TexT – Text Extractor Tool for Handwritten Document Transcription and Annotations

Anders Hast^, Per Cullhed* & Ekta Vats^

^Department of Information Technology,

Division of Visual Information and Interaction

*Uppsala University Library

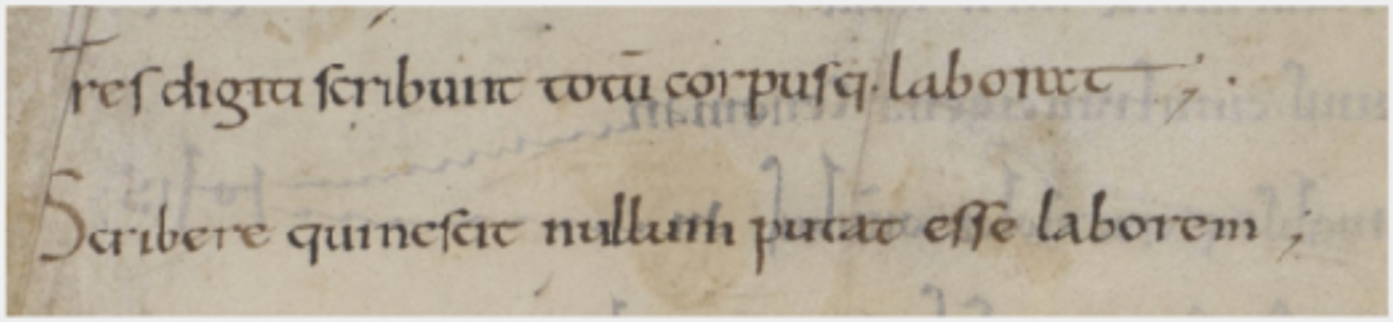
14:th Italian Research Conference on Digital Libraries, Udine, Italy

Outline

- Transcription is a **tedious** task that can be made **simple** by handwritten text recognition techniques, such as **word spotting**
- It will be shown how transcription can be done in a **new way**, but also how researchers can get **easy access**, using similar techniques, to manuscripts that are otherwise not searchable
- Semi-Automatic transcription
  - In arbitrary word order instead of line by line, word by word
  - (Problem solving through a visualisation of the problem)

Copying books is an old trade

UPPSALA
UNIVERSITET

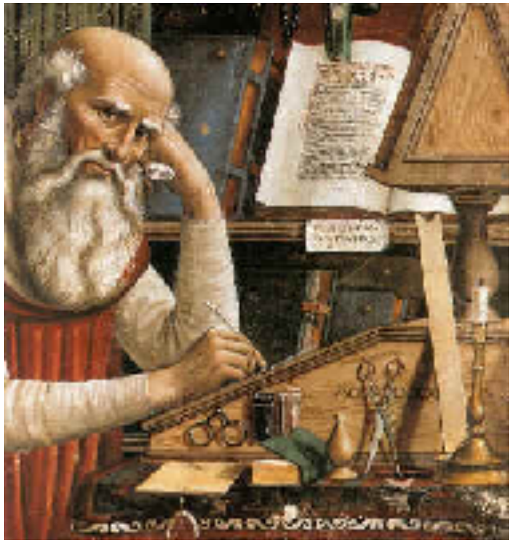


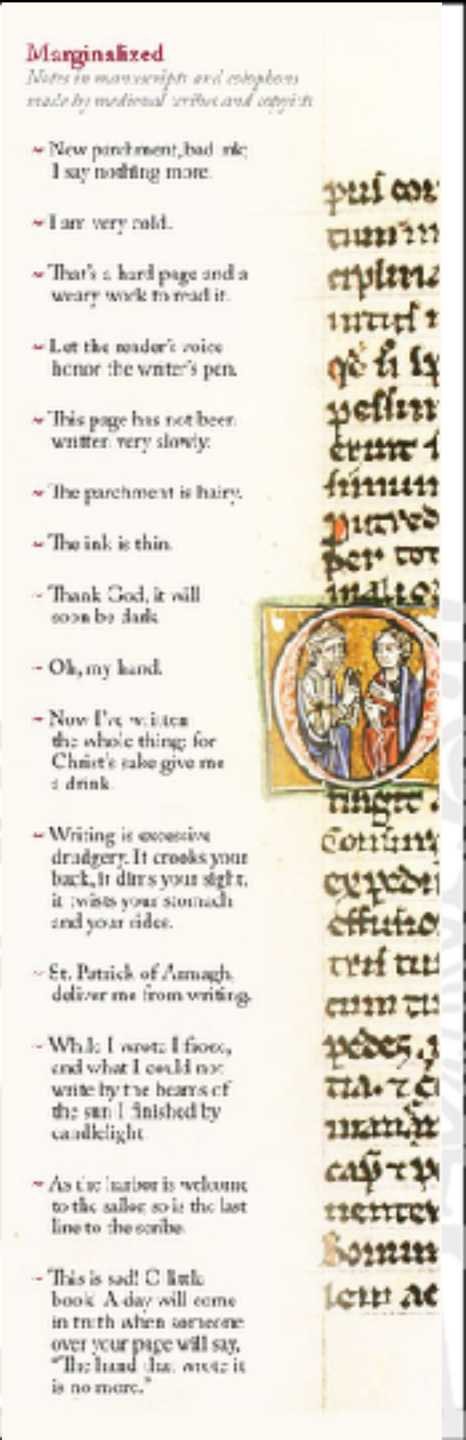
Treſ digitı ſcribunt totū corpuſq· laborat ;·

Scribere quı neſcit nullum putat eſſe laborem ;

© LONDON, BRITISH LIBRARY, ROYAL 6 A.VI, FOL. 109R

> *Tres digiti scribunt totum corpusque laborat. Scribere qui nescit nullum putat esse laborem.*
>
> *[Three fingers write and the whole body labours. He who does not know how to write thinks it is no work.]*





Marginalized

Notes in manuscripts and colophons made by medieval scribes and copyists

- New parchment, bad ink; I say nothing more.
- I am very cold.
- That's a hard page and a weary work to read it.
- Let the reader's voice honor the writer's pen.
- This page has not been written very slowly.
- The parchment is hairy.
- The ink is thin.
- Thank God, it will soon be dark.
- Oh, my hand.
- Now I've written the whole thing: for Christ's sake give me a drink.
- Writing is excessive drudgery. It crooks your back, it dims your sight, it twists your stomach and your sides.
- St. Patrick of Armagh, deliver me from writing.
- While I wrote I froze, and what I could not write by the beams of the sun I finished by candlelight.
- As the harbor is welcome to the sailor, so is the last line to the scribe.
- This is sad! O little book! A day will come in truth when someone over your page will say, "The hand that wrote it is no more."

UPPSALA UNIVERSITET

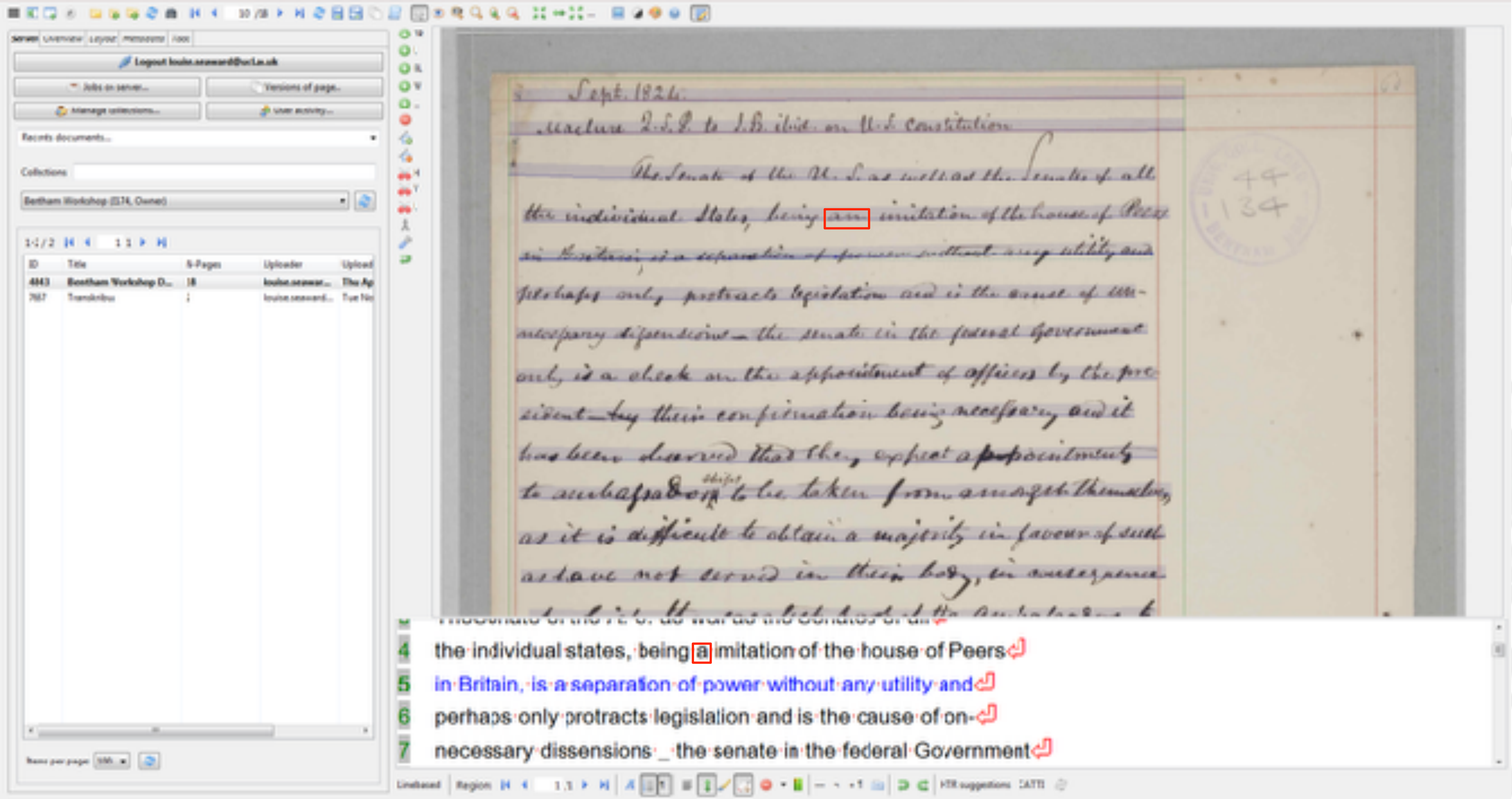
Medieval Marginalia

- The monks tried to brighten up their day by scribbling in the margins
- If copying the text was boring, transcribing it will be likewise boring…

