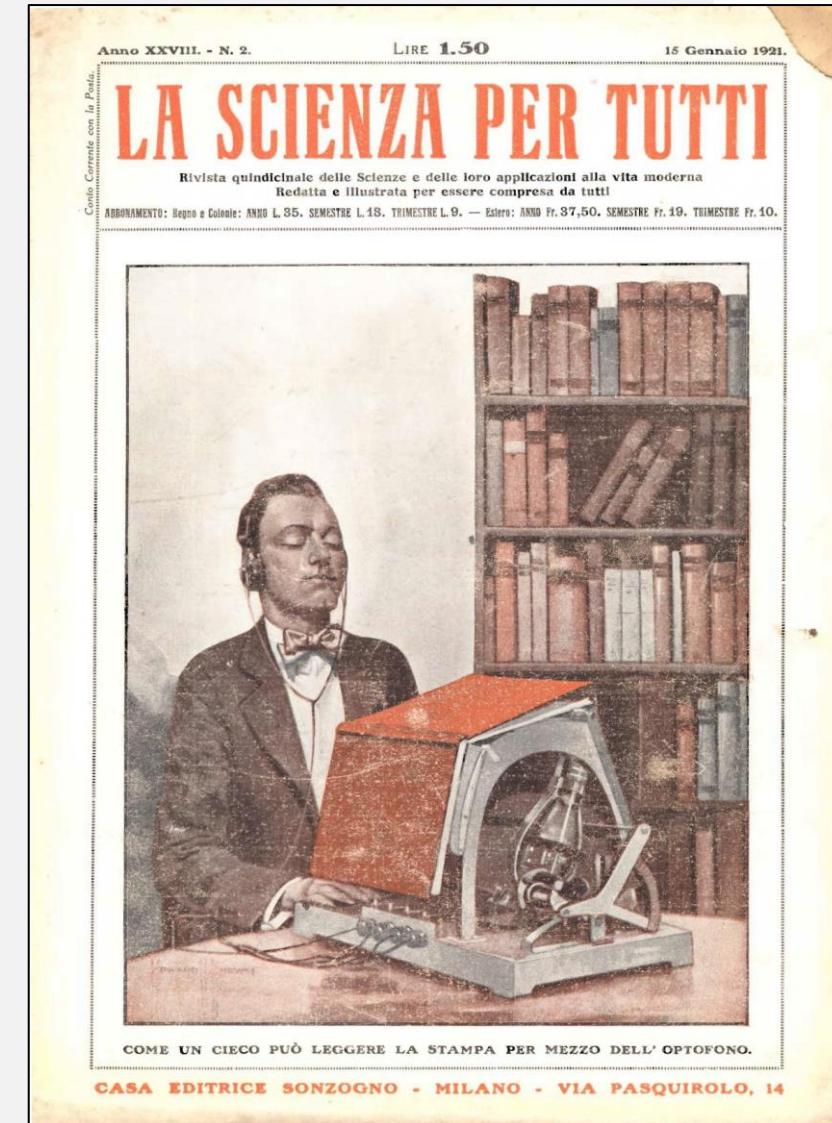
Overview [edit]	
Time period	Summary
1870–1931	Earliest ideas of optical character recognition (OCR) are conceived. Fournier d'Albe's Optophone and Tauschek's Reading Machine are developed as devices to help the blind read. ^[1]
1931–1954	First OCR tools are invented and applied in industry, able to 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

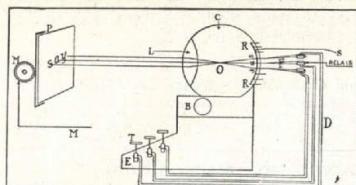
A brief history of Optical Character Recognition (OCR)

First steps:

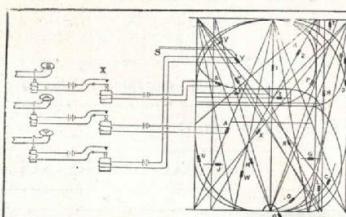
- 1870 Charles R. Carey invents the **retina scanner** (an image transmission system that used photocells)
- 1885 P. Nipkow invents the **Nipkow Disc** (an image scanning device; breakthrough for several devices, f.e. television)
- 1912 E. Fournier d'Albe conceived the **Optophone** (a scanner that produced 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'orizzontalità; *L*, scappamento elettrico che regola lo spostamento del foglio; *O*, rotolo 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 eliosioide, 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 primari.



Capitali azionanti ognuno un'elettricalamita (*E*) comandante un tasto (*T*); *B*, rotolo su cui è avolta 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 oscurete dall'immagine della lettera.

Dare « un occhio » alla macchina da scrivere e farle leggere e copiare uno scritto senza bisogno, diciamo così, della « dettatura della ditta »! — 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 nella sua forma è un punto caratteristico che non si confonde con nessun'altra lettera. Giò, 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 sférica 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 elicatica, 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, qualunque

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'eliosioide.

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 un'elettricalamita, 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 presenti 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 impresso soltanto la 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 quello 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 elettricalamita 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 sfilaria 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 esse 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 della forma. Ecco basta perché i punti caratteristici corrispondenti della retina e dell'immagine

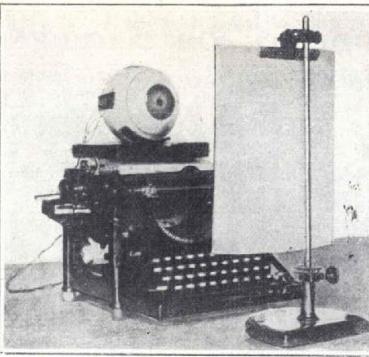
I CACCIAATORI DI SOMMERGIBILI

La guerra dei sommergibili inaugurata dalla Germania ha provocato, come tutte le guerre 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 passare da parte a parte quasi sfuggendo di sommontarle. Mobilitissimi e quindi capaci di scansare facilmente le torpedini: armati di un cannonecino sufficiente a produrre in un sommersibile un foro per cui affonda, se è immerso o non può più immergersi se sornuato.

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 sommersibili 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 contrarie misure che con l'andar del tempo si saranno prese. Ad esempio, nel nostro numero del 1° febbraio annunciammo che 40 cacciatori di sommersibili erano stati fabbricati in America per l'Inghilterra, e molti altri nei cantieri inglesi; ma d'allora in poi, il loro numero è grandemente cresciuto e crescerà ancora, perché nessuno si fidà 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 molti numerosi che le seconde.



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

gine non s'incontrino più: così avverrebbe, ad esempio, fra una lettera minuscola e la stessa lettera maiuscola. Che se a ciò si può rimediare complicando maggiormente la retina, sorge il problema dei caratteri stampati — sia pure quelli comuni di testo — i quali, pur non essendo numerosi nei loro aspetti generali (romano, elzevir, bodoniano, ecc.), si complicano per le proporzioni rispettive fra la larghezza e l'altezza d'ogni lettera, i corivi e i neretti. Sorgerebbe anche il problema di regolare il movimento del leggio, riducendolo forse a far sfilarie solo mezza lettera per volta, salvo alla retina trovare nell'una o nell'altra il punto caratteristico. Perché, mentre nella macchina da scrivere ogni lettera occupa il medesimo spazio, dalla 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* e le maiuscole in genere salvo *I*, *J*, che sono della grandezza normale per le maiuscole); altre quasi due volte, (*æ*, *œ*, *M*, *W* ed anche più *Æ*, *Œ*).

Quanto allo scarto a mano, non è nemmeno da pensare a riprodurlo in tal modo. Perciò l'utilità immediata dell'invenzione è discutibile. Pure, nessuno potrebbe negare il pregio della genialità; e nessuno può escludere che, come già avvenne altre volte per novità che parvero follie, 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.

We can talk about OCR systems referring to their actual meaning only from the 1940s (OCR for commercial purposes)

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 linotypes (or fonts) / first standardized graphic system for commercial use (OCR A)

1970s – **Second Generation of OCR hardwares:** recognition extended to other printed linotypes 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)

OPTICAL CHARACTER RECOGNITION and DH

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 relational databases)**
- ❖ **Expansion of Text Mining projects**

(quantitative text studies: *algorithmic query, Natural Language Processing, Stylometry, 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

From the very beginning, humanists noted some problems concerning OCR reliability, for 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)

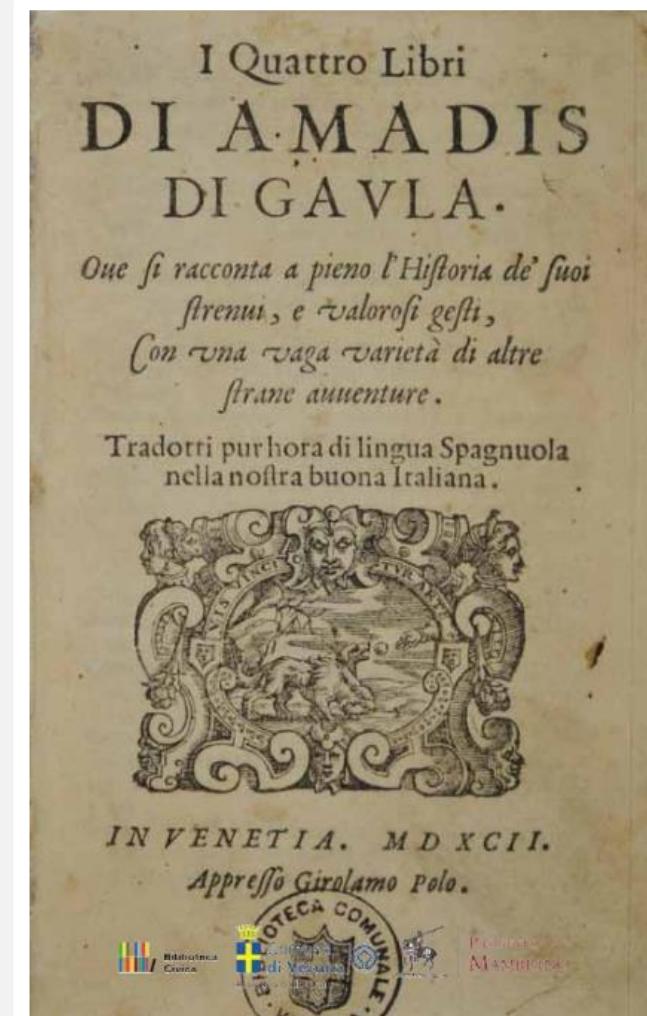
Progetto Mambrino

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

Objectives: book census, libraries digitalization programmes, bibliographical databases, (recently) **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.



