

Transkribus: Automated Text Recognition for historical documents

READ

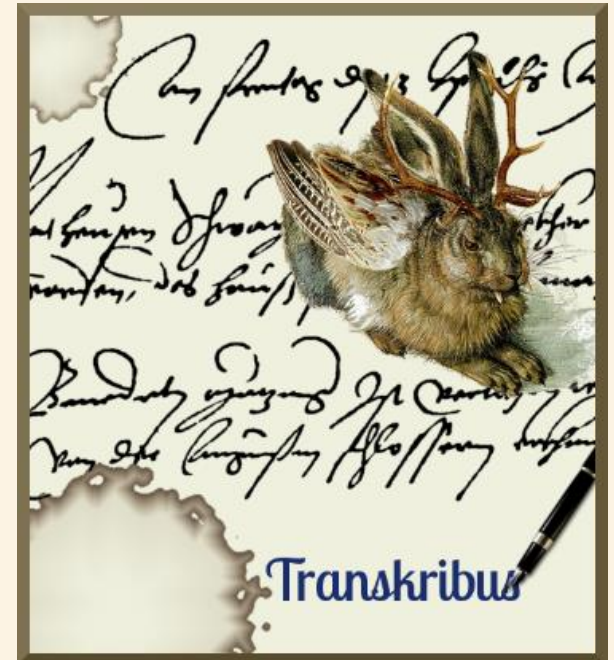


Dr Louise Seaward
Bentham Project, University College
London

@Transkribus

- Recognition and Enrichment of Archival Documents project
- Coordinated by University of Innsbruck – plus 13 other partners
- Building **Transkribus** as new research infrastructure
- Automated transcription and searching of handwritten and printed historical documents
- Services provided free of charge
- 10,000 Transkribus users

READ



Other useful sites

<https://transkribus.eu/>

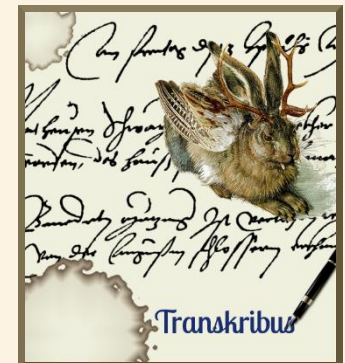
<https://read.transkribus.eu/>

<https://transkribus.eu/wiki>

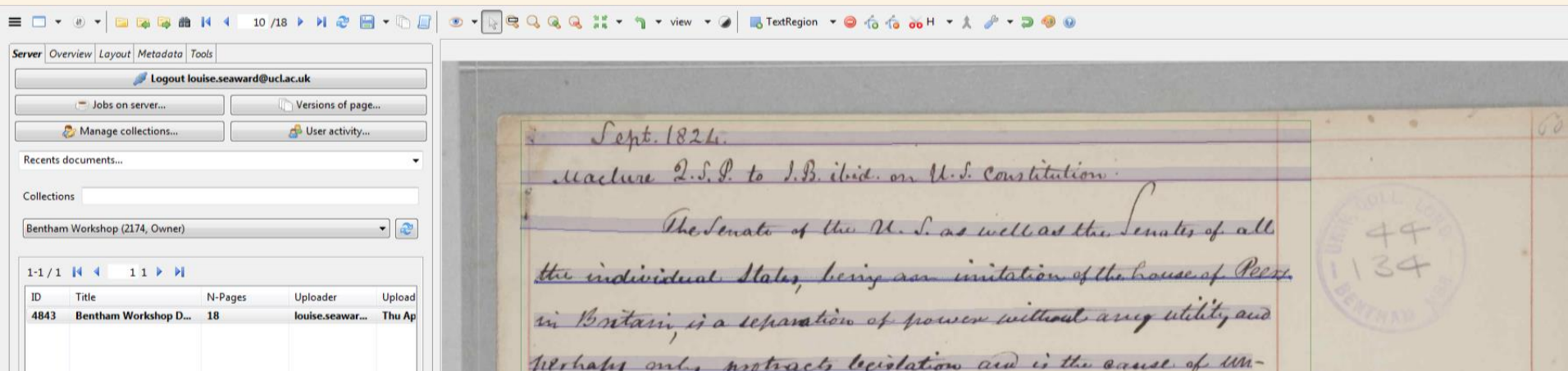
@Transkribus



READ



- **Automated Text Recognition (ATR)** and other tools
 - Enabling computers to **automatically transcribe and search** handwritten historical documents – of any date, language and layout!



