i|d|e

UNI GRAZ

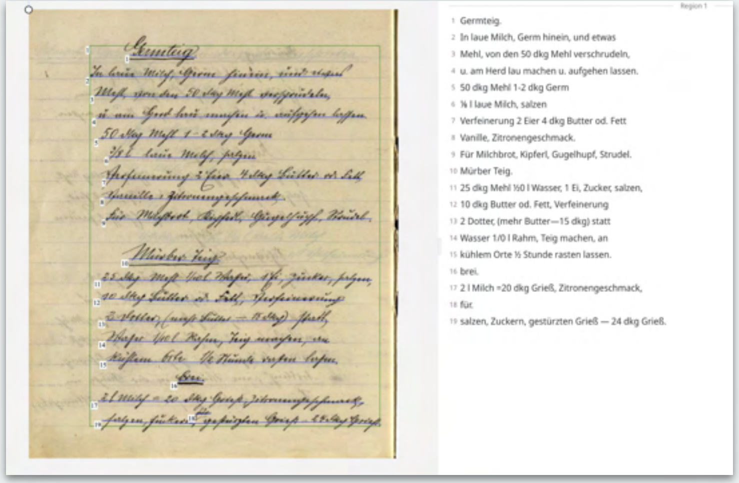
From Image to Transcription

Selina Galka

Master Class

‚Digital Scholarly Editing‘ 2024

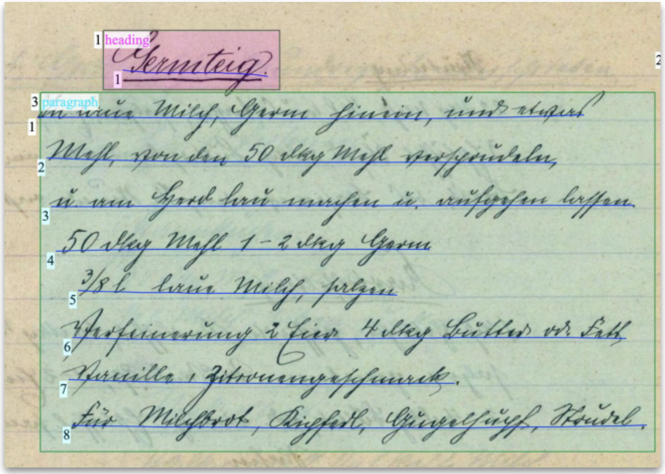
Universität des Saarlandes

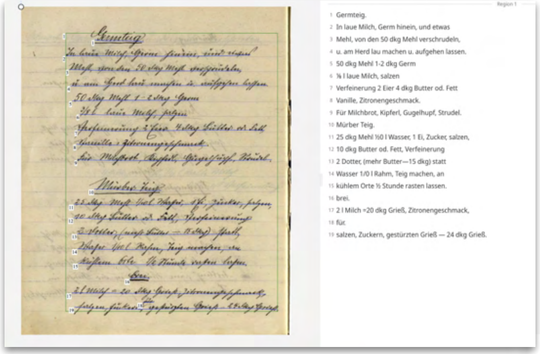


Germteig.
In laue Milch, Germ hinein, und etwas
Mehl, von den 50 dkg Mehl verschrudeln,
u. am Herd lau machen u. aufgehen lassen.
50 dkg Mehl 1-2 dkg Germ
⅛ l laue Milch, salzen
Verfeinerung 2 Eier 4 dkg Butter od. Fett
Vanille, Zitronengeschmack.
Für Milchbrot, Kipferl, Gugelhupf, Strudel.
Mürber Teig.
25 dkg Mehl ½0 l Wasser, 1 Ei, Zucker, salzen,
10 dkg Butter od. Fett, Verfeinerung
2 Dotter, (mehr Butter—15 dkg) statt
Wasser 1/0 l Rahm, Teig machen, an
kühlem Orte ½ Stunde rasten lassen.
brei.
2 l Milch =20 dkg Grieß, Zitronengeschmack,
für
salzen, Zuckern, gestürzten Grieß — 24 dkg Grieß.