Progetto Mambrino: perspectives in DH field

Mapping chivalry: digital archives based on the REPERTORIOs. The elaboration of these data (summaries, characters, places, themes motifs) will produce different sets of maps, on different levels. The results of this mapping process will be retrieved, processed and displayed in the **DL** and in the **DSE**.

Automatic text transcription: testing and training of an OCR / HTR technology for Manuzio italics font. Collaboration with Transkribus project. The OCR / HTR model has been widely implemented, tested and trained in order to start the systematic transcription of the corpus.

Modelling: before entering the **DL** resulting texts will be coded and marked according to the XML /TEI standard. For this reason we are developing a specific XML/TEI marking model.



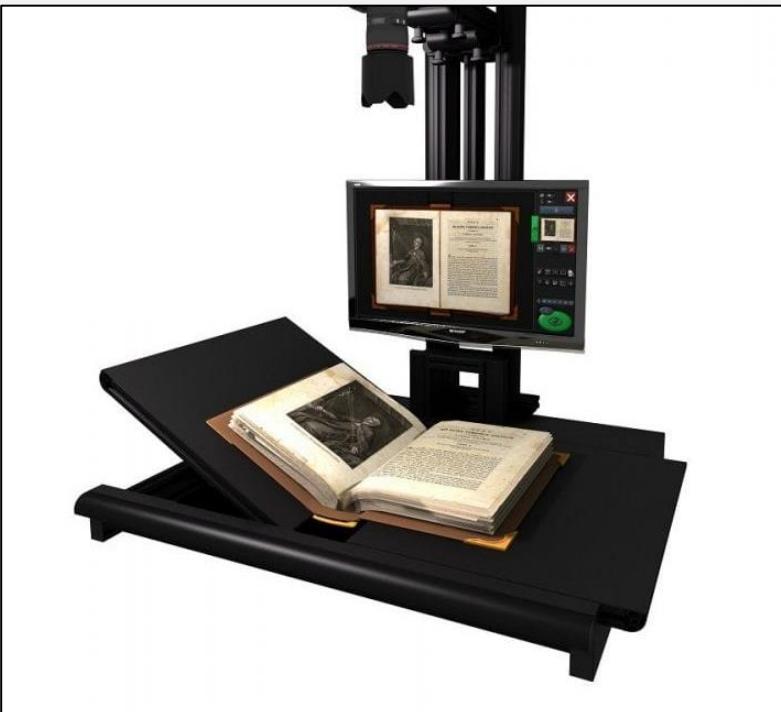
Bibliographic research: transformation of the existing bibliographic census “Spagnole Romanzerie” in a dynamic bibliographic DB integrated in the **DL** that will also share data with digital aggregators as Red Aracne and Europeana.

Digital Scholarly Editions (DSE) are the most important part of the **DL**. Front-end display will allow a parallel view of the text (according to the TEI standards) and the images of the original edition. DSE shall also retrieve and display information from the other data-set expressly created.

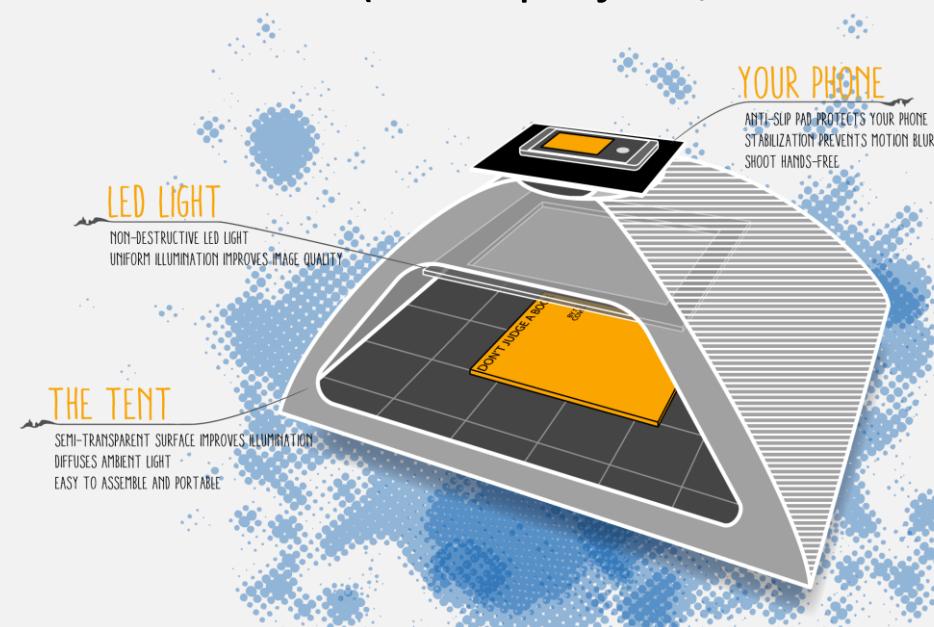
1. DIGITALIZATION

Consists in the **scanning of the textual materials**

Usually this operation is undertaken by libraries with high precision scanners



Recently handly personal scanners and other digitalization instruments have been developed
f.e. Scan Tent (READ project / Transkribus)

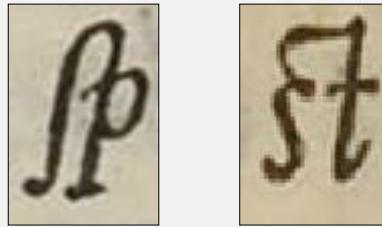


2. Problems with AUTOMATIC TRANSCRIPTION

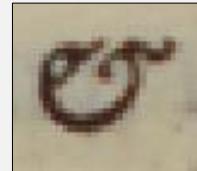
Specific issues of historical writings

f.e. in Italics scripts:

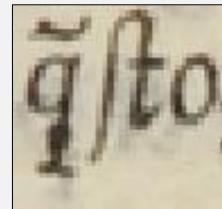
LIGATURES



TIRONIAN NOTE



ABBREVIATIONS



State of preservation of the sources

WARPED PAGES

uenit in mentem.
lum.
tatem, sine præpositione

INK TRANSFER

Tu dormis

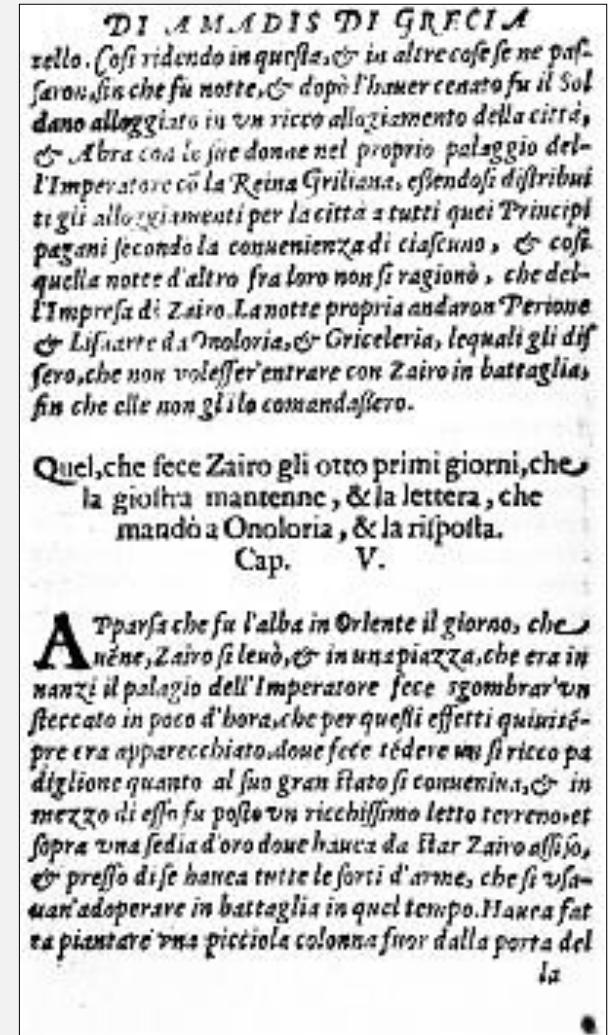
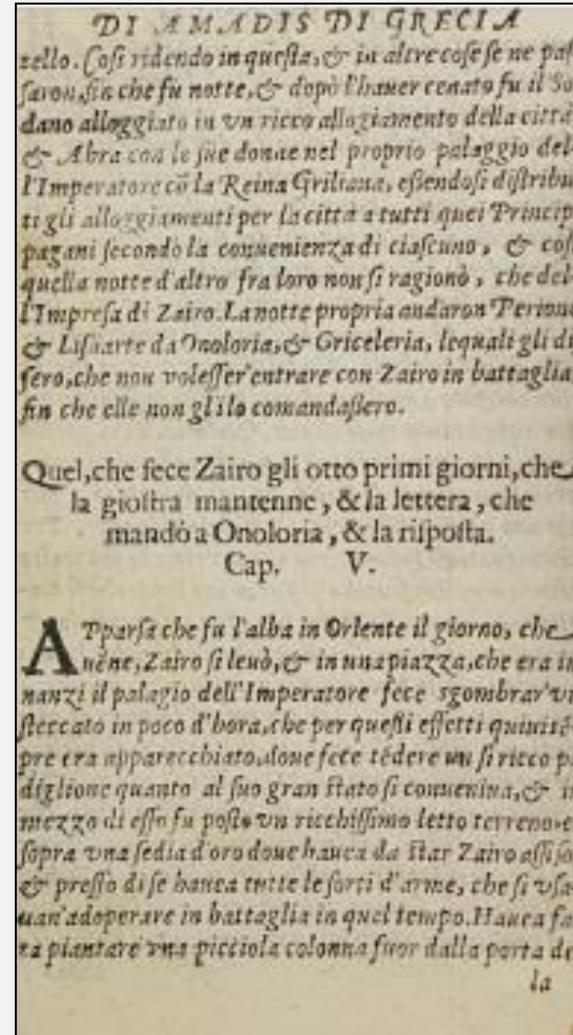
STAINS