The screenshot shows a digital document viewer interface. On the left, there is a sidebar with a table of document metadata. The main area displays a handwritten document page with a blue highlight over a section of text. A circular stamp is visible on the right side of the document page.

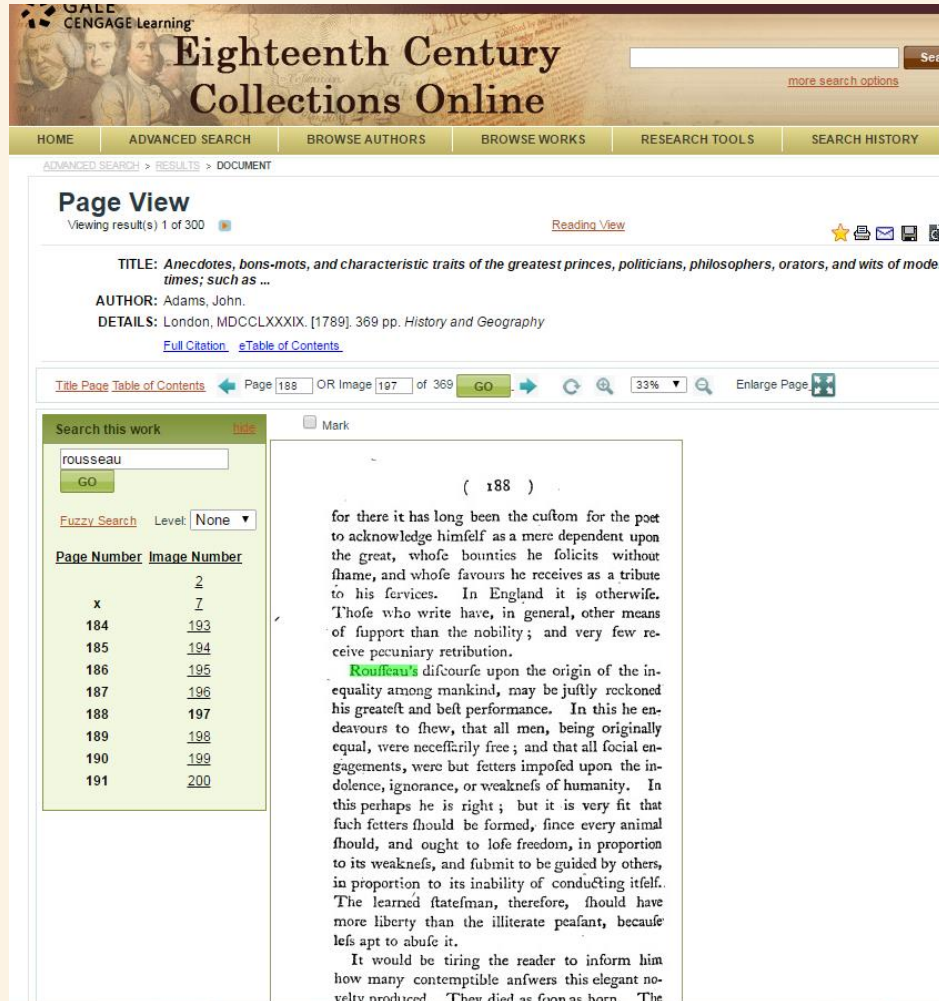
| ID | Title | N-Pages | Uploader | Upload |
|------|-----------------------|---------|-----------------|--------|
| 4843 | Bentham Workshop D... | 18 | louse.seawar... | Thu Ap |

Handwritten text on the document page:

Sept. 1824.
 Lecture 2. S. P. to J. B. ibid. on U. S. Constitution.
 The Senate of the U. S. as well as the Senates of all
 the individual States, being an imitation of the House of Peers
 in Britain, is a separation of power without any utility, and
 perhaps only obstructs legislation and is the cause of un-

Stamp: UCL LIBRARY 49 134 BENTHAM WORKSHOP

Optical Character Recognition (OCR)



The screenshot displays the ECCO website interface. At the top, there is a navigation bar with links for HOME, ADVANCED SEARCH, BROWSE AUTHORS, BROWSE WORKS, RESEARCH TOOLS, and SEARCH HISTORY. Below this, the page title is "Page View" with a sub-header "Viewing result(s) 1 of 300". The main content area shows the following details:

TITLE: *Anecdotes, bons-mots, and characteristic traits of the greatest princes, politicians, philosophers, orators, and wits of modern times; such as ...*

AUTHOR: Adams, John.