From Image to Text

1. General introduction to transcribing
    a. transcription + editorial guidelines
    b. manually vs. automatic (HTR)
    c. transcription tools
2. Transkribus
    a. General Introduction
    b. Automatic Text Recognition (Applying a model)
    c. Training a HTR model
    d. Tagging
    e. Export
    f. Publishing
3. Conclusion and Resources





Germteig.
In laue Milch, Germ hinein, und etwas
Mehl, von den 50 dkg Mehl verschrudeln,
u. am Herd lau machen u. aufgehen lassen.
50 dkg Mehl 1-2 dkg Germ
⅜ l laue Milch, salzen
Verfeinerung 2 Eier 4 dkg Butter od. Fett
Vanille, Zitronengeschmack.
Für Milchbrot, Kipferl, Gugelhupf, Strudel.
Mürber Teig.
25 dkg Mehl ½0 l Wasser, 1 Ei, Zucker, salzen,
10 dkg Butter od. Fett, Verfeinerung
2 Dotter, (mehr Butter—15 dkg) statt
Wasser 1/0 l Rahm, Teig machen, an
kühlem Orte ½ Stunde rasten lassen.
brei.
2 l Milch =20 dkg Grieß, Zitronengeschmack,
für
salzen, Zuckern, gestürzten Grieß — 24 dkg Grieß.

Transcription

Transcribing

In editorial studies, **transcription** is understood as the transfer of a historical source text into a modern medium, nowadays usually **machine-readable text**.

machine-readable → searchable, further processable with computers (automatic processing, manipulating, analysis)

“The result of a transcription is based on the **specific questions** and historically evolved **guidelines** of the individual discipline.”

Klug, Helmut W. 2021. *Transkription*. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". Aufgerufen am: 8.2.2024. Handle: hdl.handle.net/11471/562.50.197. PID: o:konde.197