mini hauendoj usurpata la tana de li più antichi e bravi leoni, ne faranno con crudeli beccate e morte di molti di loro cacciati via con l'aiuto de gli duo orsi piu sauij: E finalmente i paseri bianchi con l'aquila reale, e molte altre aquile minori, insieme col Gallo incoronato uerranno in soccorso de gli affitti leoni: E con questa uenuta si causerà l'ultima rouina de li corvi marini: E poi che cosi è, per gli iddi ui prego, che non ui induciate tal cosa nel core: e non crediate, che il timore de la morte mi faccia dire q'sto per che mai ne le guerre grandi e spaueteuoli ne temetti, quando il Re Armato andò sopra Costantinopoli con nō minore esercito di quello, che potrete condurre uoi: Credatemi o alto Re, che io ueggo il Dio Marte in opposizione con Saturno, ischermire con una spada contra di uoi: E se pure tanto ui è à core la uenetta de le ingiurie uostre e nostre, come di uostri uasalli, aspettate un' altro miglior tempo: e per che crediate, che io non ui dico bugia; aspettate che questi duo pianeti giungano nel mezzo de la opposizione loro, e uedrete il mio fine e essere uenuto. E detto questo si assise tuttaua lamentandosi assai dolorosamente. Furono molti, che si accostarono al consiglio del sauij Re di Cal-

3. PREPARATORY OPERATIONS (PREPROCESSING)

OCR / HTR results largely depends on the quality of scanned images

- Layout management:**
page splitting, page orientation, deskewing, selection of text regions
- Noise removal/reduction:**
bad light effects, contrast, despeckling, dewarping
- (in some cases) **Bynaryzation:**
bi-tonal images



4. CHARACTER RECOGNITION PROCESS

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

open access

Windows, Linux, MacOS

single characters recognition

training: per glyphs

output: .txt .doc .pdf .xml .html

ABBYY Fine Reader

proprietary

Windows, Linux, MacOS

single characters recognition

training: per glyphs

output: .txt .doc .pdf .xml .html

Tesseract

Created by Hewlett Packard (1984-94)

Released in 2005 as an open source platform

Implemented by Google, after the launch of Google Book Search

Tesseract can be trained by each single glyph and has been extended to more than 100 different languages, but still have problems with Historical texts

The main issue is the constant **variability of glyphs** (fluctuation):

- different graphical representation of the same character (even inside the same source text)
- scanning distortions (skewing, blurring, ...)



Tesseract OCR

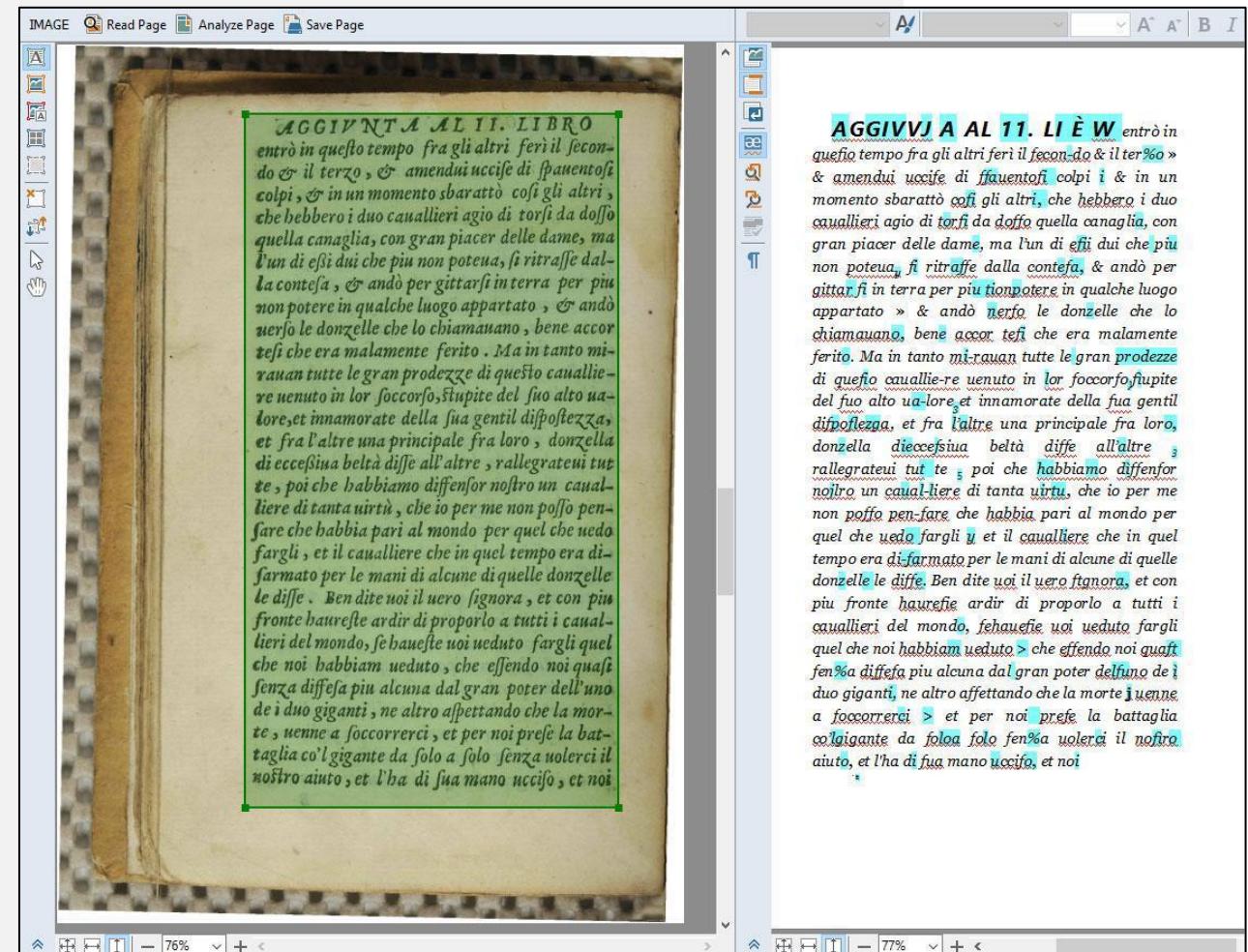
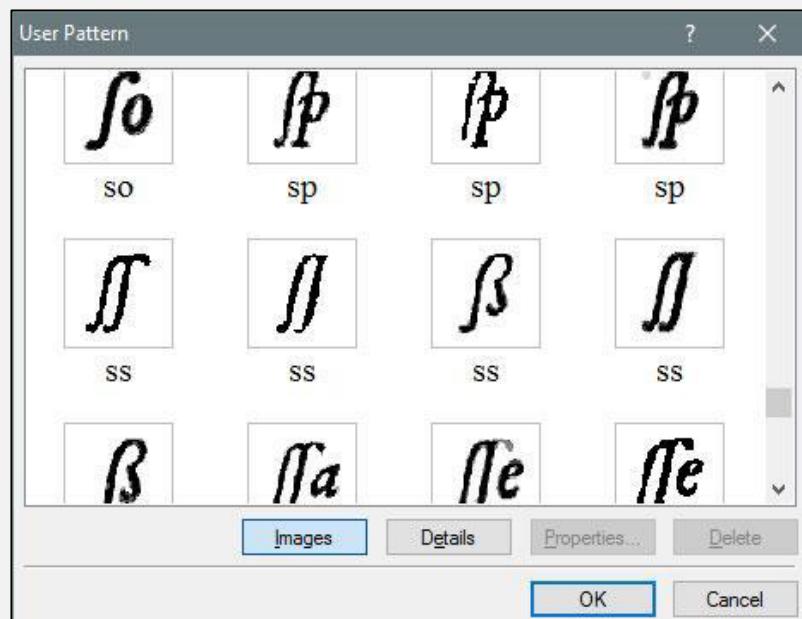
github.com/tesseract-ocr/tesseract

ABBYY FineReader

Specifically conceived for pdf's recognition

Includes an efficient **Layout Analysis** system

Can be trained, but not assures good transcription results with Historical Texts
(character error rate around 20-30%)



OCR / HTR softwares for Historical / Handwritten texts

OCR4All

open access

Linux
VM for Windows and MacOS

Recognition: characters in context

Golden Standard training
neural networks and LSTM

output: .txt .xml

Transkribus (READ)

open access / from 2020
recognition costs expected

Windows, Linux (VM), MacOS

Recognition: words in context

Golden Standard training
deep learning neural networks

output: .txt .doc .pdf .xml .html

eScriptorium

open access (?)