DETAILS: London, MDCCLXXXIX. [1789]. 369 pp. *History and Geography*

Below the details, there is a search bar with the text "rousseau" and a "GO" button. A "Fuzzy Search" section is also visible. To the right of the search bar is a table with two columns: "Page Number" and "Image Number".

| Page Number | Image Number |
|-------------|---------------------|
| | 2 |
| x | 7 |
| 184 | 193 |
| 185 | 194 |
| 186 | 195 |
| 187 | 196 |
| 188 | 197 |
| 189 | 198 |
| 190 | 199 |
| 191 | 200 |

The main text area shows the OCR of the document page, which is page 188. The text reads:

(188)

for there it has long been the custom for the poet to acknowledge himself as a mere dependent upon the great, whose bounties he folicits without flhame, and whose favours he receives as a tribute to his services. In England it is otherwise. Thofe who write have, in general, other means of fupport than the nobility; and very few receive pecuniary retribution.

Rouffeau's difcourfe upon the origin of the inequality among mankind, may be juftly reckoned his greateft and beft performance. In this he endeavours to fhew, that all men, being originally equal, were neceffarily free; and that all focial engagements, were but fetters impofed upon the indolence, ignorance, or weaknefs of humanity. In this perhaps he is right; but it is very fit that fuch fetters fhould be formed, fince every animal fhould, and ought to love freedom, in proportion to its weaknefs, and fubmit to be guided by others, in proportion to its inability of conducting itfelf. The learned ftatesman, therefore, fhould have more liberty than the illiterate peafant, becaufe lefs apt to abufe it.

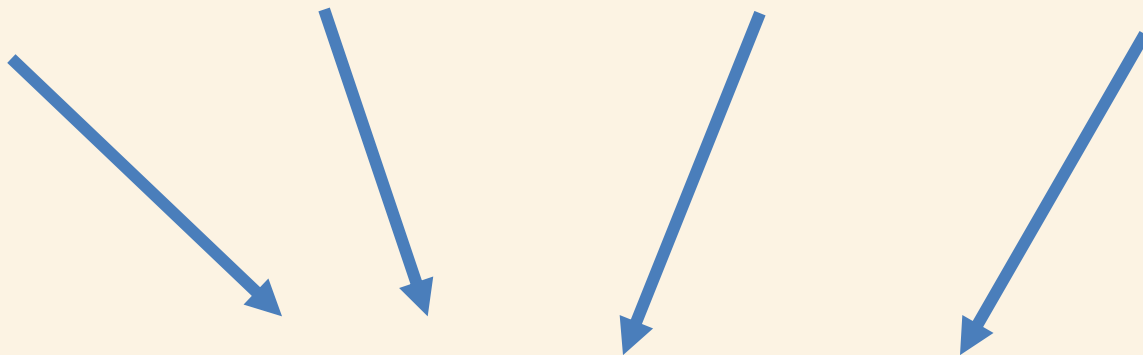
It would be tiring the reader to inform him how many contemptible anfwers this elegant novelty produced. They died as foon as born. The