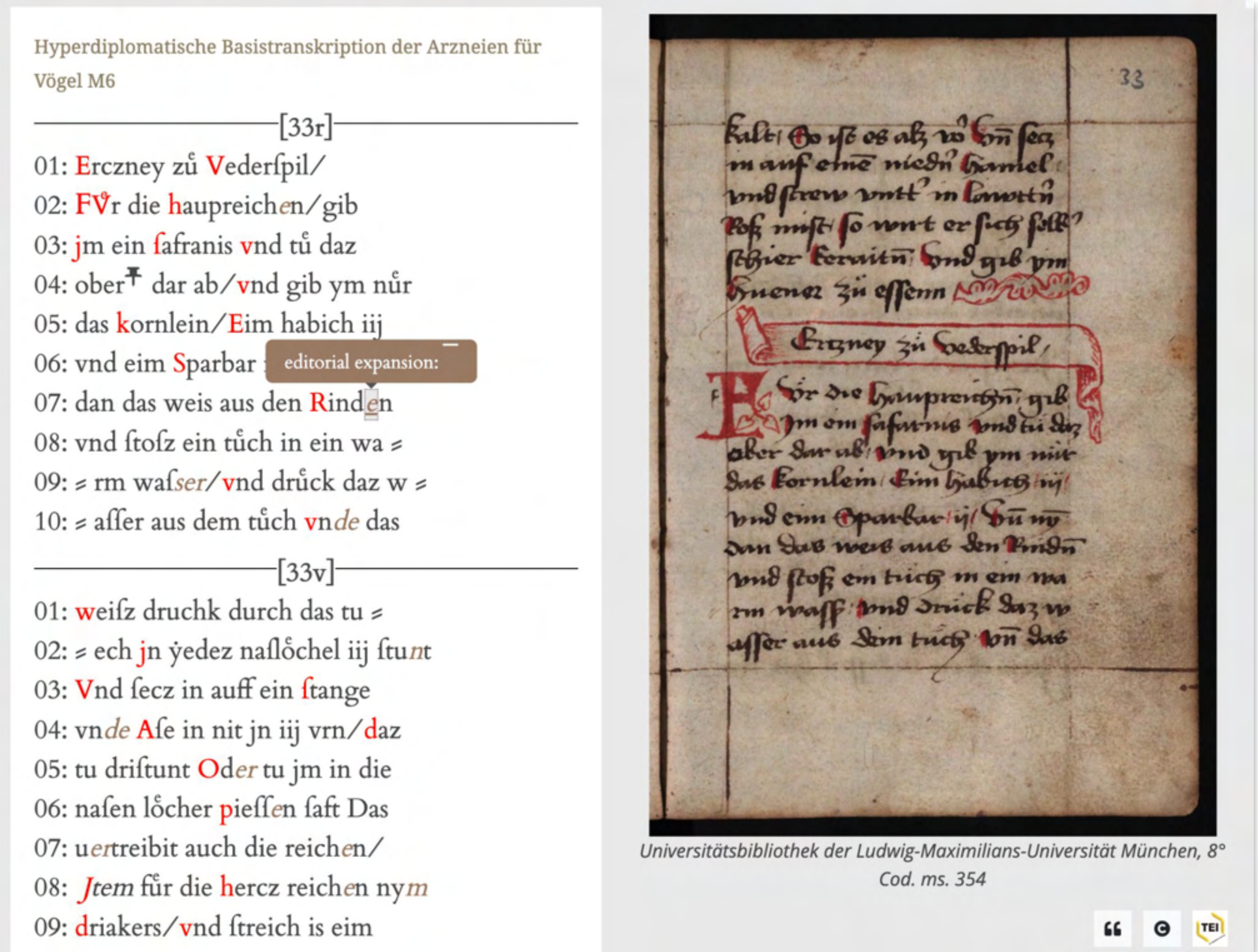
→ edition guidelines

Transcription guidelines / editorial guidelines

- depending on the type of the text and the intended audience of the edition, editorial decisions must be made
- these should be explained in the transcription/editing guidelines (maintenance of a guidelines document during the transcription process is recommended)

- Example questions:
  - Are the writer's errors documented, and if so, is there a classification of these errors?
  - Are abbreviations reproduced or resolved (tacitly)?
  - How are headings or rubrics, if any, presented in the text?

Example of a diplomatic transcription



Hyperdiplomatische Basistranskription der Arzneien für Vögel M6

[33r]

01: Erczney zuͤ Vederſpil/
02: FVͤr die haupreich*e*n/gib
03: jm ein ſafranis vnd tuͤ daz
04: ober† dar ab/vnd gib ym nuͤr
05: das kornlein/Eim habich iij
06: vnd eim Sparbar
07: dan das weis aus den Rinden
08: vnd ſtoſz ein tuͤch in ein wa ⸗
09: ⸗ rm waſ*ser*/vnd druͤck daz w ⸗
10: ⸗ aſſer aus dem tuͤch vn*de* das

editorial expansion:

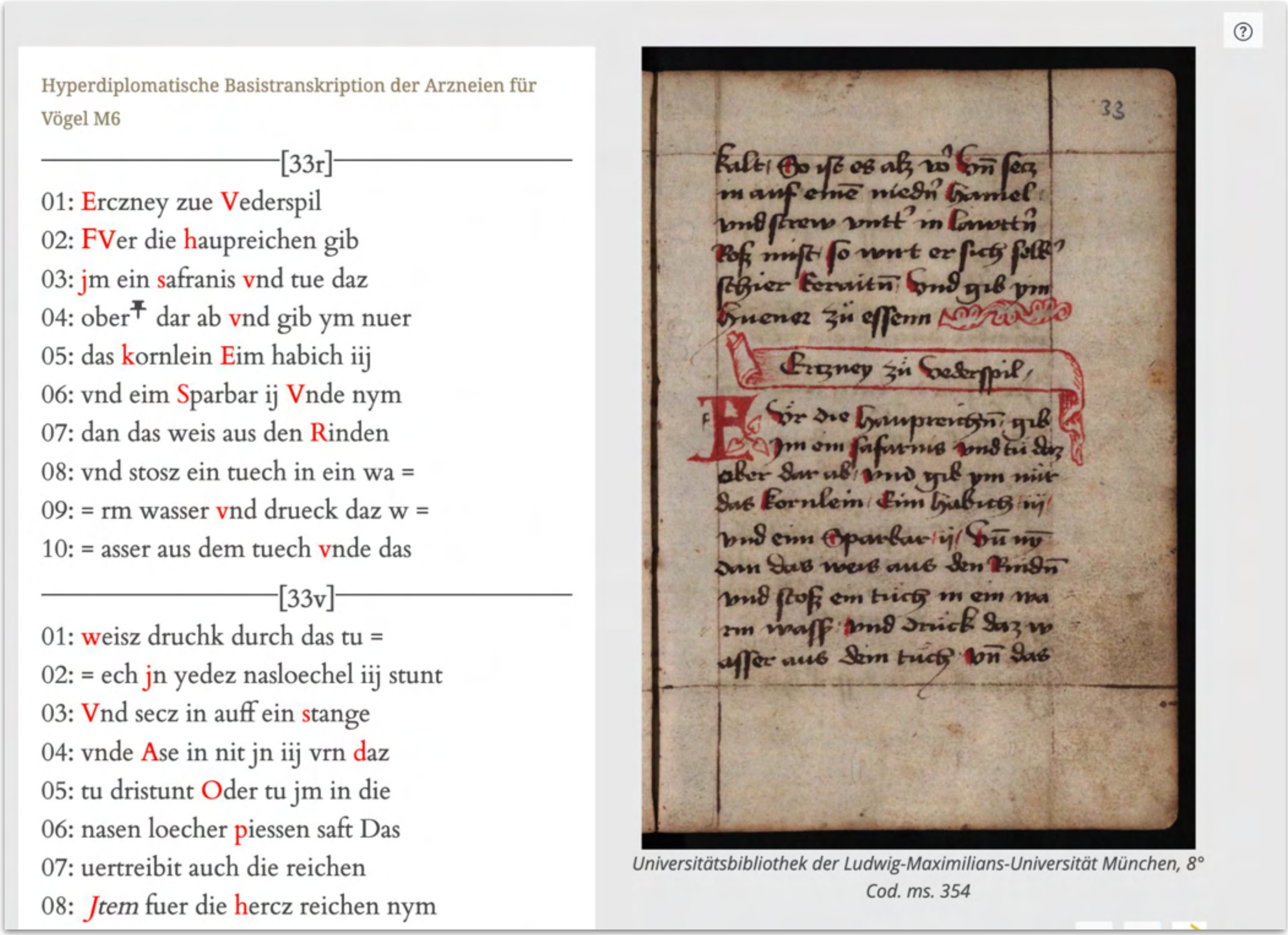
[33v]

01: weiſz druchk durch das tu ⸗
02: ⸗ ech jn ẏedez naſloͤchel iij ſtu*n*t
03: Vnd ſecz in auff ein ſtange
04: vn*de* Aſe in nit jn iij vrn/daz
05: tu driſtunt Od*er* tu jm in die
06: naſen loͤcher pieſſ*e*n ſaft Das
07: u*er*treibit auch die reich*e*n/
08: *Jtem* fuͤr die hercz reich*e*n ny*m*
09: driakers/vnd ſtreich is eim

Universitätsbibliothek der Ludwig-Maximilians-Universität München, 8° Cod. ms. 354

Schultz-Balluff, Simone; Bülters, Timo; George, Anaïs; Orfgen, Lukas. 2022. Hyperdiplomatische Basistranskription der Arzneien für Vögel M6, In: Hyperdiplomatische Transkriptionsplattform. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka. GAMS. PID: o:hyper.jagdM6.3 (Accessed 2024-02-16)