Windows, Linux, MacOS

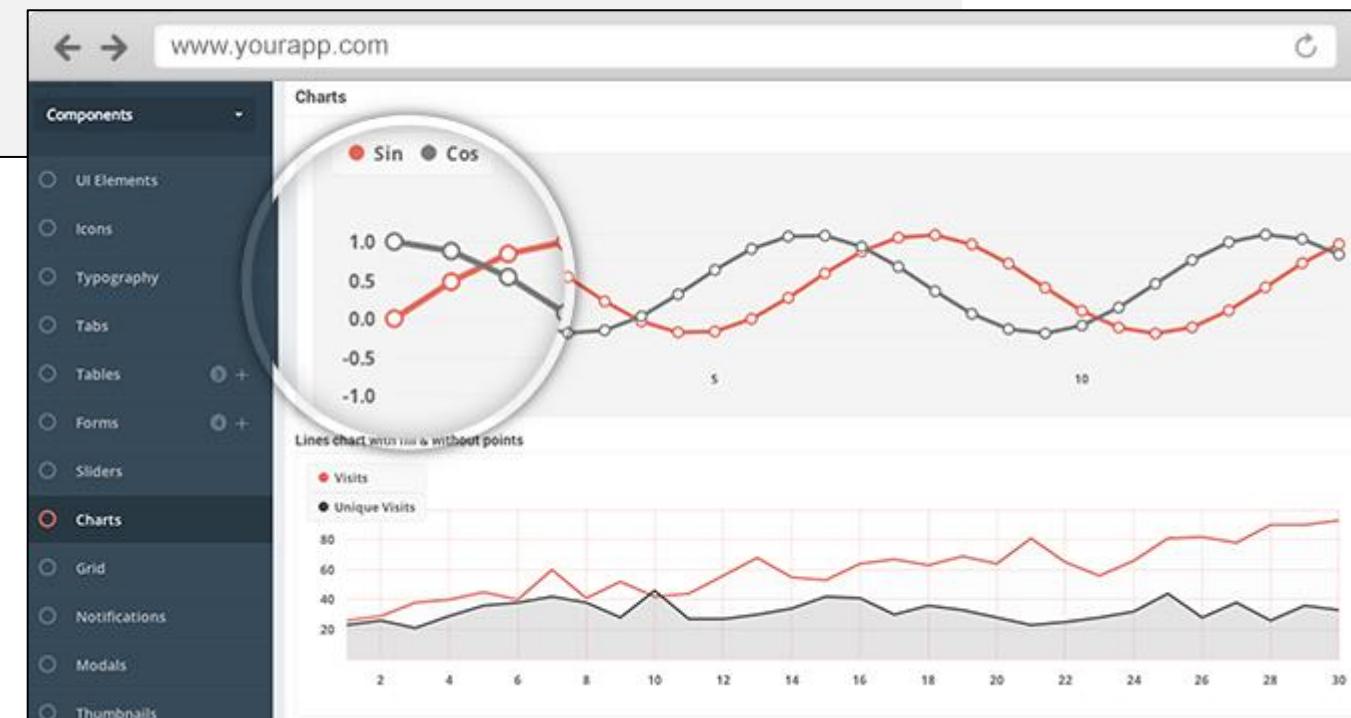
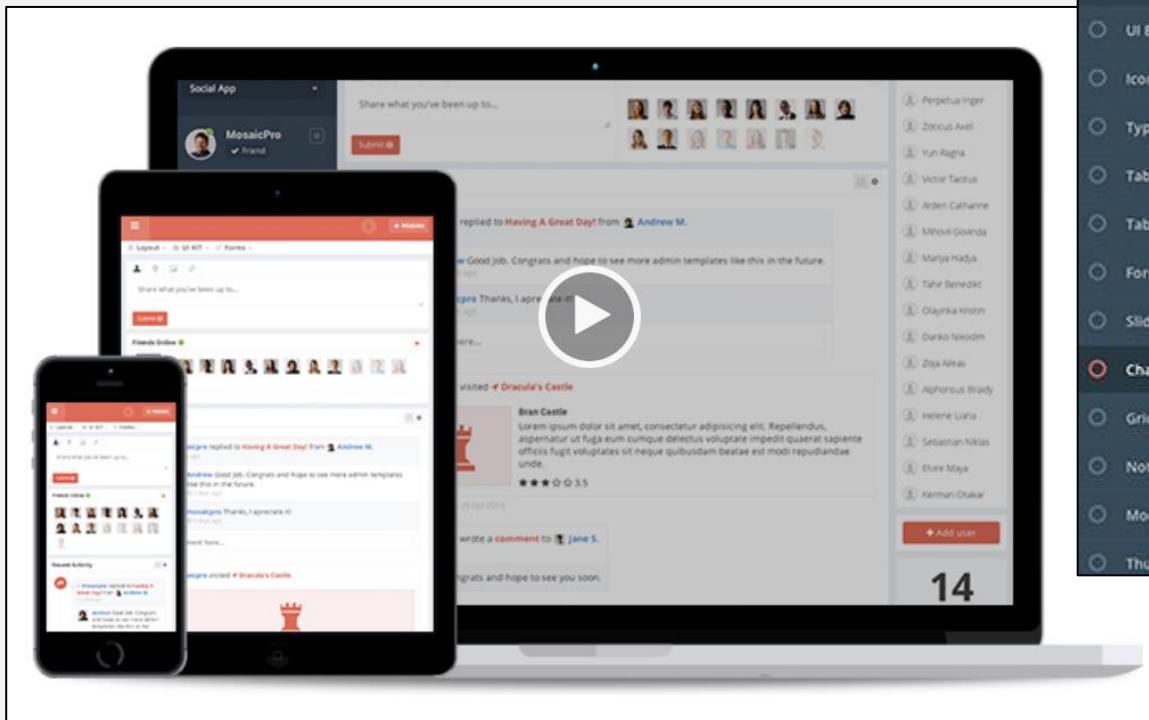
/

Golden Standard training
deep learning neural networks

output: rtf file

Developed by Roger Musings (blog: <https://escriptoriumblog.wordpress.com/>)

Not available, site work-in-progress



OCR4All

<https://github.com/OCR4all>

Developed by the University of Würzburg Centre of Philology and Digitality

Released in 2018 as an open source platform

Integrates the features of past open source platforms
(OCRopus / OCRopy, Kraken e Calamari)



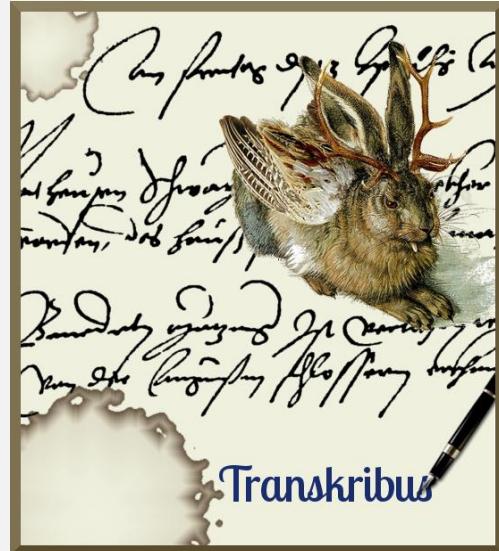
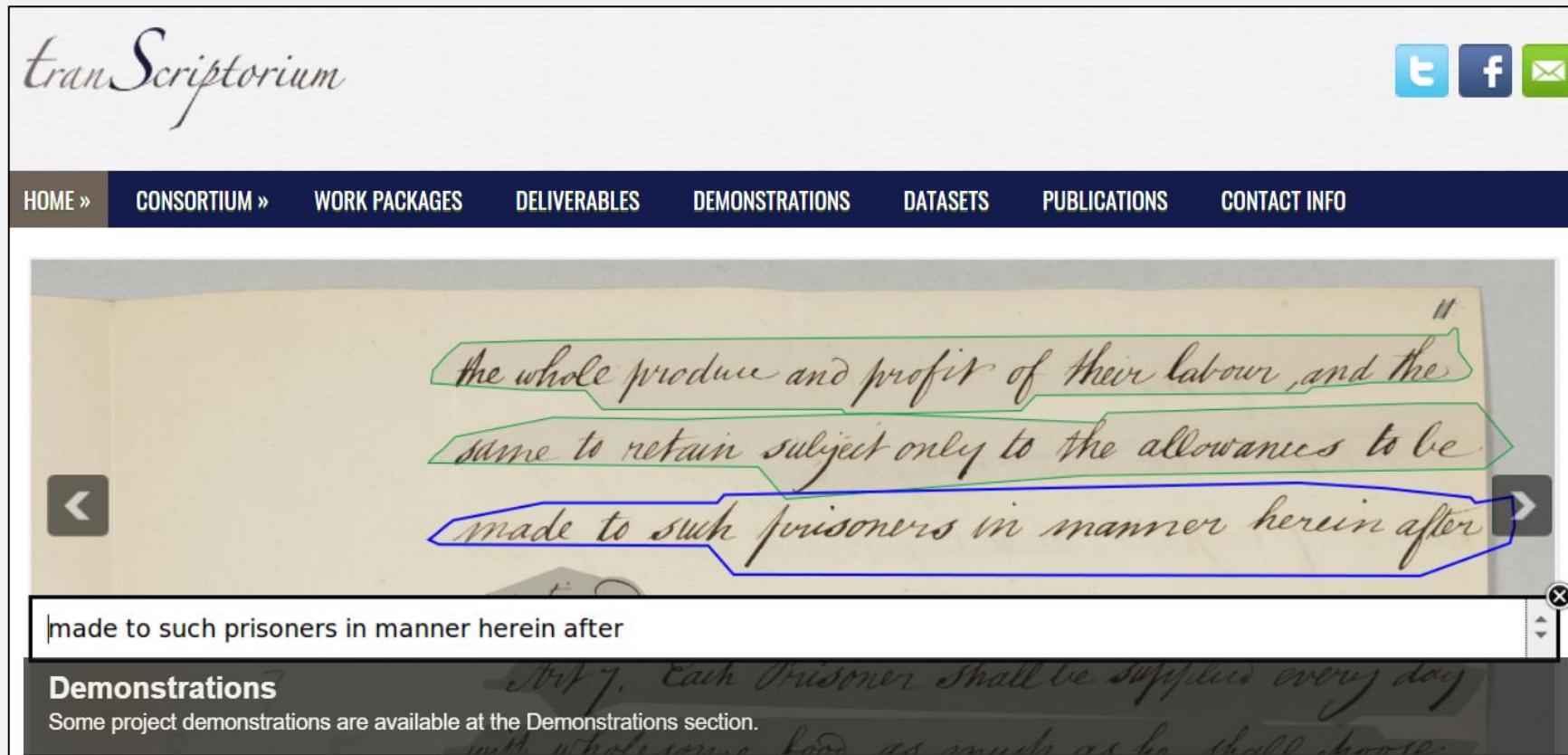
pros	cons
Open Access software Neural Networks and LSTM training: Good transcription results (less than 7% CER for printed texts)	By default works only on Linux OS (VM per Win, MacOS) It requires command line knowledge
OCR models are stored locally (user personal objects)	Not stable (published in 2019)