l

e

q

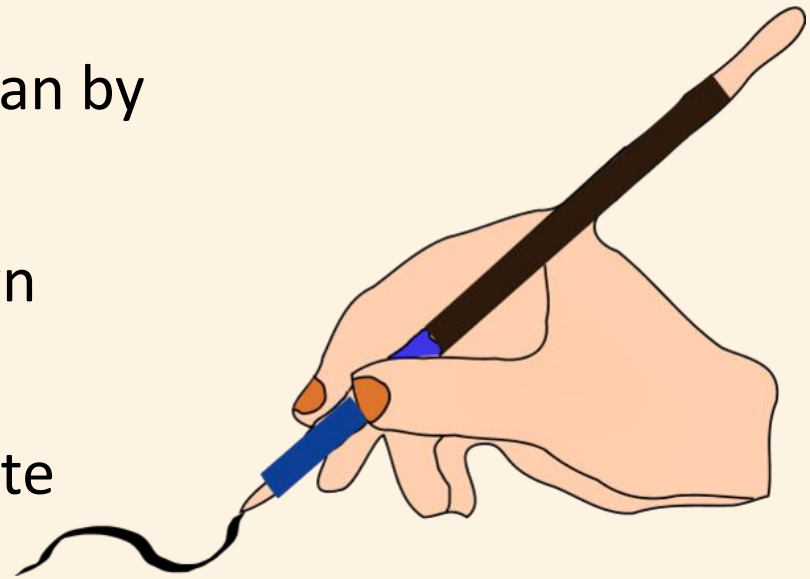
l



lequel

Automated Text Recognition

- Machine learning using neural networks
- Processes writing by line, rather than by character
- Needs to be trained by being shown document images and transcripts
- More training data → more accurate recognition
- Train a model to transcribe and search a collection of documents



Creating training data

- Start with at least 15,000 words (75 pages)
- Prepare training data in 3 stages:
 - 1. Upload images to Transkribus**
 - 2. Segment images into lines**
 - 3. Transcribe each page accurately**
- The Transkribus team will use this training data to train a model to recognise your collection



If you have existing transcripts, you can also use these to train a model!

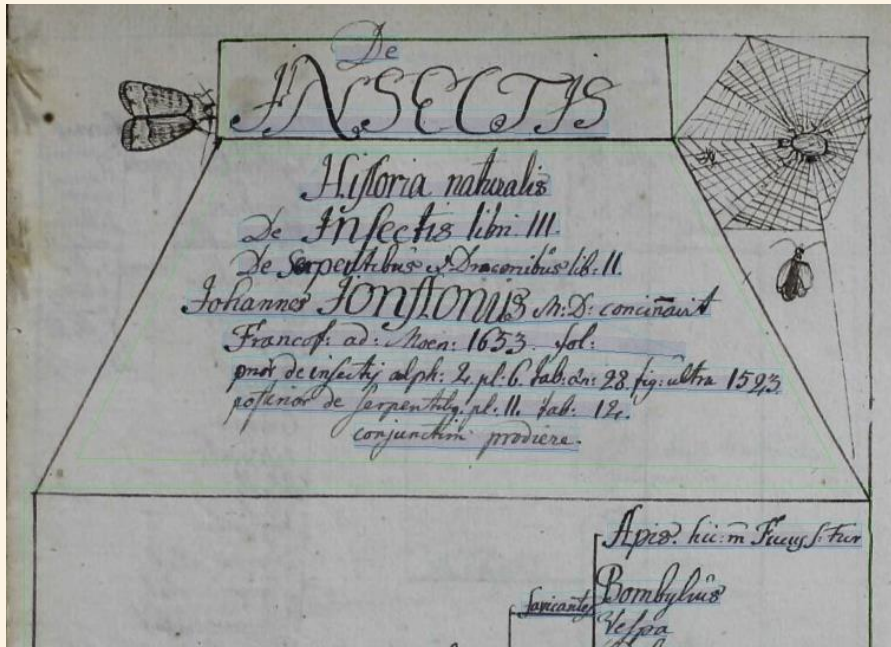
Check the Transkribus wiki for more
info

<https://transkribus.eu/wiki>

Automated Text Recognition models

- Once training is complete, access your model in Transkribus
- 300+ models already trained
- Apply your model to other pages from the same collection – **transcribe and search**
- Foundation for further research and scholarly editing
- Work in teams, correct and edit transcripts, add tags and metadata, export transcripts

Accuracy?



- Measure accuracy in Transkribus via Character Error Rate (CER)
- Best results = transcripts with CER of 10% or less, i.e. 90% of characters in a transcript are correct
- These transcripts can be understood, searched and corrected quickly!