Normalized version



Hyperdiplomatische Basistranskription der Arzneien für Vögel M6

[33r]

01: Erczney zue Vederspil
02: FVer die haupreichen gib
03: jm ein safranis vnd tue daz
04: ober₸ dar ab vnd gib ym nuer
05: das kornlein Eim habich iij
06: vnd eim Sparbar ij Vnde nym
07: dan das weis aus den Rinden
08: vnd stosz ein tuech in ein wa =
09: = rm wasser vnd drueck daz w =
10: = asser aus dem tuech vnde das

[33v]

01: weisz druchk durch das tu =
02: = ech jn yedez nasloechel iij stunt
03: Vnd secz in auff ein stange
04: vnde Ase in nit jn iij vrn daz
05: tu dristunt Oder tu jm in die
06: nasen loecher piessen saft Das
07: uertreibit auch die reichen
08: *Jtem* fuer die hercz reichen nym

Universitätsbibliothek der Ludwig-Maximilians-Universität München, 8° Cod. ms. 354

Schultz-Balluff, Simone; Bülters, Timo; George, Anaïs; Orfgen, Lukas. 2022. Hyperdiplomatische Basistranskription der Arzneien für Vögel M6, In: Hyperdiplomatische Transkriptionsplattform. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka. GAMS. PID: o:hyper.jagdM6.3 (Accessed 2024-02-16)

Transcription: How?

Transcription can be done **manually** (keying, double-keying) or **automatically** (OCR, HTR).

MANUAL TRANSCRIPTION

keying: The manual capture, i.e. typing, of a text in the course of its digitization.

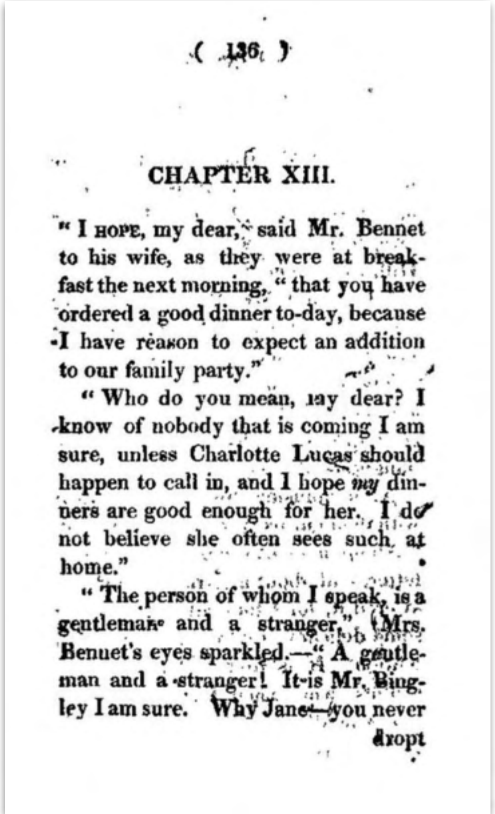
double-keying: Two people type out the content of a document; a computer program then searches for differences between the two versions. Any typing errors found are then corrected by a third person.

Transcription: How?

AUTOMATIC TRANSCRIPTION

OCR (Optical Character Recognition)

- Automatic text recognition of **printed texts**
- i.e. a computer "reads" a scanned document, recognizes and captures the text in it and then generates an electronic version.



(136)

CHAPTER XIII.

"I HOPE, my dear," said Mr. Bennet to his wife, as they were at breakfast the next morning, "that you have ordered a good dinner to-day, because I have reason to expect an addition to our family party."

"Who do you mean, my dear? I know of nobody that is coming I am sure, unless Charlotte Lucas should happen to call in, and I hope *my* dinners are good enough for her. I do not believe she often sees such at home."