Transkribus

https://transkribus.eu/Transkribus/

Even modern transcribers makes mistakes…



ID	Title	#-Pages	Uploader	Upload
4443	Bentham Workshop D...	18	louise.seawar...	Thu Ap
7867	Transkribus	1	louise.seaward...	Tue No

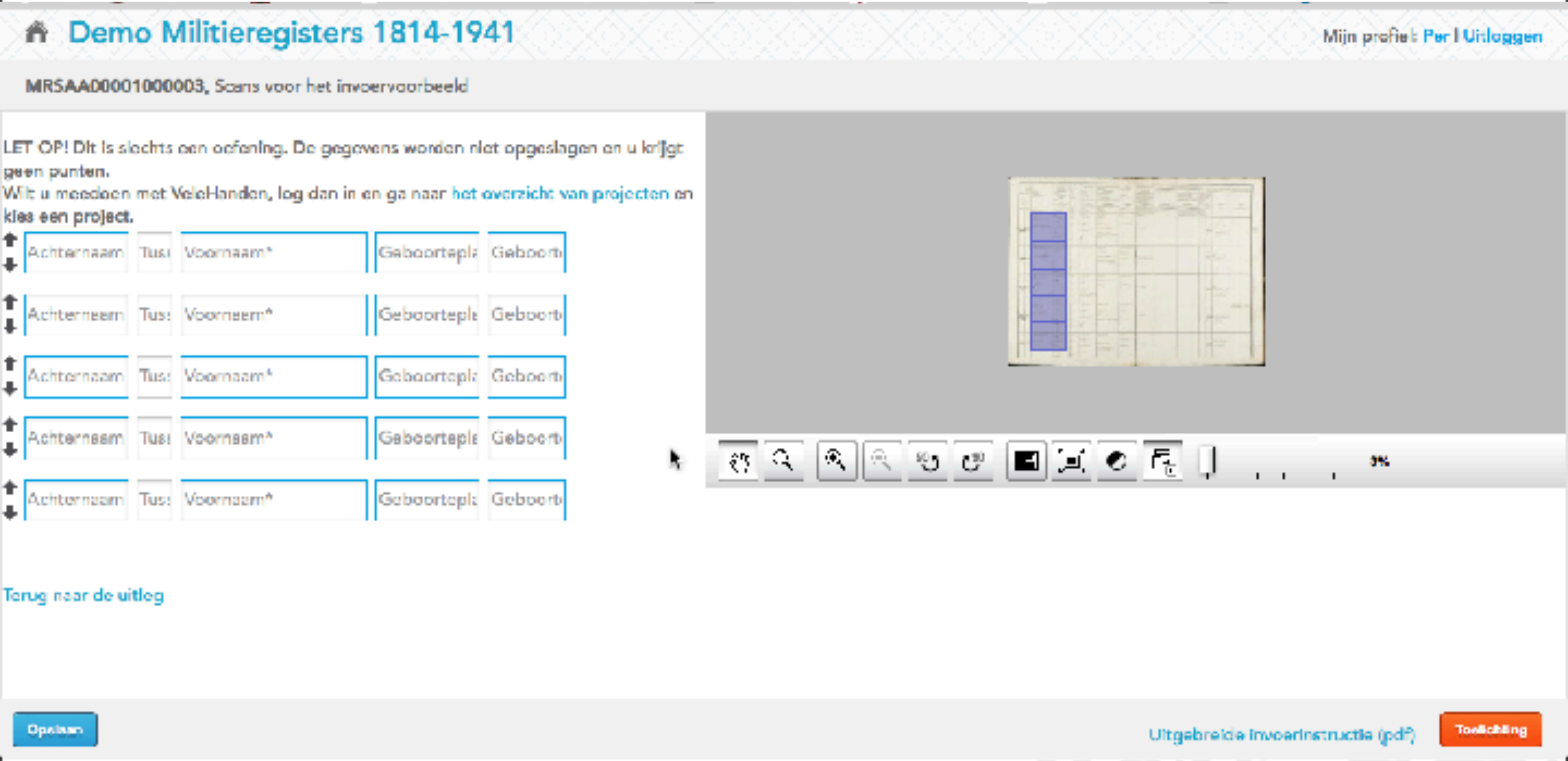
4 the individual states, being a imitation of the house of Peers
5 in Britain, is a separation of power without any utility and
6 perhaps only protracts legislation and is the cause of on-
7 necessary dissensions _ the senate in the federal Government

Transkribus

- Now, this sounds interesting, but it starts to get really exciting if you consider that
  - **once you have properly transcribed e.g. 100 images** you may inform us and we will train an HTR engine from the Computational Intelligence Technology Lab (CITlab) of the University of Rostock on your documents and
  - you will be able to transcribe further pages of your documents with the **support of automatically produced handwritten text**.

Veele handen

Is a Dutch initiative where volunteers can sign in and start transcribing in a simple manner



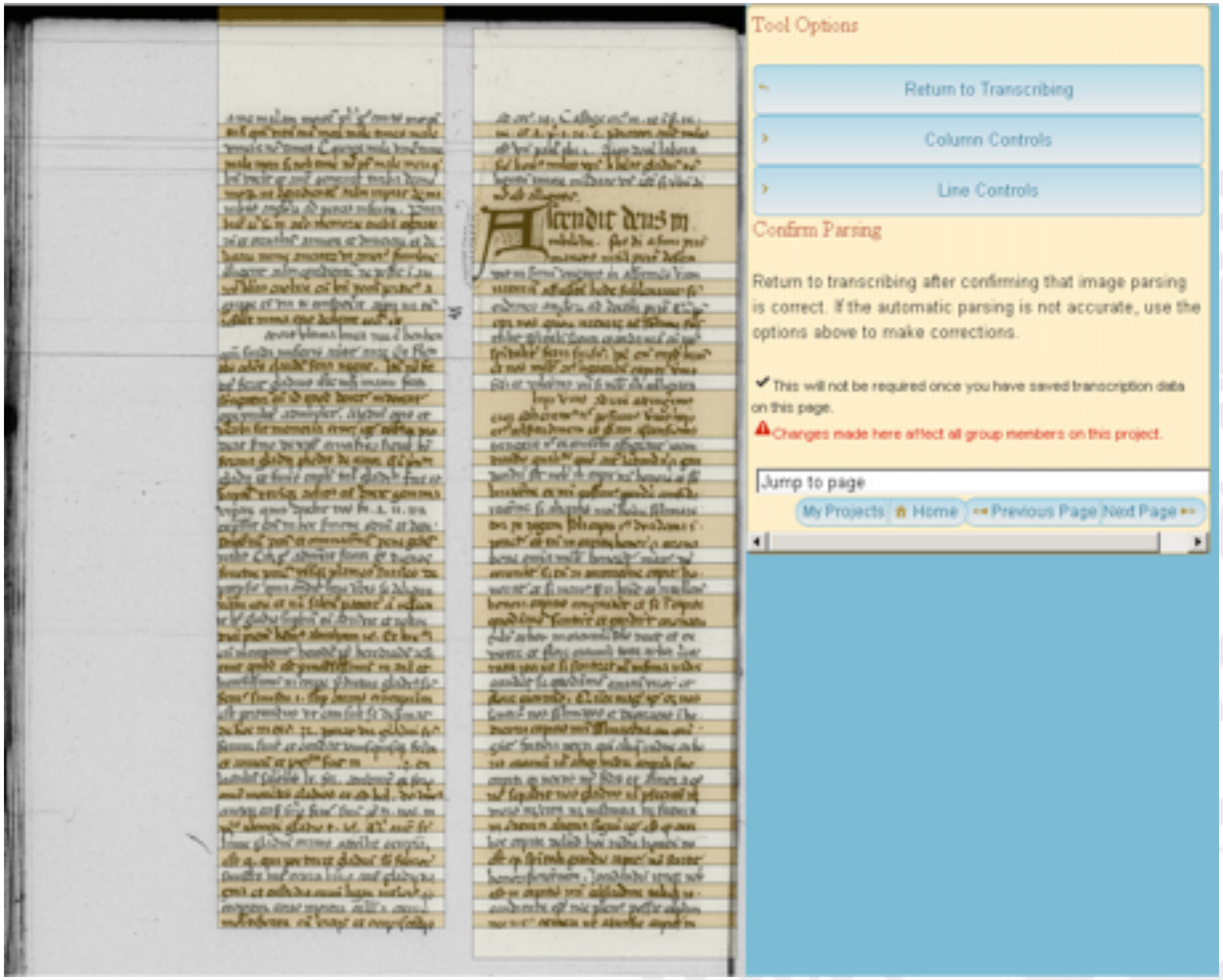
Demo Militieregisters 1814-1941

MRSAA00001000003, Scans voor het invoervoorbeeld

LET OP! Dit is slechts een oefening. De gegevens worden niet opgeslagen en u krijgt geen punten.
Wilt u meedoen met VeleHanden, log dan in en ga naar het overzicht van projecten en kies een project.