Bentham model

- Based on Jeremy Bentham's papers (c.18-19 English philosopher)
- Written by Bentham and his secretaries
- Trained on 800 pages
- 5-10% CER is possible
- Working on a new model based on Bentham's most difficult handwriting – 28% CER



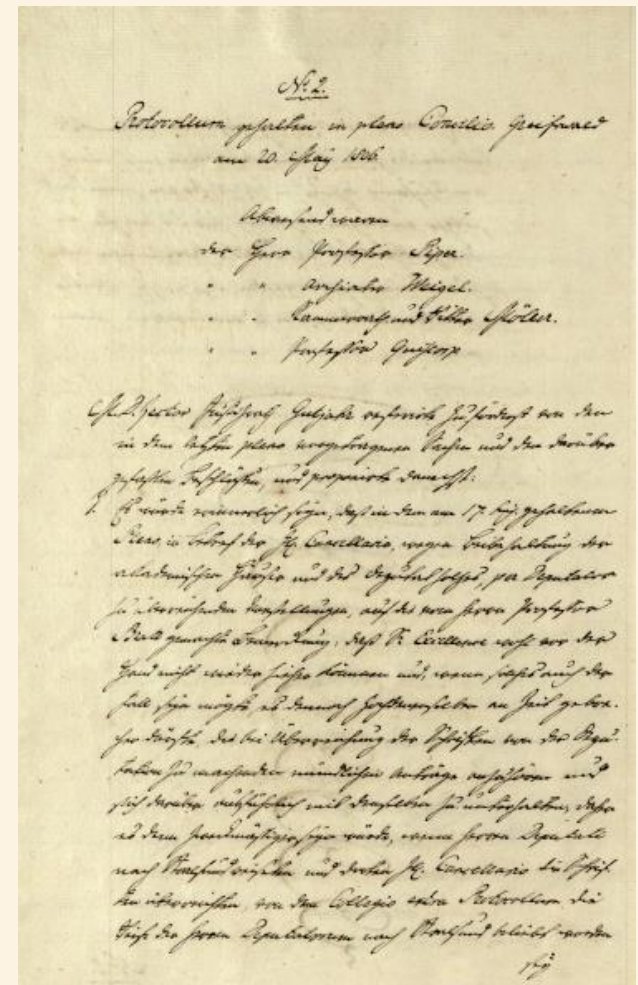


circumstance. Of all figures, however, this you will observe, is the only one that affords a perfect view, and the same view, of an indefinite number of apartments of the same dimensions: that affords a spot from which, without any change of situation, a man may survey, in the same perfection, the whole number, and without so much as a change of posture, the half of the ^{whole} ~~same~~ number, at the same time: that, within a boundary of a given extent, contains the greatest quantity of rooms: that places the center at the least distance from the light: that gives the Cells most width, at the part where, on account of the light, most

0 circumstance. Of all figures, however, this, you will observe, is the only one that ↵
 1 affords a perfect view, and the same view, of an indefinite number of apartments ↵
 2 of the same dimensions: that affords a spot from which, without any, change ↵
 3 of situation, a man may survey, in the same perfection, the whole number, ↵
 4 whole ↵
 5 and without so much as a change of posture, the half of the are numteyar ↵
 6 the same time that, within a boundary of a given extent, contains the greater ↵

Konzilsprotokolle model

- Papers from University of Greifswald archive (c.18-19 German)
- Minutes of University council – written in several hands
- Trained on 2,700 pages (410,000 words)
- CER of less than 5% is possible
- Results now integrated in archival repository



View

- Image view
- Contents
- Thumbnail gallery
- Bibliographic data
- Full text
- Named Entities
- DFG-Viewer
- OPAC

Search in: Konzilsprotokolle 1806 - 1807

Named Entities

Overview Area

Tag range: 10 pages Filter tags:

- 1-10 Stralsund Candidatus Cancellario Medow Medo
- 11-20 Albrecht Julius Greifswald Cancellario M Ziems-
- 21-30 Jordan Berlin Stockholm Greifswald Friedrich O
- 31-40 Haselberg Haselberg Conclusa Senioratus Wall
- 41-48 Rudolphi Rudolphi