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



HTR / OCR

Transkribus is an **Handwritten Text Recognition (HTR) software**

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

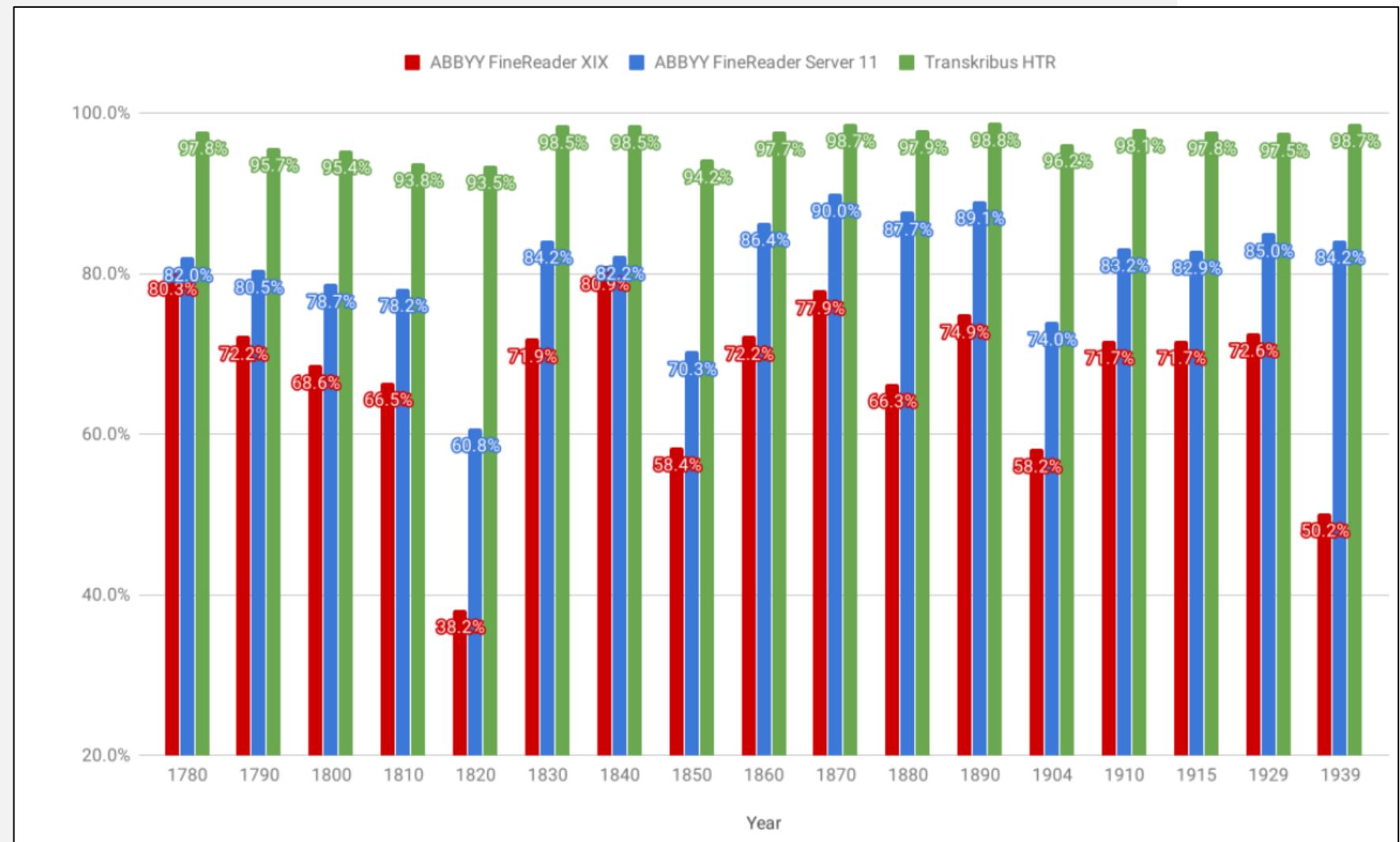
OCR	HTR
Focus on single characters	Focus on both single characters, words and context (sentence-based with an n-gram)
Preferably bi-tonal images (bynaarized)	Preferably clear background, but also full colour and greyscale images assures good results
Language/character variability can cause trouble	Less troubles with language/character variability
Trained-fixed tools	Permits to create individual or extended 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 results)



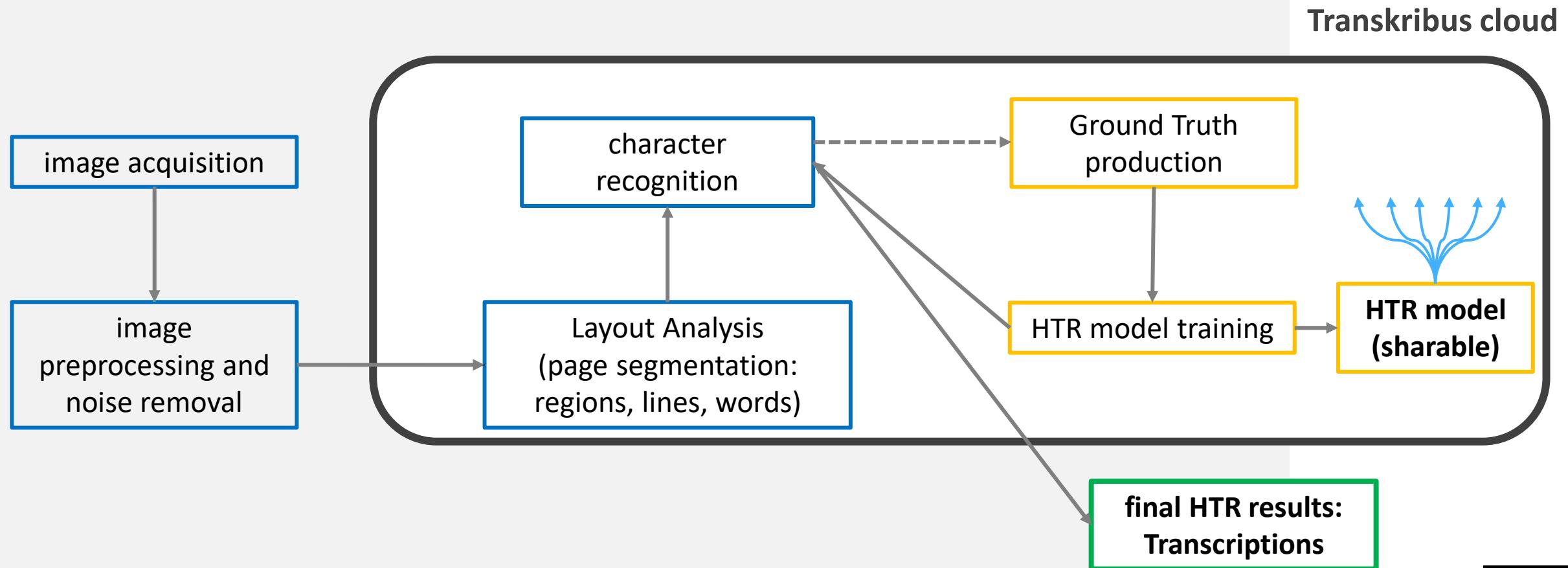
<<https://dev.clariah.nl/files/dh2019/boa/0694.html>>

Main Transkribus pipeline

- 1. Image import:** upload image preprocessed files to the cloud
- 2. Layout Analysis:** page segmentation (automatically: regions, lines / manually: words)
- 3. Groud Truth production:** manual transcription of a text portion (1500-2000 words)
- 4. Model Training:** creation of a text recognition model (HTR, HTR+, PyLaya)
- 5. Text Recognition:** automatic transcription with a suitable model
- 6. Export:** get an encoded text transcription (in different formats: XML, txt, word, pdf)

Transkribus workflow

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



Transkribus interface resembles a transcriber work desk

Transkribus v1.6.0 (28_02_2019_14:59), Loaded doc: LeandroESP1, ID: 96141, Page 2, file: 0005_1L.tif [Image Meta Info: (Resolution:600.0, w*h: 4588 * 6828)] [current line: w*h: 3476 * 129]

Server Overview Layout Metadata Tools

Logout stefano.bazzaco.1@gmail.com

Document Manager User Manager

Versions Jobs

Recent Documents... User activity

Collections: ESPLeandro (26094, Owner)

1-4 / 4 1 1 1 1 1 1 Doc-ID

ID	Title	Pages	Uploader	Uplo...
138468	TRAINING_TESTSET_Gothic1_Leandro	2	stefano.bazzac...	Wed
137626	TRAINING_TESTSET_GothicEsp1500_1	1	stefano.bazzac...	Mon
96142	LeandroESP2	137	stefano.bazzac...	Tue
96141	LeandroESP1	133	stefano.bazzac...	Tue

Epistola.

Epistola en la qual el Auctor dirige la obra

presente al Illustríssimo y muy excedente señor don Juan Claros de Guzmán, Conde de Niebla e c. Primogénito del muy exelente señor don Juan Alonso de Guzmán Duque de Medina Sidonia, e c. mi señor.



La causa mas principal muy exelente señor por donde nos comovemos a amar y dessear seruir a vn principe o gran señor. Es por su misma benibolencia a fa bieidad: cõ que acostúbra tratar a los inferiores suso agra deciendo los pequeños servicios con crecidas obras de benevolencia: tomando exemplo en nuestro señor dios: lleno de toda bondad y clemencia. Que los pequeñitos servicios: no solamente a el, sefchos mas a nuestros proximos por respecto suo: paga con tan crecido galardón que nos da por ellos la eterna bien auenturāça. Y esta es la causa por que sua persona se determina a ofrecer aun principe vna cosa aunque basconcedendo una mala respeto de honor al transcribirlo.