T-pen

From the university of St. Louis in the United States is much used for the transcriptions of medieval manuscripts.



Confirm Parsing

Return to transcribing after confirming that image parsing is correct. If the automatic parsing is not accurate, use the options above to make corrections.

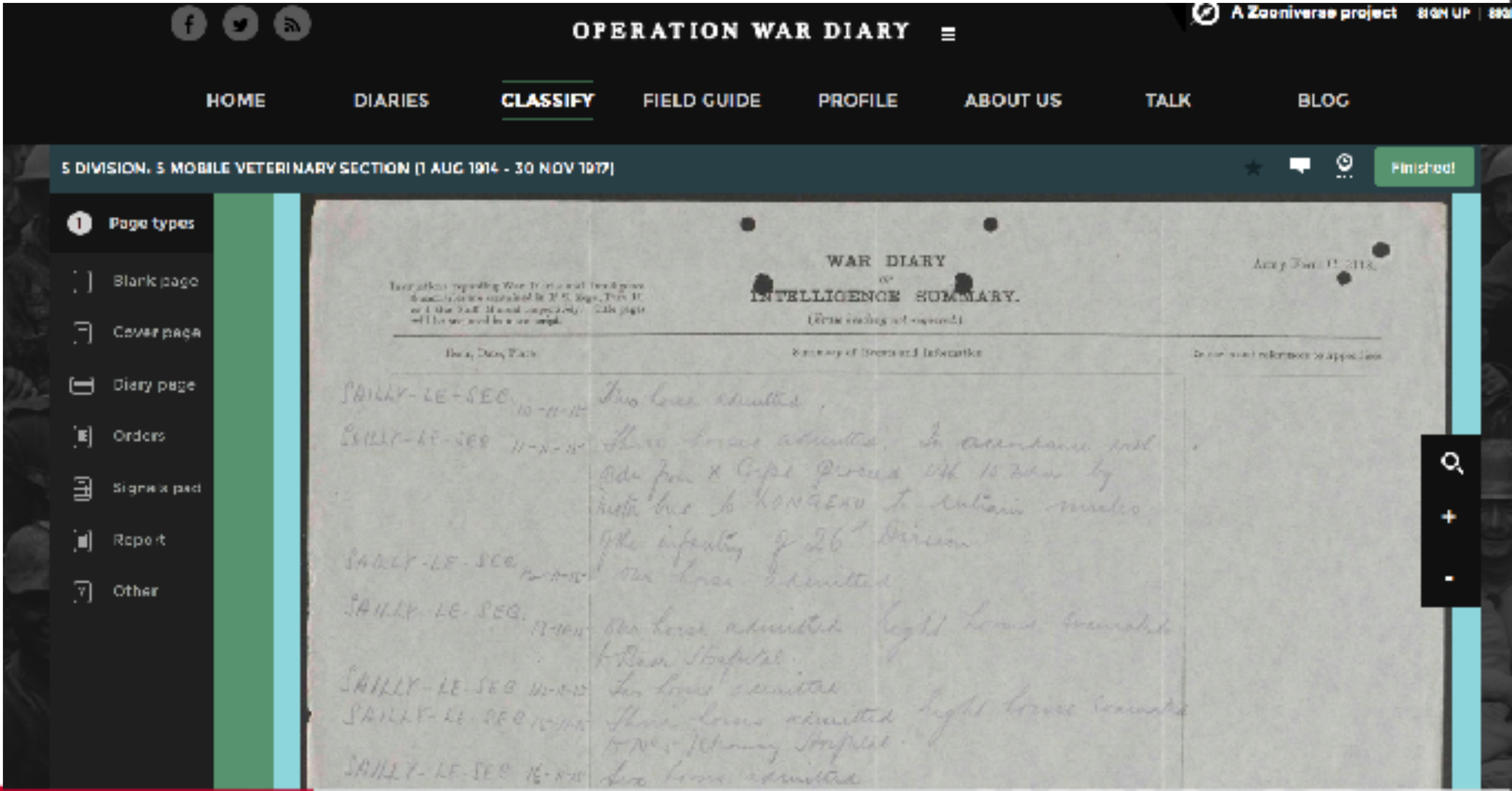
✓ This will not be required once you have saved transcription data on this page.

⚠ Changes made here affect all group members on this project.

The Zooniverse transcription project

The Oxford based Zooniverse is a general crowdsourcing platform used for many purposes, not only transcription. It has one million users-mostly from English speaking countries.

Operation War Diary



5 DIVISION. 5 MOBILE VETERINARY SECTION (1 AUG 1914 - 30 NOV 1917)

Instructions regarding War Diaries and Intelligence Summaries are contained in F. S. Regs., Part II. and the Staff Manual respectively. Title pages will be prepared in manuscript.

WAR DIARY
or
INTELLIGENCE SUMMARY.
(Erase heading not required.)

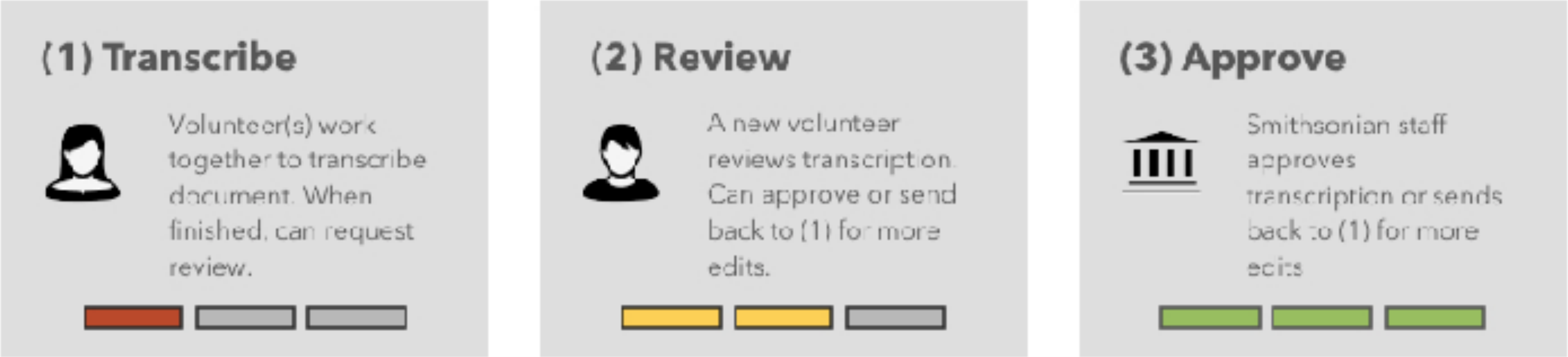
Army Form C. 2118.

Hour, Date, Place	Summary of Events and Information	Remarks and references to Appendices
SAILLY-LE-SEC 10-11-15	Two horses admitted.	
SAILLY-LE-SEC 11-11-15	Three horses admitted. In accordance with orders from X Corps proceeded with 15 horses by motor bus to LONGEAU to [illegible] mules of the infantry of 26th Division	
SAILLY-LE-SEC 12-11-15	One horse admitted	
SAILLY-LE-SEC 13-11-15	One horse admitted. Eight horses evacuated to Base Hospital.	
SAILLY-LE-SEC 14-11-15	Ten horses admitted	
SAILLY-LE-SEC 15-11-15	Three horses admitted. Eight horses evacuated to No 1 Veterinary Hospital.	
SAILLY-LE-SEC 16-11-15	Six horses admitted	

UPPSALA UNIVERSITET

Smithsonian

SMITHSONIAN DIGITAL VOLUNTEERS: TRANSCRIPTION CENTER



(1) Transcribe

Volunteer(s) work together to transcribe document. When finished, can request review.