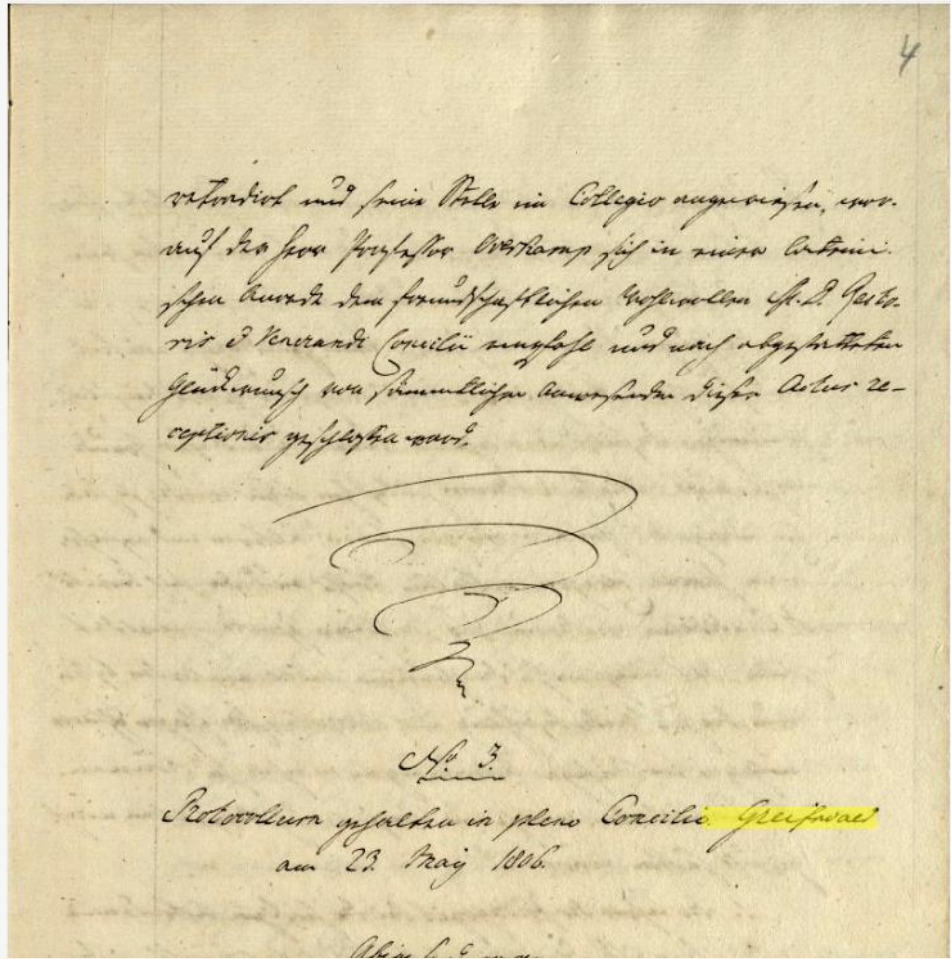
Search hits

1 / 10

Back to hit list

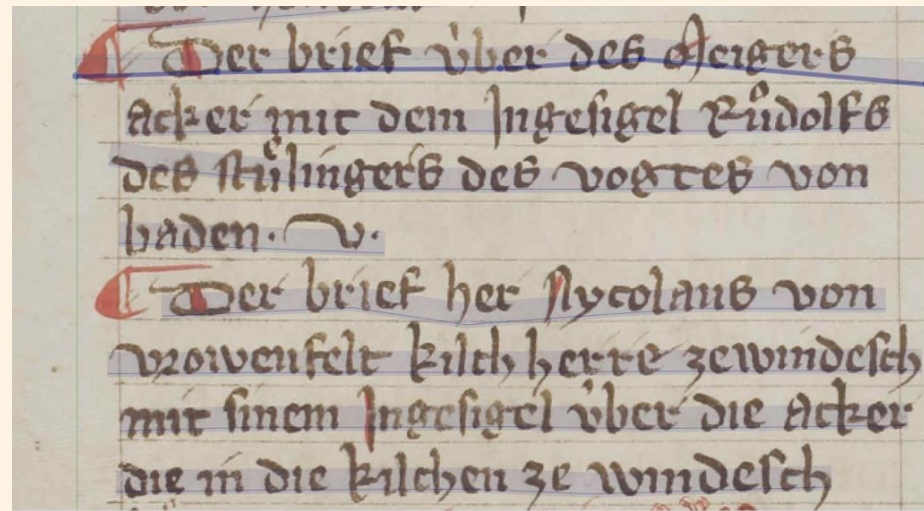
Konzilsprotokolle 1806 - 1807

Navigation icons: Home, Previous, 9:4r, Next, Full Screen, Rotate, Refresh, Zoom In, Zoom Out, and a slider.