1-1 Epistola.
1-2 ¶ Epistola en la cual el auctor dirige la obra
1-3 presente al illustríssimo y muy excedente señor don Juan Claros de
1-4 Guzmán, Conde de Niebla e c. Primogénito del muy exelente señor
1-5 don Juan Alonso de Guzmán, duque de Medina Sidonia e c. mi señor.
1-6 La causa más principal, muy exelente
1-7 señor, por donde nos comovemos a amar y dessear servir a
1-8 un principe, o gran señor. Es por su misma benibolencia a sa-

User, Collection and Documents management tabs

Transkribus v1.6.0 (28_02_2019_14:59), Loaded doc: LeandroESP1, ID: 96141, Page 2, file:

Server Overview Layout Metadata Tools

Logout stefano.bazzaco.1@gmail.com

Document Manager User Manager

Versions Jobs

Recent Documents... User activity

Collections:
ESPLeandro (26094, Owner)

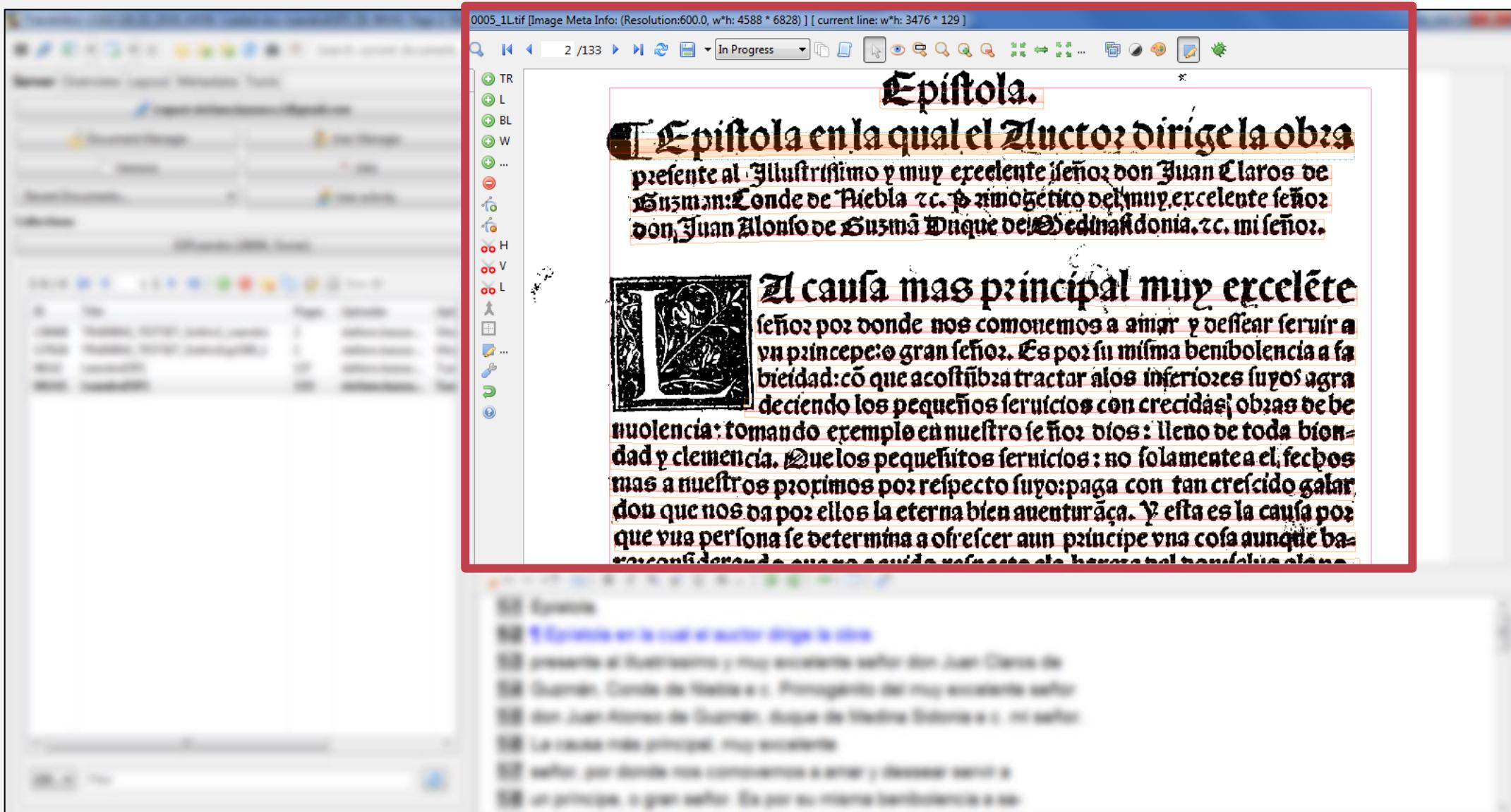
1-4 / 4 1 1 1 1 1 1 1 1 Doc-ID

ID	Title	Pages	Uploader	Uplo
138468	TRAINING_TESTSET_Gothic1_Leandro	2	stefano.bazzac...	Wed
137626	TRAINING_TESTSET_GothicEsp1500_1	1	stefano.bazzac...	Mon
96142	LeandroESP2	137	stefano.bazzac...	Tue
96141	LeandroESP1	133	stefano.bazza...	Tue

100 Filter

30

Canvas panel and image tabs



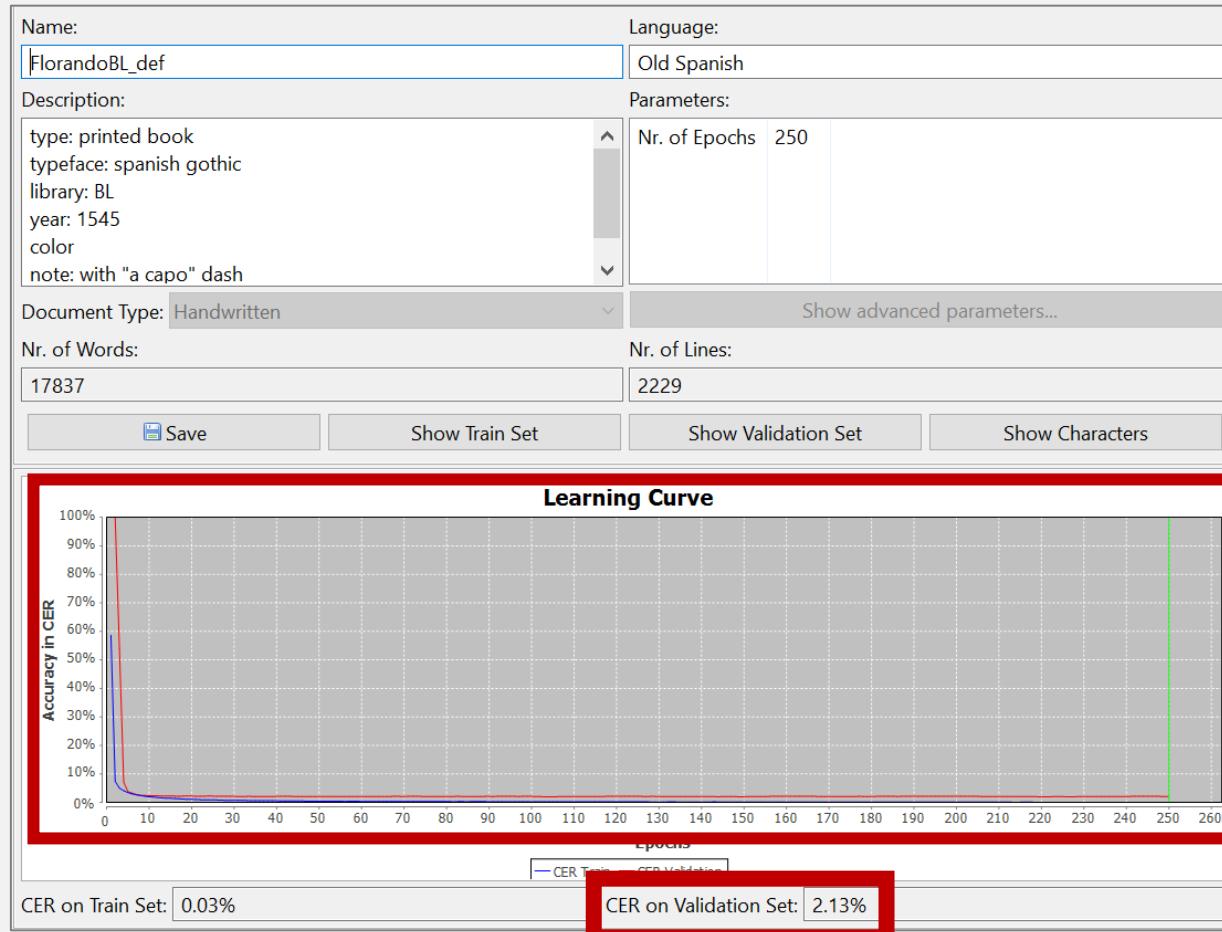
Transcription panel and tabs (virtual keyboard)

A screenshot of a Microsoft Word document window. The title 'Epistola en la cual el autor dirige la obra' is at the top. Below it is a large initial letter 'L'. The main text starts with 'La causa más principal, muy excelente señor por donde nos comovemos a amar y desear servir a un príncipe o gran señor. Es por su misma benibolencia a la humanidad que autor dirige tratar sobre instrucción...'. A red box highlights the first sentence. At the bottom, a numbered list from 1-1 to 1-8 is shown, corresponding to the text above.