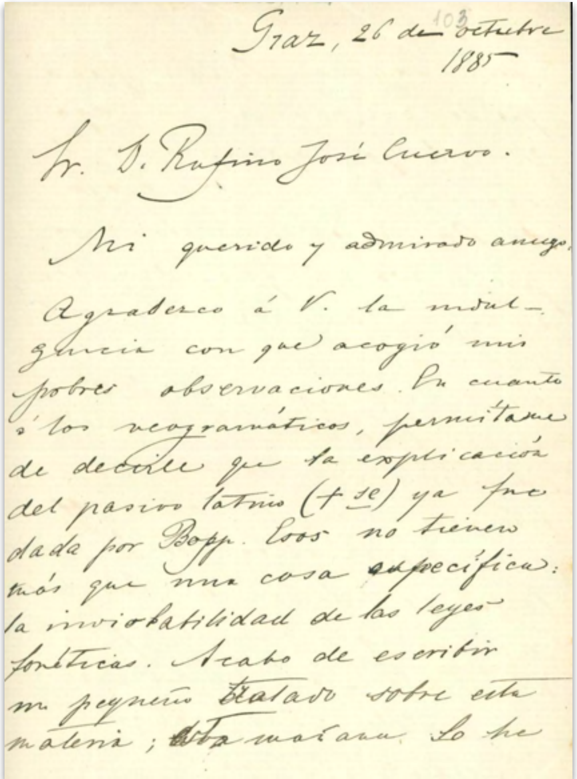
"The person of whom I speak, is a gentleman and a stranger." Mrs. Benuet's eyes sparkled.—"A gentleman and a stranger! It is Mr. Bingley I am sure. Why Jane—you never

dropt

Transcription: How?

HTR (Handwritten Text Recognition)

- converting handwritten text into machine-readable and editable text
- involves using various techniques from artificial intelligence, machine learning, and computer vision to analyze and interpret handwritten documents
- Process:
  - preprocessing / feature extraction / recognition / post-processing



103

Graz, 26 de octubre
1885

Sr. D. Rufino José Cuervo.

Mi querido y admirado amigo.

Agradezco á V. la indulgencia con que acogió mis pobres observaciones. En cuanto á los neogramáticos, permítame de decirle que la explicación del pasivo latino (+ se) ya fue dada por Bopp. Esos no tienen más que una cosa específica: la inviolabilidad de las leyes fonéticas. Acabo de escribir un pequeño tratado sobre esta materia; ~~otra~~ mañana lo he