Königsfelden model

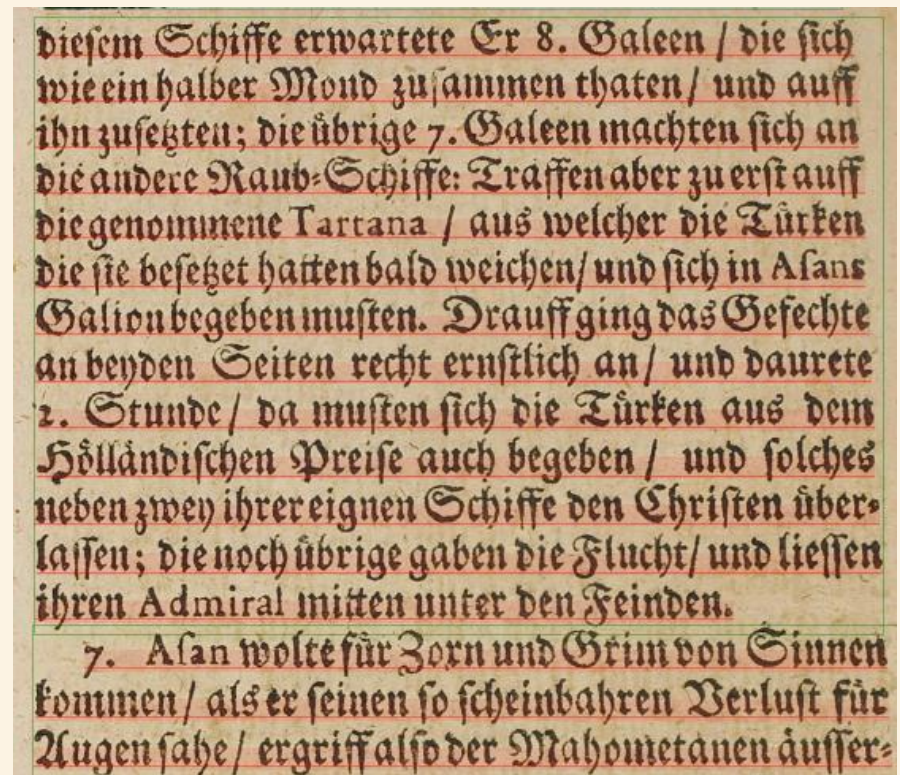
- Charters from a 14th century abbey in Switzerland (Gothic script)
- Written by 2-3 scribes
- University of Zurich and State Archives of Zurich
- Trained on 26,000 words
- Character Error Rate of 10%
- Can deal with abbreviations and unusual symbols



¶Der·brief·über·des·Meigers↵
 acker·mit·dem·ingesigel·Rüdolfs↵
 des·stülingers·des·vogres·von↵
 baden.v↵
 Der·brief·her·scolaus·von↵
 vrowenfelt·kilch·herre·zewindesch↵