Results

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, changes)



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

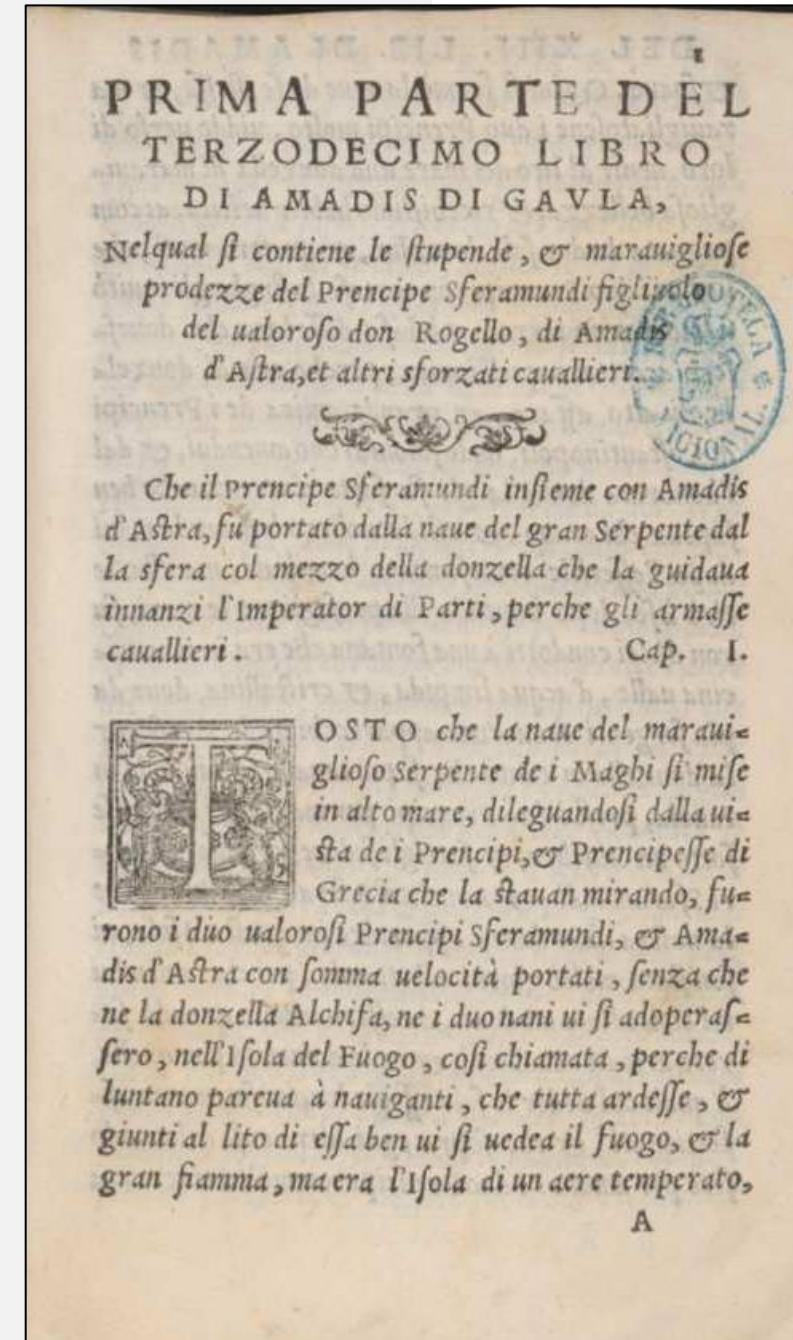
Each training suppose 2 datasets:
Train Set / Validation Set

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

Some results with printed texts

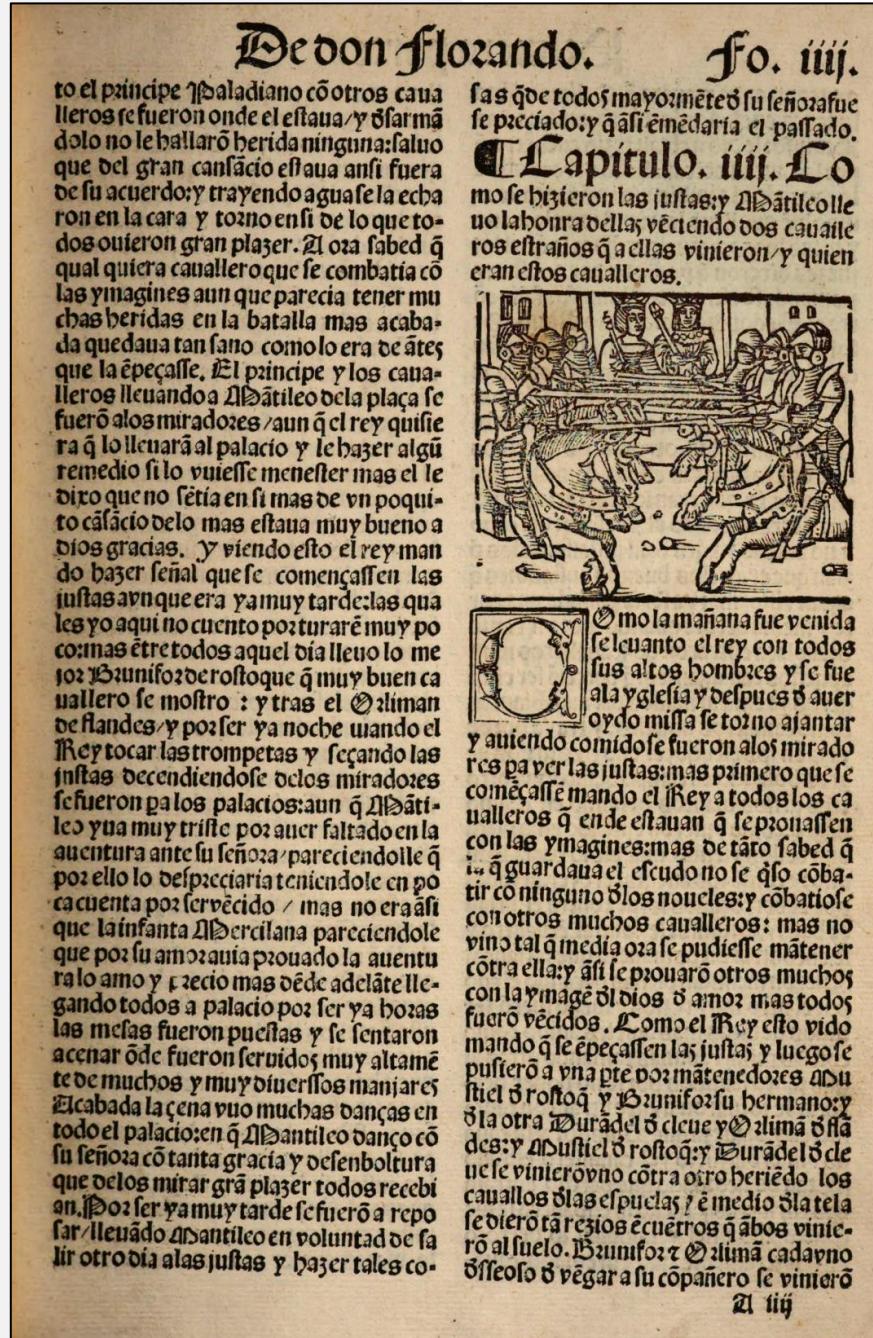
Italics (Aldine, XVIc)

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%



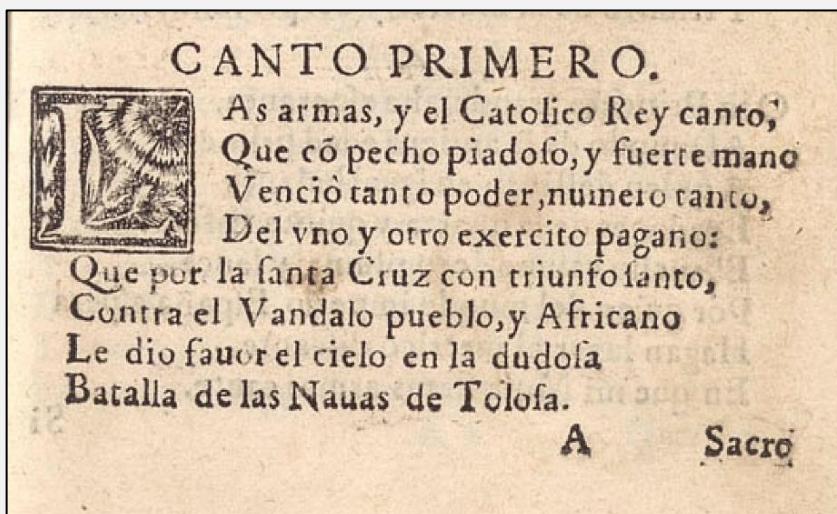
Spanish Gothic script (XV-XVIc)

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



Spanish Round script (XVI-XVIIc)

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



C A N T O

2

Sacrosanta Princessa, cuyo imperio
 Incapaz de mudanças de fortuna,
 No se estrecha en tan infimo emisferio,
 Mas vestida del Sol pisas la Luna:
 Tu nuestro Norte y luz, tu refrigerio,
 Tu nuestro escudo, y vnica coluna,
 Dame espiritu tal, que en gentil arte
 Pueda entonar mi canto el son de Marte.

3

Los versos que en sus músicos clarines,
 Citaras dulces, y organos sonoros
 Te cantan sin cessar los Serafines,
 Te cantan sin cessar los nueue coros:
 Sin fin celebren los ecelsos fines
 Del triunfo de la Cruz contra los Moros,
 Dela Cruz que abrio el cielo, y libró el mundo,
 Y triunfò de la muerte, y del profundo.

4

O tu Principe grande, alta esperança,
 Al mundo de la antigua edad primera,
 A quien destinan en igual balança
 Los dones de la quarta, y quinta Esfera:
 El ancho campo de tu pluma, y lança,
 Por quien del mundo imperio Espana espera
 Hagan lugar al metrico discante,
 En que mi Musa fieras armas cante.

Si

Costs

From november 2020, text recognition in Transkribus requires some costs:

1. to each new user they provide 500 free credits (more than 2500 pages for printed books)
2. when the 500 credits expires, various credit-packages are available
3. some special credit-packages are conceded for thesis

The screenshot shows the READ coop website interface. At the top, there is a navigation bar with links to Transkribus, ScanTent, read&search, The COOP, News, Sign in, and a shopping cart icon with '0' items. The main content area is divided into two sections: 'On-demand' on the left and 'Subscription' on the right.

On-demand section (Save 10%):

- With this package you can process 120 Pages (PyLaia Handwritten). Options: 120 Credits (16 €) or 500 Credits (59 €, limited sharing and buying options).
- With this package you can process 1200 Pages (PyLaia Handwritten). Options: 1200 Credits (13,5 € / year).

Subscription section (Save 25%):

- With this package you can process 1500 Pages (HTR+ Print). Options: 300 Credits (15,90 € / month, minimum subscription duration: 6 months) or 120 Credits (13,5 € / year).