(2) Review

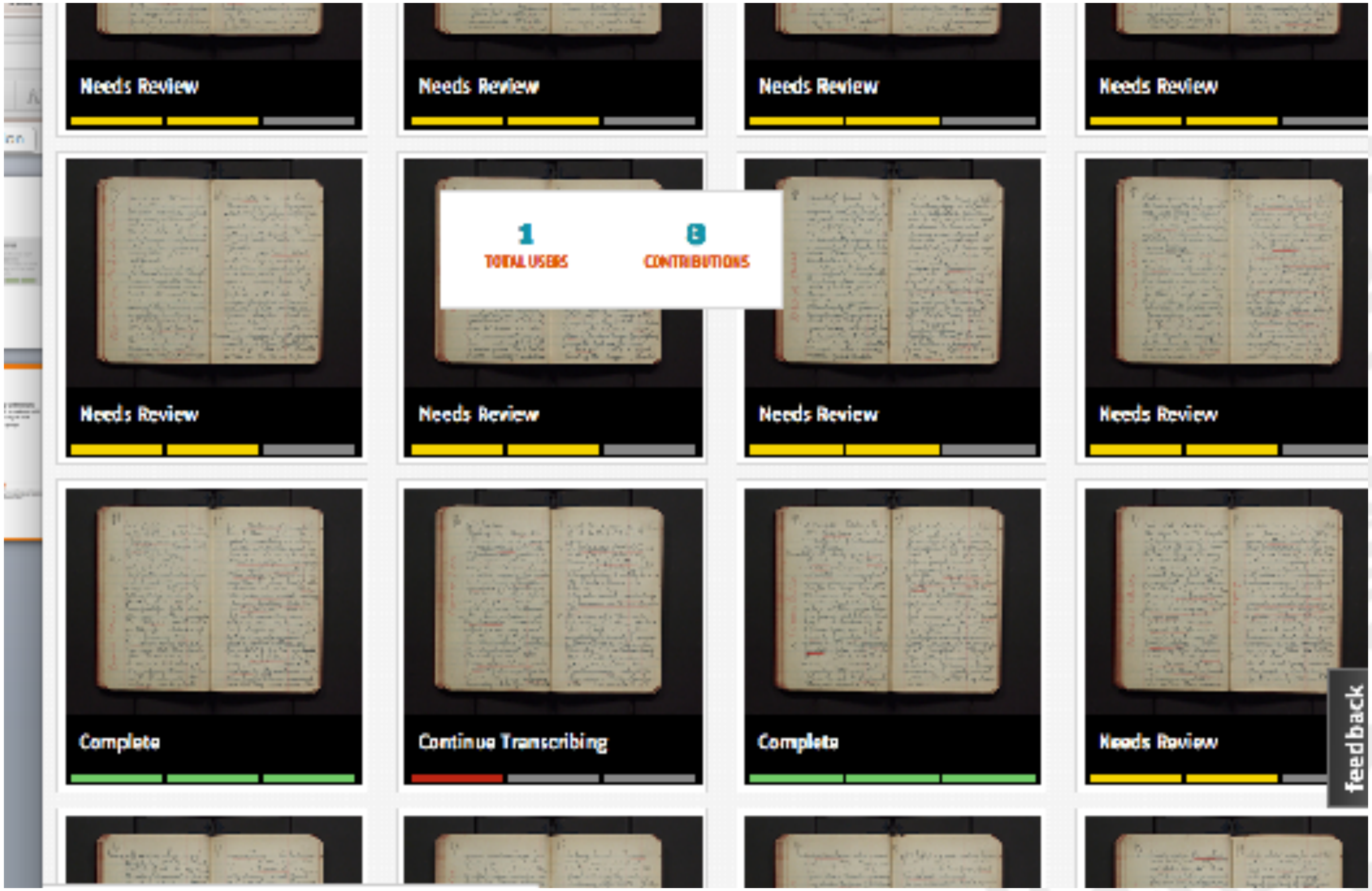
A new volunteer reviews transcription. Can approve or send back to (1) for more edits.

(3) Approve

Smithsonian staff approves transcription or sends back to (1) for more edits.

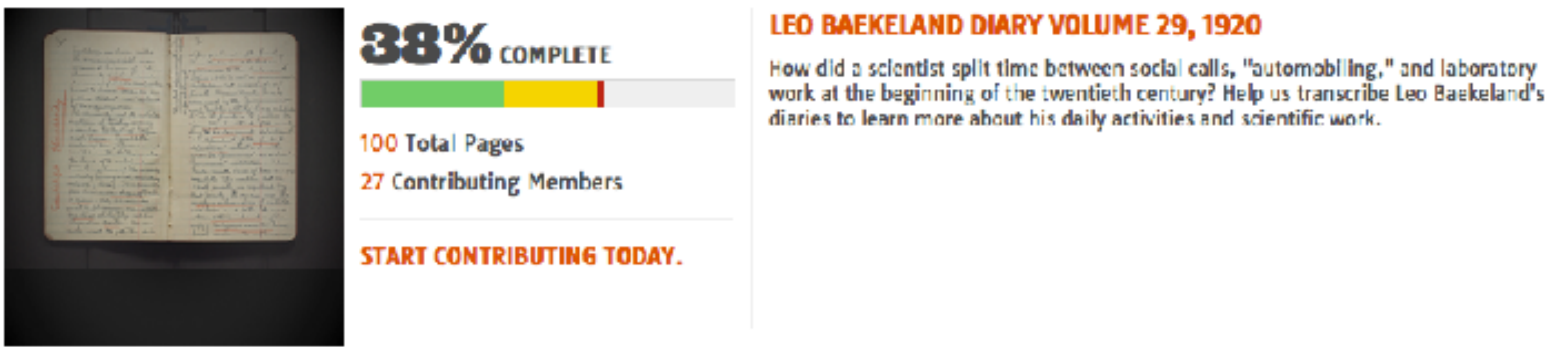
Example

Every transcriber knows exactly where to start contributing as this is obvious from the GUI



The Purpose of Transcription

- Our goal with this project is to make our collections:
  - more accessible and useful to curators, researchers, and anyone with a curious spirit
- Because computers have a hard time understanding handwriting, many of our collections still hold many secrets and hidden knowledge inside their pages
  - With your help, we can bring that knowledge to life.
- Searchability & Readability - Usefulness over Perfection



38% COMPLETE

100 Total Pages

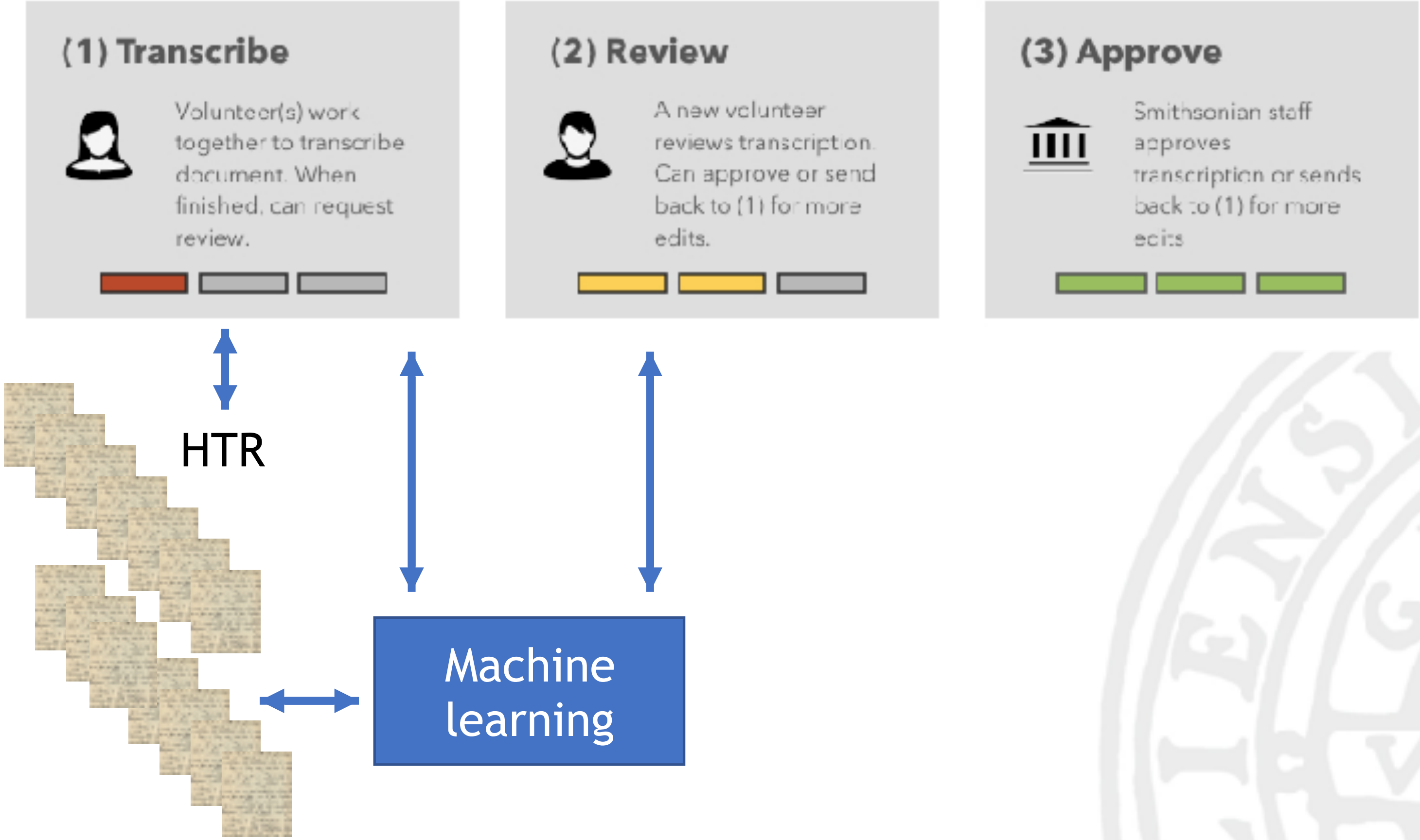
27 Contributing Members

START CONTRIBUTING TODAY.

LEO BAEKELAND DIARY VOLUME 29, 1920

How did a scientist split time between social calls, "automobiling," and laboratory work at the beginning of the twentieth century? Help us transcribe Leo Baekeland's diaries to learn more about his daily activities and scientific work.

TexT-project - Uppsala



(1) Transcribe

Volunteer(s) work together to transcribe document. When finished, can request review.

(2) Review

A new volunteer reviews transcription. Can approve or send back to (1) for more edits.