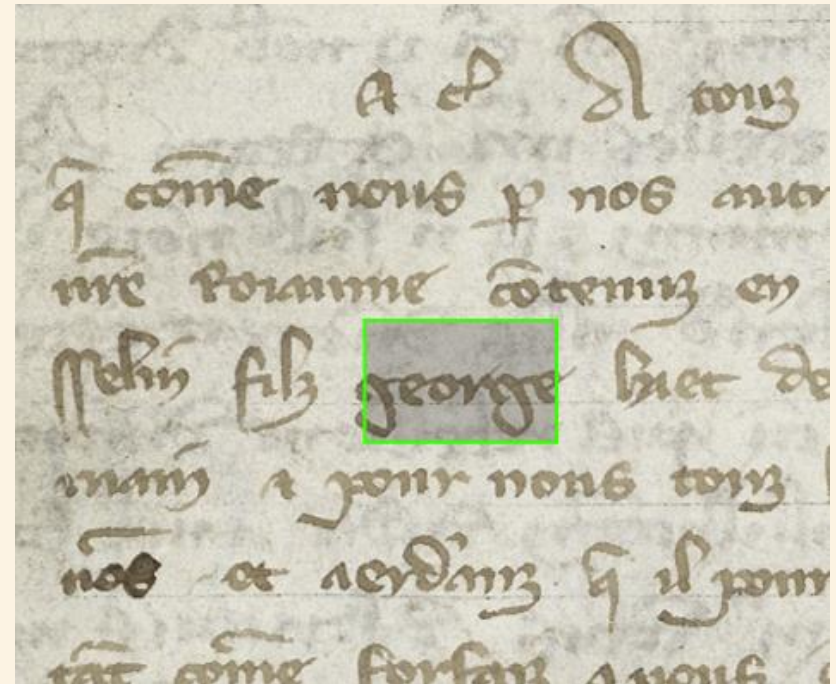
Recognising printed text

- Neural networks can also process printed text – with less training data!
- Transcribe documents or use OCR engine in Transkribus to create training data
- Only 5000 words of training data needed
- Results with 1-2% Character Error Rate are possible



Keyword Spotting

- Sophisticated form of keyword searching
- Detects similarities in images of words, rather than transcripts
- Searches through probability values attributed to each character
- Can work with outputs with higher error rates – 30% CER
- Can make precise or broad searches to find all possible matches



Benefits of our network



- The more users, the stronger the technology becomes
- Your documents remain private
- General models for English, German, Dutch etc. will be possible in the future
- Join us as an MOU partner or at our next conference in November 2018
- Sustainability plan for Transkribus after end of the READ project – freemium model

Thank you for listening!

<https://transkribus.eu/>

<https://read.transkribus.eu/>

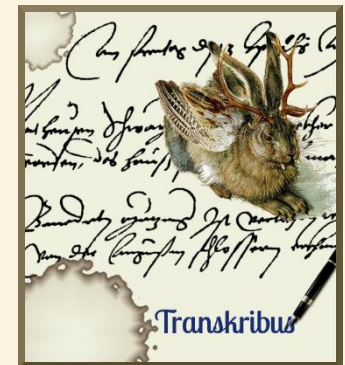
<https://transkribus.eu/wiki/>

email@transkribus.eu

@Transkribus

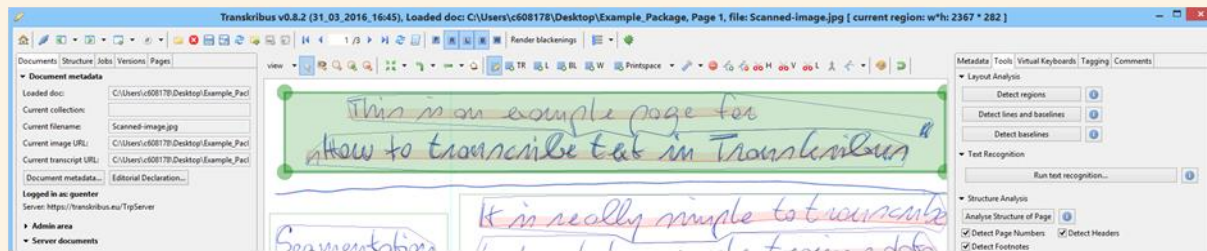


READ



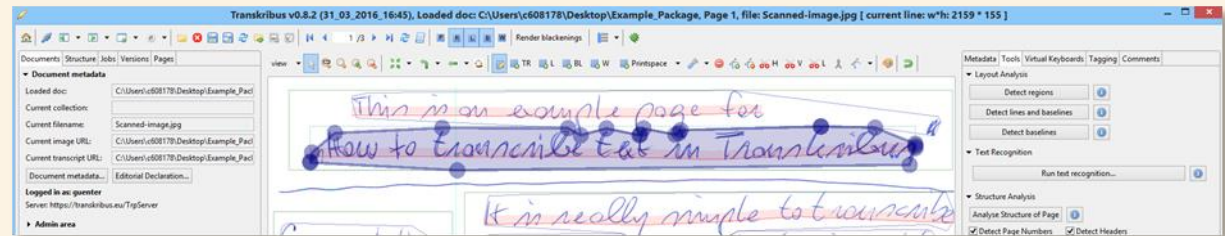
Segmentation

Lines in the manuscript image need to be **segmented** to connect them to the lines of the text transcript.



← **Text regions**

Line regions



← **Baselines**