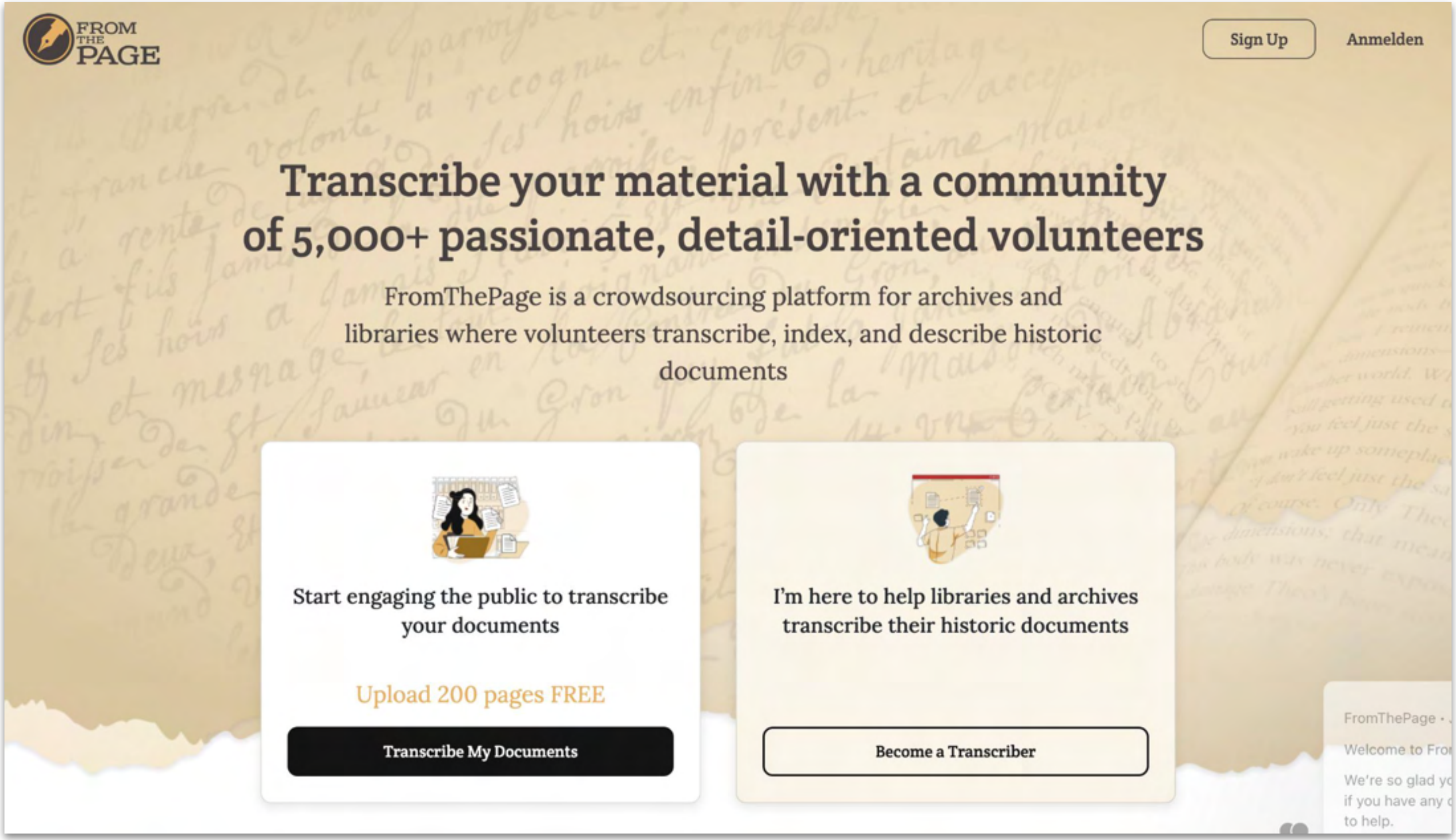
Hugo Schuchardt an Rufino José Cuervo (58-SC384H29). Graz, 26. 10. 1885. Hrsg. von Bernhard Hurch (2023). In: Bernhard Hurch (Hrsg.): *Hugo Schuchardt Archiv*. Online unter https://gams.uni-graz.at/o:hsa.letter.11272, abgerufen am 16. 02. 2024. Handle: hdl.handle.net/11471/518.10.1.11272.

Transcription tools

Transcription tools

- software applications that support the process of transcribing a historical source online or offline
- eg. providing a GUI (Graphical User Interface)
- offering layout and/or text recognition or anchoring the transcribed text in the digital facsimiles with the help of coordinates
- depending on the application the data can be saved in different formats
- a common format for mapping page structures is PAGE-XML
- usually created for individual application purposes

Klug, Helmut W. 2021. *Transkriptionswerkzeuge*. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". Aufgerufen am: 8.2.2024. Handle: hdl.handle.net/11471/562.50.199. PID: o:konde.199



Transcribe your material with a community of 5,000+ passionate, detail-oriented volunteers

FromThePage is a crowdsourcing platform for archives and libraries where volunteers transcribe, index, and describe historic documents