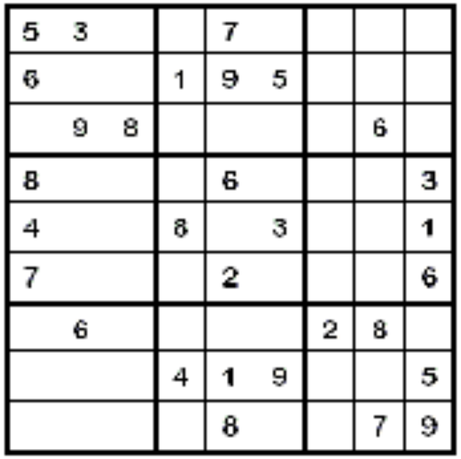
(3) Approve

Smithsonian staff approves transcription or sends back to (1) for more edits.

HTR

Machine learning

How can we make it fun?



5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

- Why do people love Sudoku?!?
- Is it really a game?
  - Or rather about problem solving?
  - And the satisfaction of reaching a goal?
- You learn NOTHING?!
- From crosswords you might at least learn something…

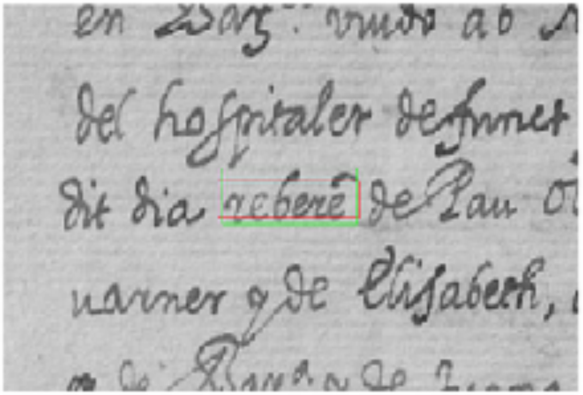
Can Transcription be “fun”?

- Can we make it a game?
  - How and why? Should we even??
- Or is it rather about problem solving?
- Can you learn something while doing it?
- This is a challenging task!

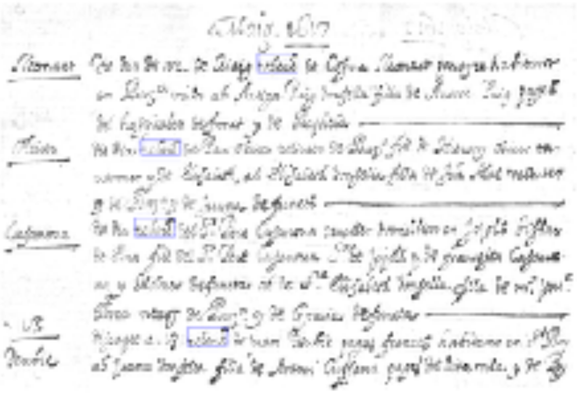
Gamification - is that the Solution?

- **Gamification** is *the application of game-design elements and game principles in non-game contexts*
- Probably no first person shooter game…
  - not even medieval fighting games… or?!
- What about problem solving?
  - If Sudoku can be fun..

Word Spotting

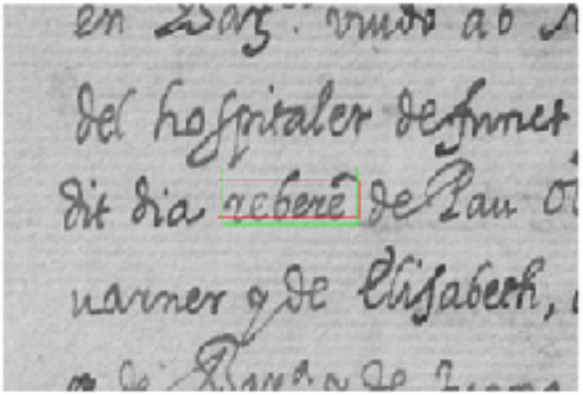


en [illegible] vindo ab [illegible]
del hospitaler defunct
dit dia acebere de Pau [illegible]
varner y de Elisabeth,



- The user marks up a word
- …and the HTR engine finds it in the text
- Can be used to make large text document collections searchable
- Ex: Monk
  - L. Schomaker, Rijksuniversiteit Groningen
  - Search & annotation tools for handwritten manuscripts
- Monk learns by user annotation

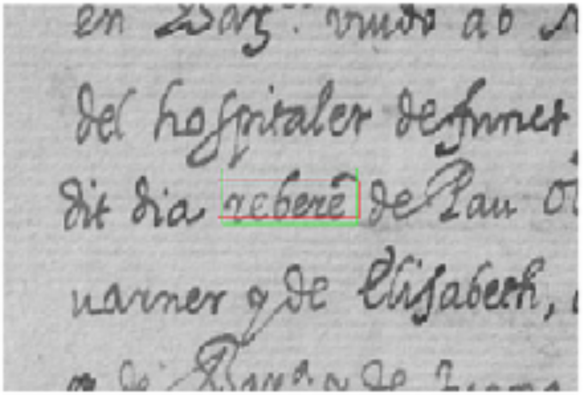
Word Spotting

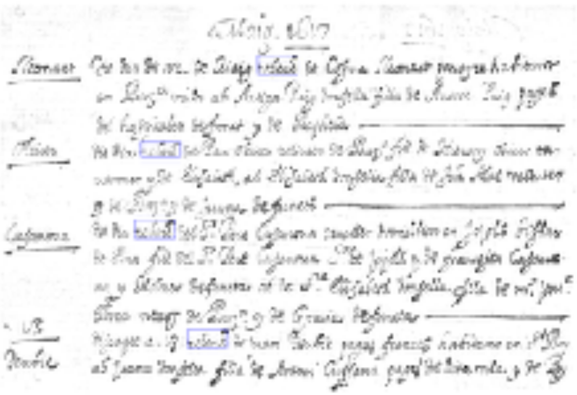


- We have developed an HTR engine for Word Spotting (WS)
- Segmentation Free
  - Based on stitching principles from Computer Vision
  - i.e pattern matching
- Learning free

A Segmentation-free Handwritten Word Spotting Approach by Relaxed Feature Matching
Anders Hast and Alicia Fornés
DAS2016, International Workshop on Document Analysis Systems, 2016.

Word Spotting





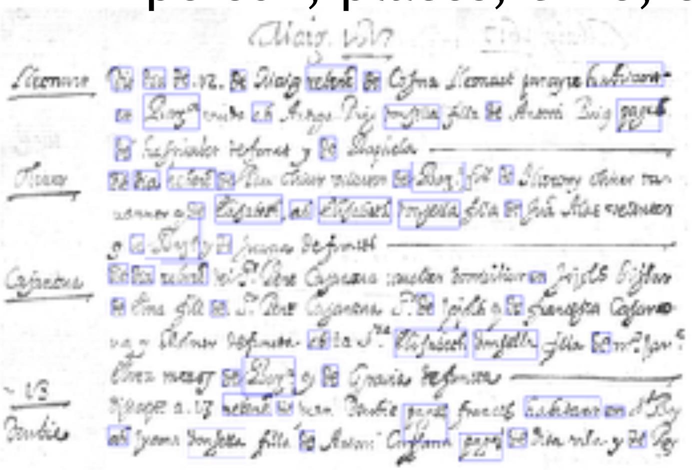
- We plan to use the WS approach for transcription as well!

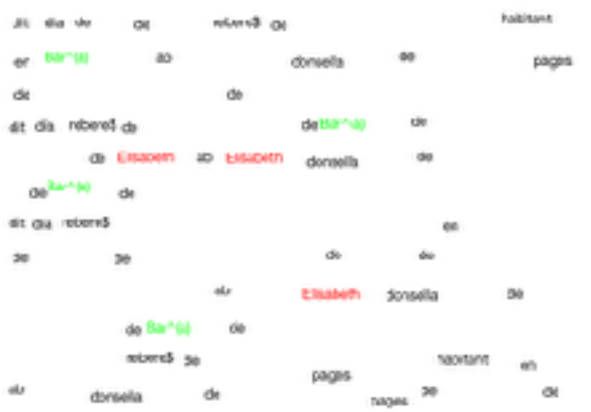
- The user marks the red rectangle
- The system finds the "perfect" bounding box
  - Principle: "all the word and nothing but the word"!

On-the-fly Historical Handwritten Text Annotation
Ekta Vats and Anders Hast
1st International Workshop on Human-Document Interaction (HDI 2017), in conjunction with ICDAR 2017

Semi Automatic Transcription

- User should mark a word and transcribe it ONCE
- Then the HTR-Engine should find all occurrences
- Annotation possibilities!
  - person, places, time, etc…