Credit-packages for Coop members: <https://readcoop.eu/transkribus/credits/>

Extended models

- Extended models are based on a defined number of pages of **different works**
- Works usually are present the same script (but different versions – **chars inner variability**)
- 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 EXTENDED MODEL

- Multiple scripts / no work prevails on the others
- Each new document can be interpreted by its nearness to trained scripts

SPANISH GOTHIC 15TH-16TH CENTURY

A screenshot of the ReadCoop platform interface. At the top, there is a sample of Spanish Gothic text from the 15th-16th century. The text is in a medieval-style font and discusses an emperor in Constantinople named Julian. Below the sample, there is a small image of a manuscript page. To the left of the sample, the ID 'ID: 33106' is displayed. On the right side of the sample, there is a blue button labeled 'HTR+'. Below the sample, the text 'SPANISHGOTHIC_XV-XVI_EXTENDED (V1.0.0) - PRINT' is shown. Underneath this, the title 'Spanish Gothic 15th-16th Century' is displayed in a large, bold, dark blue font. Below the title, the author's name 'By: Stefano Bazzaco' is listed. At the bottom of the interface, there are four circular icons with labels: 'Spanish', '15th, 16th', 'Gothic Script', and '0.92 (CER)'. To the right of these icons, the text 'VIEW MODEL' is written in a dark blue font.

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

Authors:

Stefano Bazzaco (coord.)

Nuria Aranda García

Ángela Torralba Ruberte

Pedro Monteiro

Giada Blasut

Federica Zoppi

SPANISH REDONDA 16TH-17TH CENTURY

En defension de vna auéntura estraña:
Del qual por oy tratar no de termino,
Pues ya los quattro siguen su camino.

En corte en fin llegados, no quisieron
reves entrar con la donzella:

ID: 33399

Quanto el menester graue lo requiere
(O sea, ò no, tu subdito y vassallo),
Que en campo tres dias pueda sustentallo.

Con esto tâbien pide, que aunque dado
Fuese el dia primero al que pareza.

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

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

Authors:

Stefano Bazzaco (coord.)

Gaetano Lalomia

Daniela Santonocito

Manuel Garrobo Peral

Mónica Martín Molares

Carlota Cristina Fernández Travieso

Transkribus: advanced functionalities

1. **Page to Page Layout Analysis (P2PaLA)**: automated segmentation tool based on *pre-trained layouts*
2. **Text2image**: automated import of already transcribed materials (not stable)
3. **Structural / Textual Metadata**: possibility to add structural and semantic metadata (exported in the XML file)
4. **Keyword Spotting (KWS)**: words can be searched in the recognized text and are detected even if the transcription is not reliable

Advanced tools (1) : P2PaLA

P2PaLA tool (Page to Page Layout Analysis) permits to create some **pre-trained layout models**

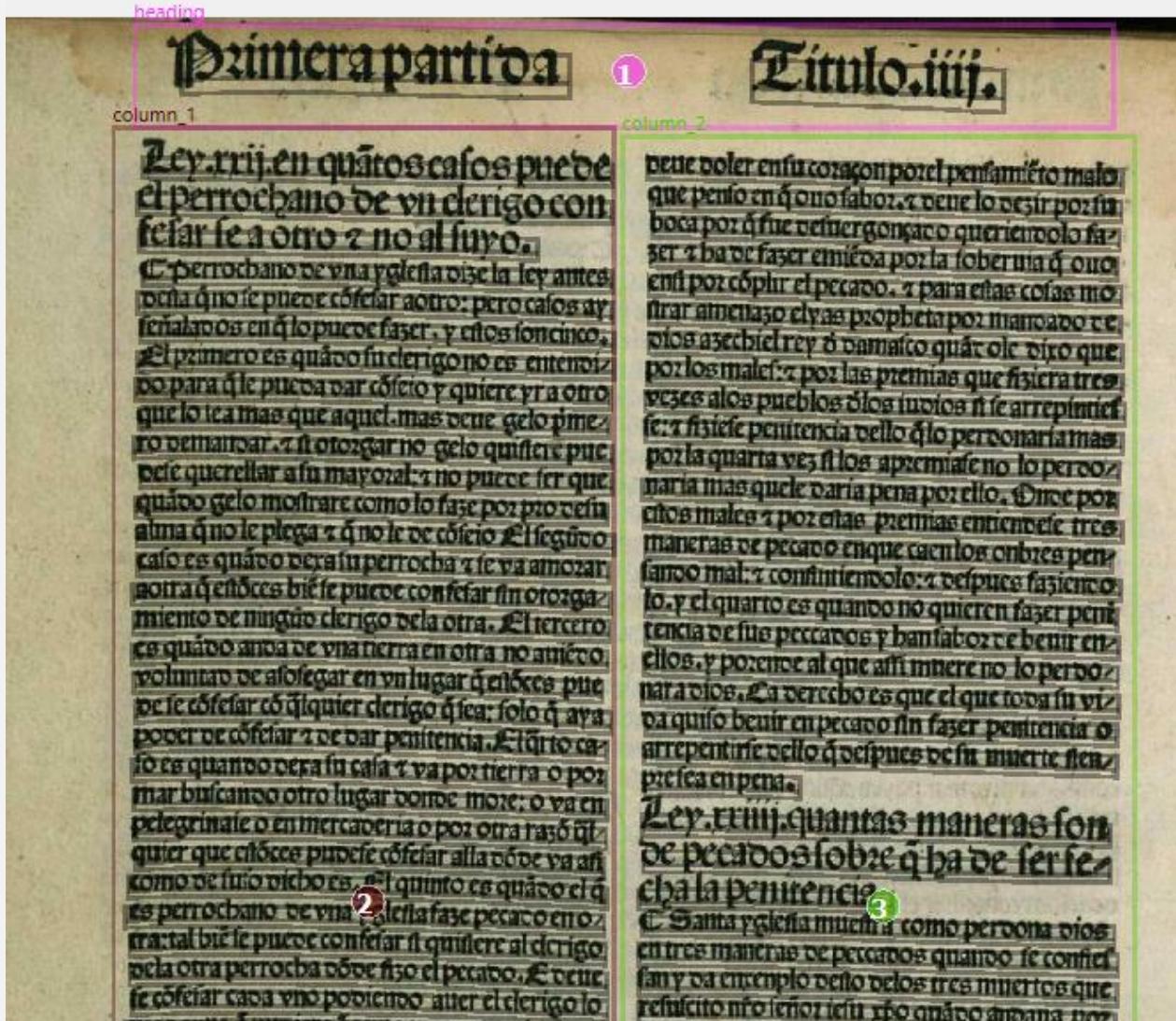
By default, inside Transkribus the Layout Analysis tool detects text lines and regions *from the top to the bottom*

In the cases of complex layouts (2 or more columns, tables, marginalias, ...) we can resort to P2PaLA tool:

- images upload
- manual segmentation and reading order check
- structural metadata for regions
- segmented pages training (200 for columns, more than 1000 for tables)
- creation of a P2PaLA model

P2PaLA limitations:

- not stable for complex sources (p.e. *littera nobilior*, tables)
- it requires some manual work



For 2 columns layout:

- 1) Heading
- 2) Column 1
- 3) Column 2

Reading order: 1, 2, 3

Model:

SpanishGothicBooks_2columns+heading

Train set: 200 pages

Advanced tools (2): KWS

KWS (KeyWord Spotting) function permits to find keywords inside the recognized text by a fuzzy search

KWS is based on the graphic representation of words, not on the recognized transcript

Results are stored aggregated to a confidence index (from 1 to 0) which express the reliability of that result in relation to the textual string we are searching:

- Confidence index between 0.8-1.0: quite sure is the same word we were searching
- Confidence index between 0.2-0.7: maybe a word close to the one we were searching for (sing/plu; gender)

Keyword Spotting Results			
"cavalliere" (771 hits)		"gigante" (87 hits)	
Confidence ^	Page Nr.	Line transcription	Prev...
0.8449	183	Ma il cavalliere, che amava di amor grande	cavallie
0.8399	112	parlar del cavalliere e dicevano che veramen-	cavallie
0.8365	198	cavalliere. Questa mischia fu grande e pe-	cavallie
0.8326	111	con giocondo riso, gentil cavalliere non vi da-	cavallie
0.8002	79	suo amato cavalliere, che era di seta pavo-	cavallier
0.7886	245	se, e le disse l'Infanta. Il vostro cavalliere è	cavallies
0.4113	109	era presso la figliuola per ricever il cavallier	cavallie
0.3644	135	tri di lancia mostrato esser uguale al cavallier	cavallie
0.3500	139	in tutte quelle qualità, che in buoni cavallieri	cavallies
0.3380	88	gentil come questa. La moglie del cavallier	cavallie
0.2390	153	na con questo cavallier ragionava tanto più se	cavallie
0.2388	99	le del cavallier col quale eran quei principi al	cavallie
0.2364	344	può far il più lieto cavallier che viva. Di-	cavallies
0.2355	106	to il poter loro. Il cavallier che avea mem-	cavalliu
0.2304	92	ta il cavallier che tanto vi ama, che solo per	cavalliu
0.2133	63	rebbe gran pregiudizio apportato. Il cavallier	cavallie
0.2116	165	in arme eccellente, che amor fa il cavallier ar	cavallie
0.2098	132	di questa giostra. Questo cavallier ha tanto	cavallie
0.2094	99	e che il cavallier dall'arme turchine ben a-	cavallie
0.2089	107	na tutta ridente disse, poi che il cavallier dal-	cavallie
0.2078	106	il cavallier non era di quel tronco ferito, che	cavalli
0.2061	296	soccorrevi, e questo è il cavallier nostro com-	cavallie
0.2039	8	per marito, ma il cavallier con franco animo	cavallie

Preview



Transkribus: sustainability > 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 HTR recognition models (individual / extended)
 - new P2PaLA models of *pre-trained layouts*
- **Release of new infrastructures and tools:**
 - Scan tent + DocScan app (portable digitalization tools)
 - Transkribus Lite* (web browser version – new!)
 - Transkribus Learn* (to train young transcribers)
 - Transkribus Read&Search* (web pub. platform)