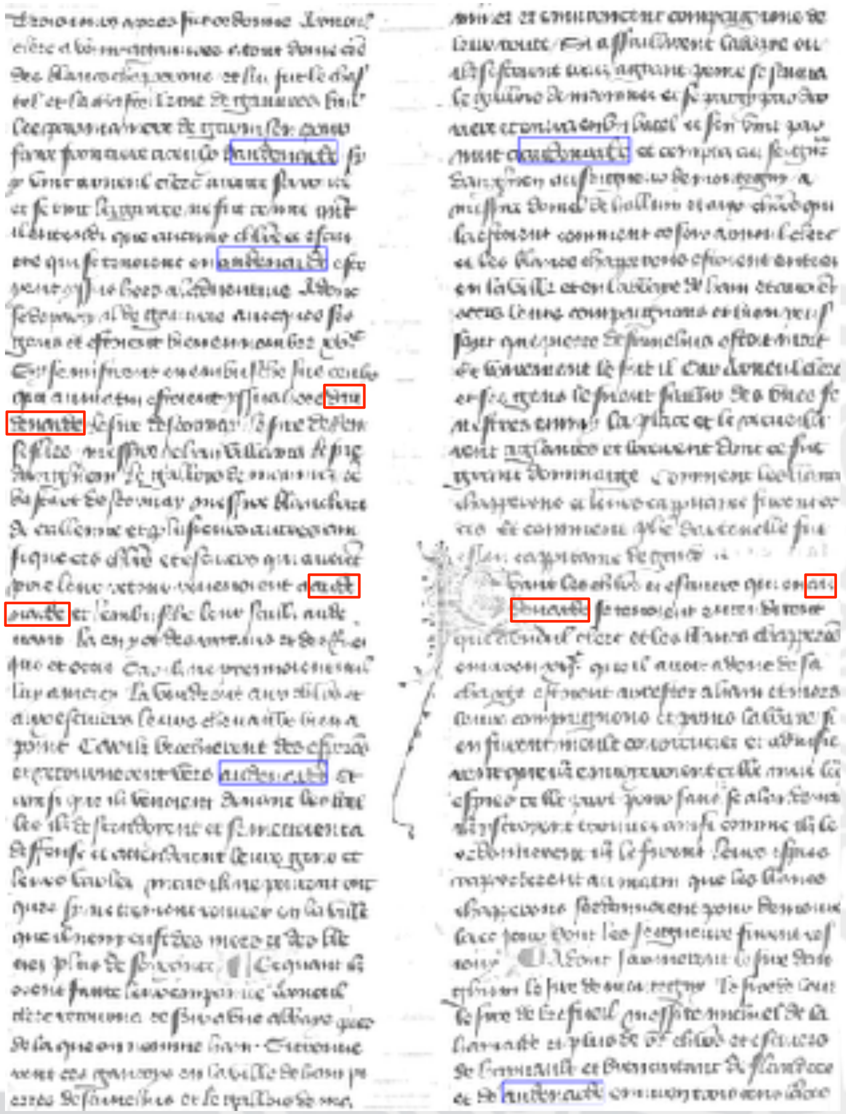




TexT - Text Extractor Tool for Handwritten Document Transcription and Annotation
Anders Hast, Per Cullhed and Ekta Vats
14:th Italian Research Conference on Digital Libraries, IRCDL 2017

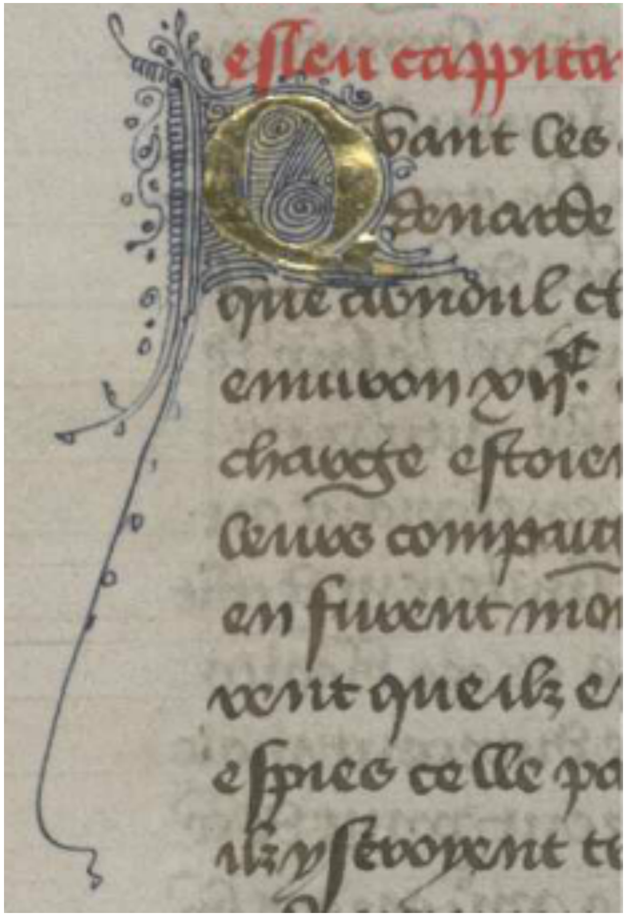
Problems

- Only whole words are found
- Note how “d’Audenarde” and “a Audenarde” are contracted to one word
  - This is usually a small problem
- A bigger problem are the split words
- Solvable by letter spotting?

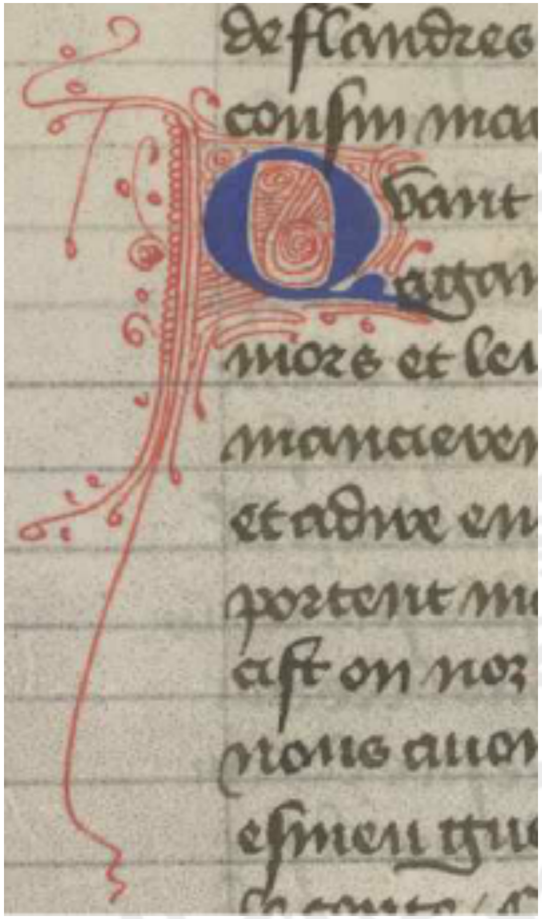


Problems with Initials

Close to searched word (Audenarde)

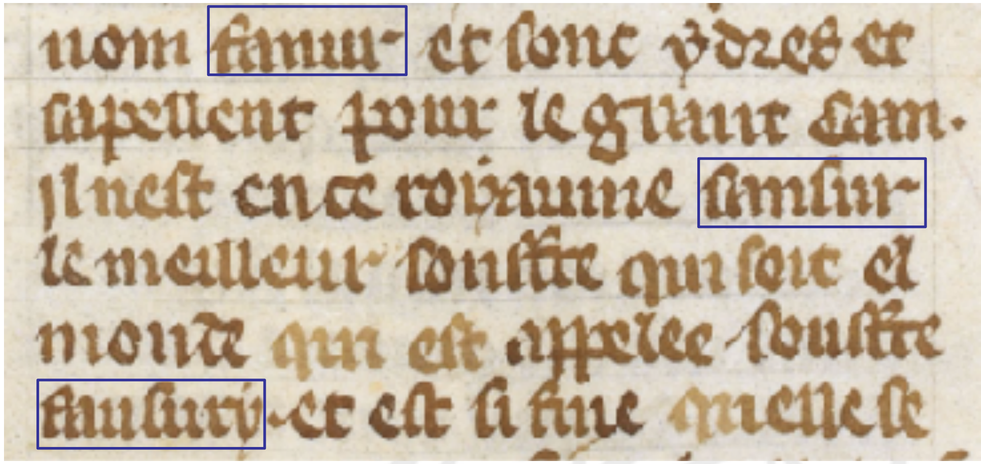


Or if it is the word we search (Quant)



An Example of problems with Orthography

- A section of Marco Polo’s account dealing with Fansur
- Here spelled three different ways:
  - <fanur>
  - <sansur>
  - <fausur(y)>



nom fanur· et ſont ꝑdres et
ſapellent pour le grant kaan·
il neſt en ce roiaume ſanſur·
le meilleur ſoulfre qui ſoit el
monde qui eſt appelee ſoulfre
fauſury· et eſt ſi fine que elle eſt

British Library Royal MS 19 D I, f.123r

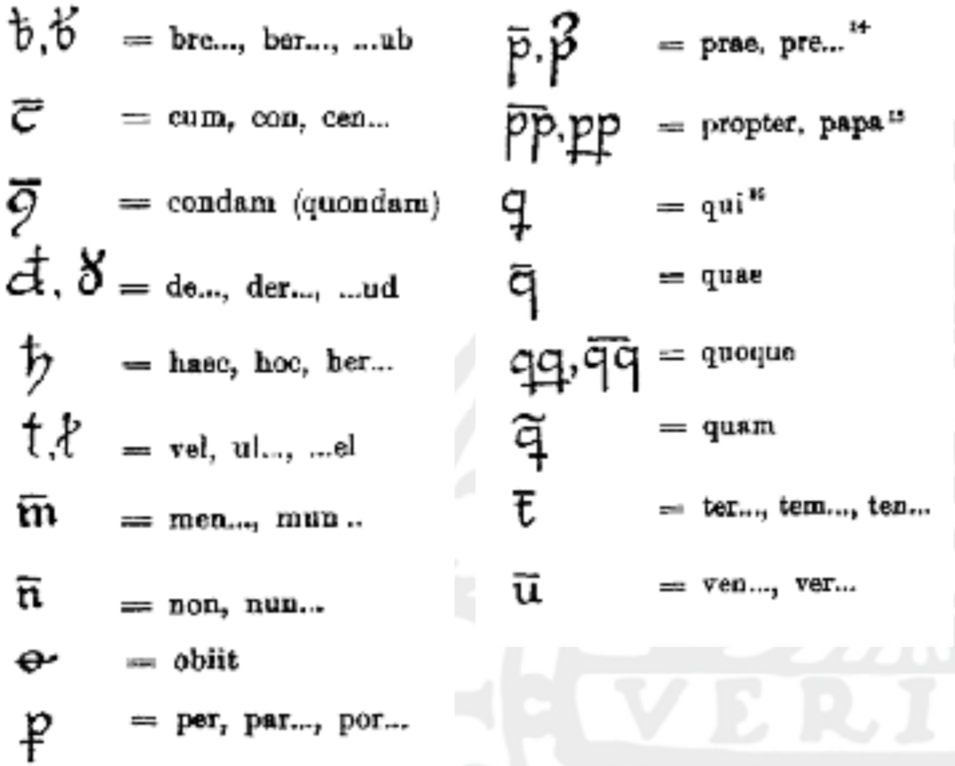
https://medium.com/@siwaratrikalpa/barus-sumatra-in-a-medieval-egyptian-text-edcfb7be9517

Abbreviations is not a new thing

The most common texting abbreviations, Italian-style

- m = **mi** (*I*, *me*)
- t = **ti** (*you*)
- xke = **perché** (why, because)
- cmq = **comunque** (*anyway*)
- bc = **baci** (*kisses*)
- midi = mi dispiace (*I‘m sorry*)
- pfv = per favore (*please*)
- d = **da** (from, since, of)
- grz = **grazie** (*thanks*)
- tn = **tanto** (a lot, much, long time)
- k = **chi** (who, what)
- c6 = **Ci sei?** (Are you there?)
- qls = **qualcosa** (*something*)
- + = **più** (*more*)
- risp = **rispondi** (*answer*)
- nn = **non** (*no, not*)
- prox = **prossima** (*next*)
- gg = **giorno** (*day*)
- tvb = ti voglio bene (*I love you*)
- ta = **ti amo** (I love you)

Common abbreviations, Monk-style



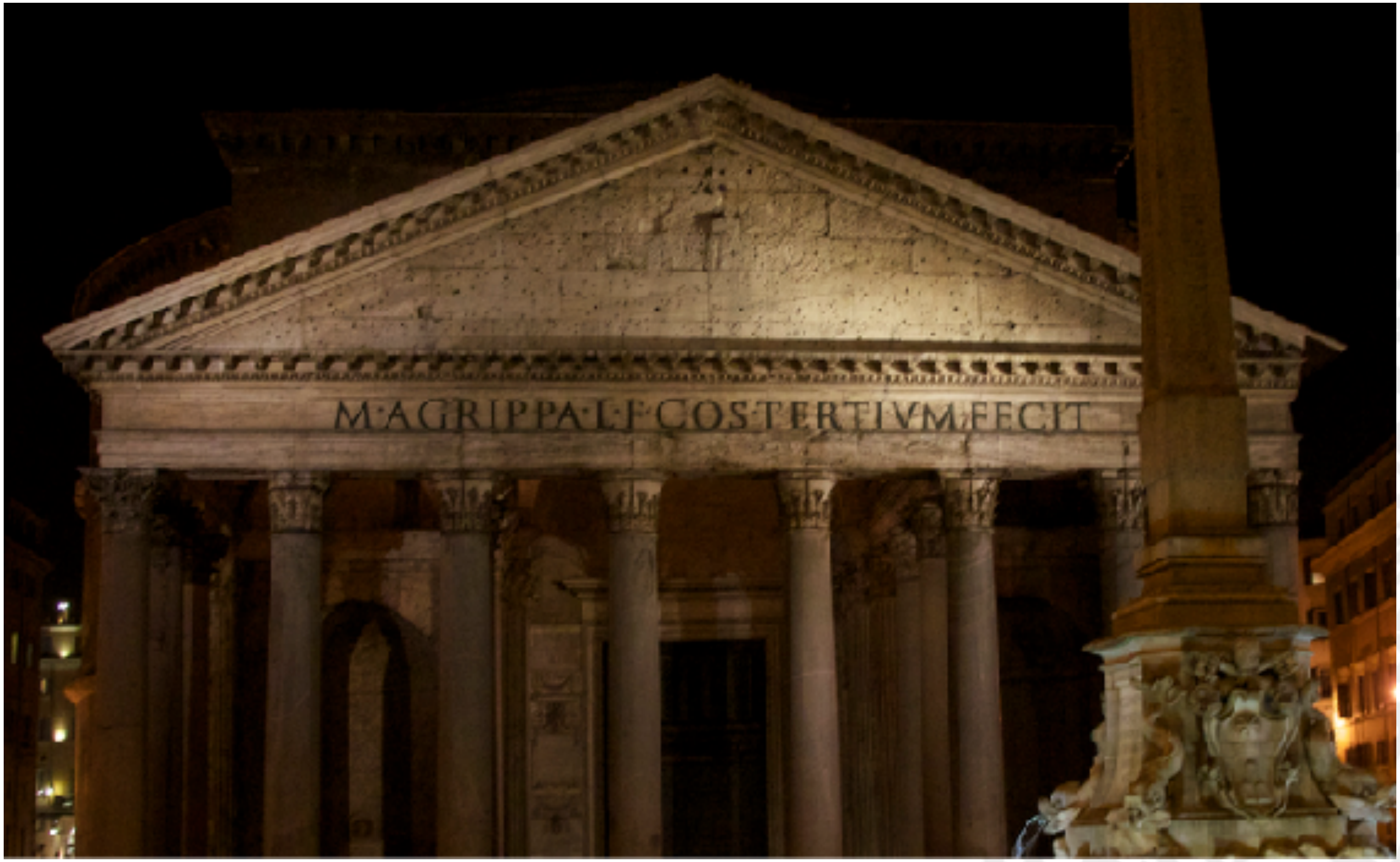
= bre..., ber..., ...ub
= cum, con, cen...
= condam (quondam)
= de..., der..., ...ud
= haec, hoc, her...
= vel, ul..., ...el
= men..., mun ..
= non, nun...
= obiit
= per, par..., por...
= prae, pre...[14]
= propter, papa[15]
= qui[16]
= quae
= quoque
= quam
= ter..., tem..., ten...
= ven..., ver...

http://www.dummies.com/languages/italian/texting-and-chatting-in-italian/
https://kuscholarworks.ku.edu/bitstream/handle/1808/1821/47cappelli.pdf

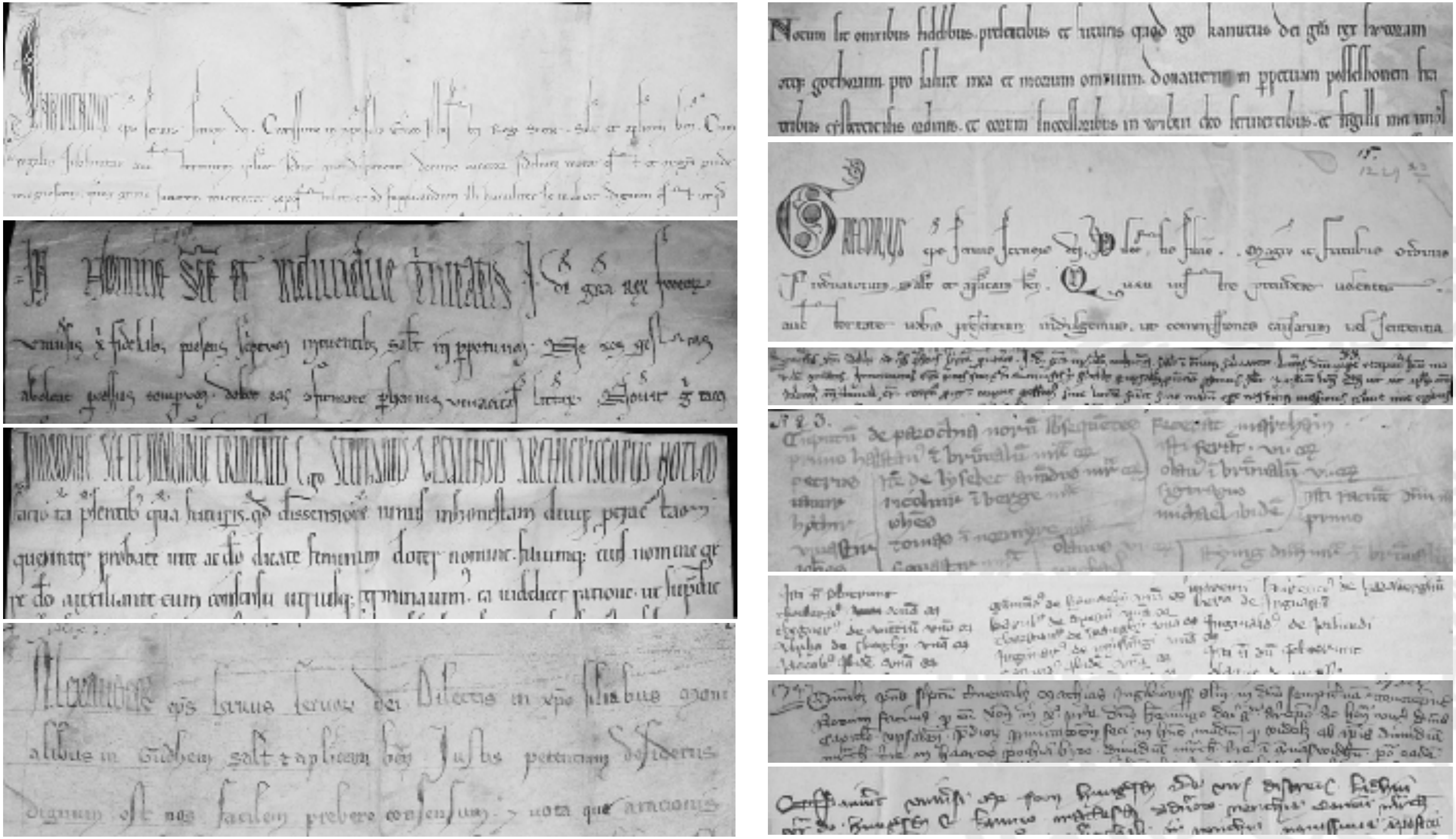
Actually it's not a new thing at all

"M[arcus] Agrippa L[ucii] f[ilius] co[n]s[ul] tertium fecit"

Marcus Agrippa, son of Lucius, made [this building] when consul for the third time

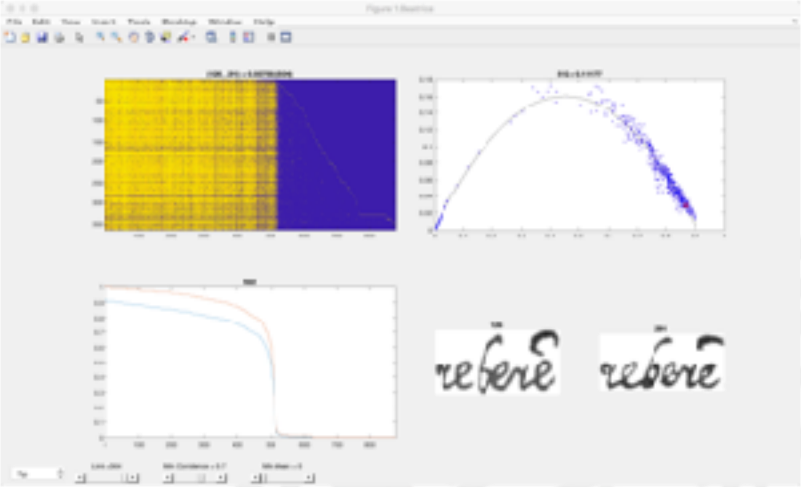


Text and Writing Variations

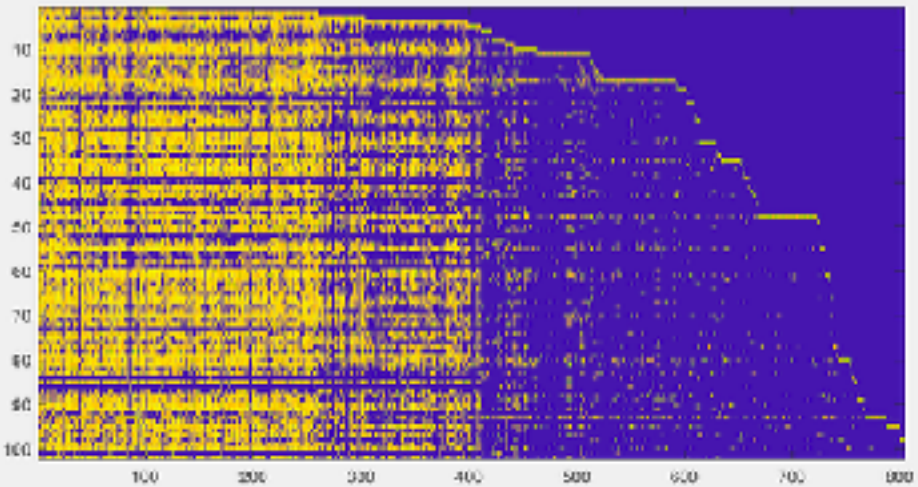


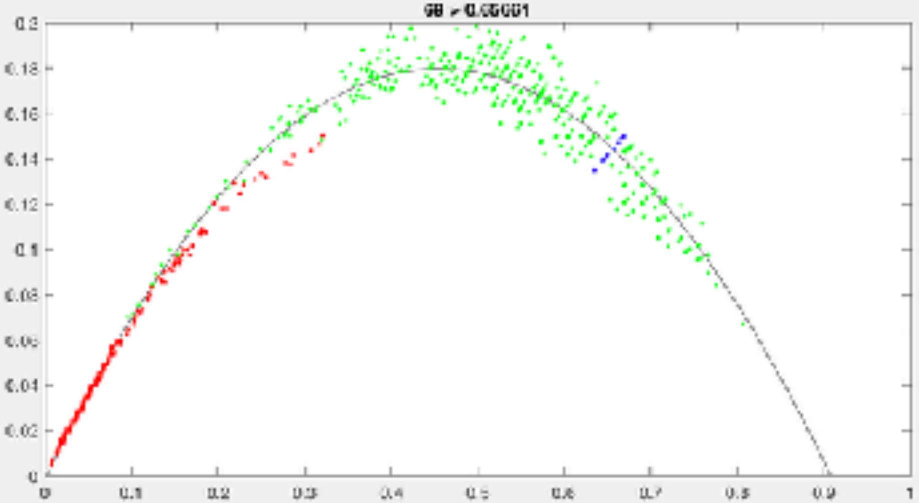
Query Expansion

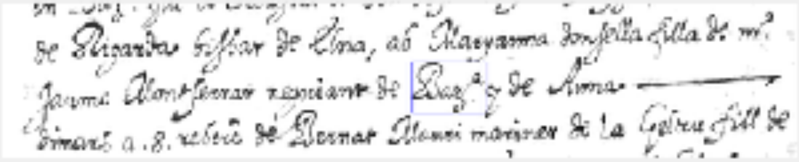
- Search for a word
- Repeat the search for each and every word found
- Old thing… but we do not have a feature for each word to use in order to improve there result
- Use the Visualisation tool: Beatrice



‘Human in the loop’





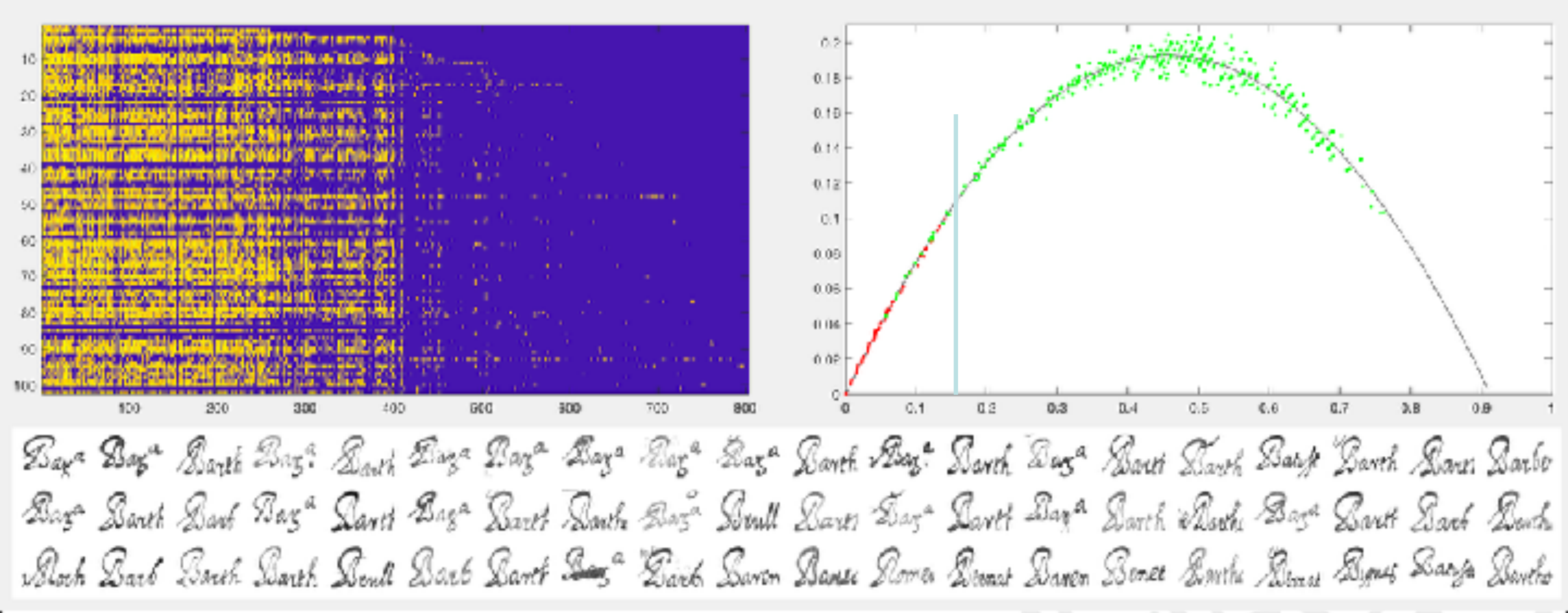


de Rigarda bisbar de Elna, ab Maryanna donsella filla de mº
Jaume Montserrat negociant de Barª y de Anna
dimarts a .8. reberē de Bernat Alemi mariner de la Geltru fill de

Barª Barª Barª
Barª Barª Barª
Barª Barª Barª

Barª

Results



10
20
30
40
50
60
70
80
90
100
100
200
300
400
500
600
700
800
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Why?

- Use the tool to find a cut between inliers and outliers
- Examine further:
  - the worst ‘correct’ and
  - the best ‘incorrect’
  - in doubt: look at the sentence where the word occurs
- Let the system LEARN from this
- The advantage is that it is MUCH easier than looking through the 50 pages to check…
- Non found words will be “left over” after complete transcription

Conclusion

- WS can be used to make non transcribed collections searchable
- It can also be an efficient tool for making transcriptions easier
- Use visualisations to **solve** the problem:
  - correct vs. incorrect words
- Still many interesting challenges!