Start engaging the public to transcribe your documents

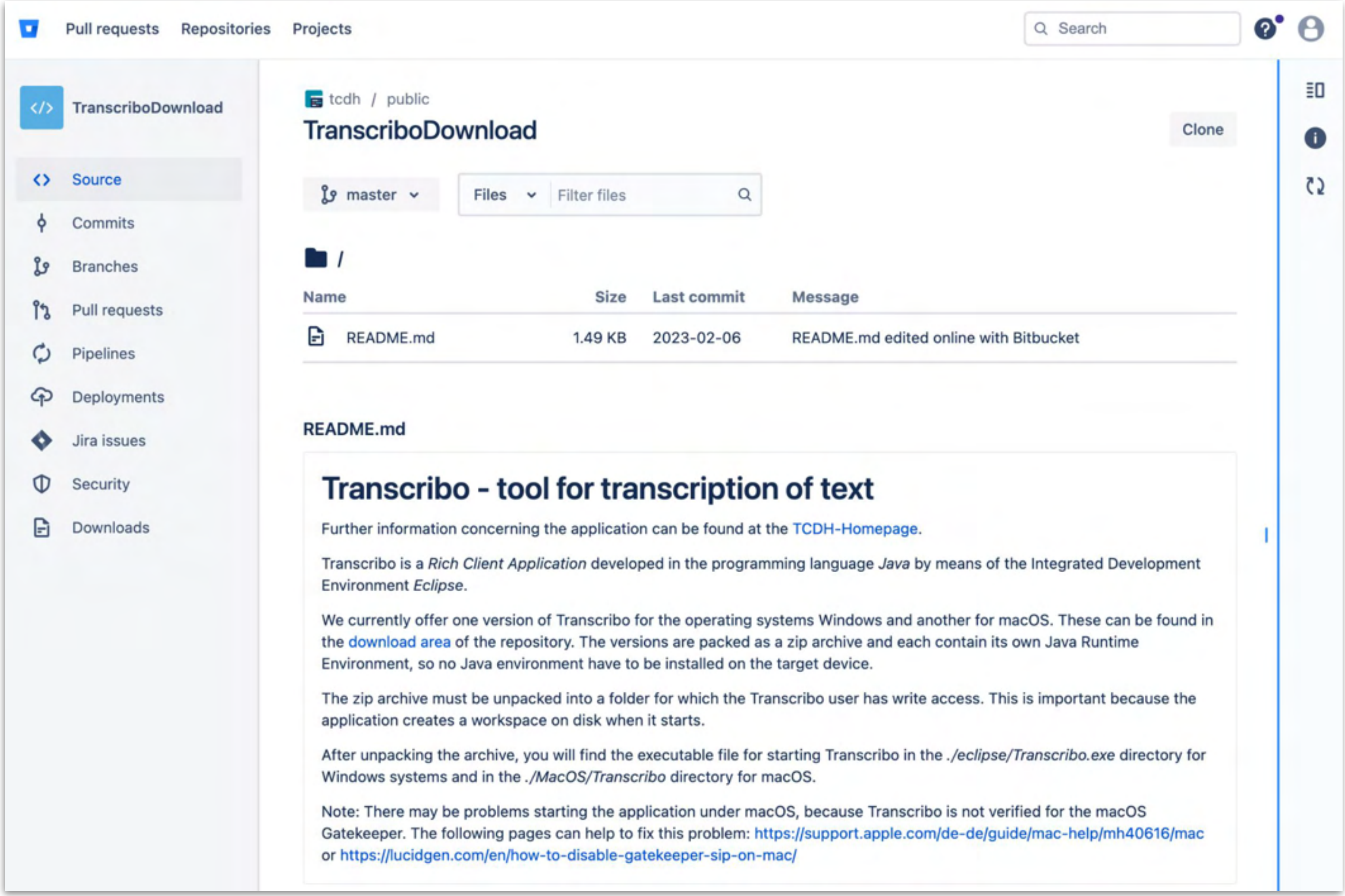
Upload 200 pages FREE

I'm here to help libraries and archives transcribe their historic documents

FromThePage • .

Welcome to Fro

We're so glad y
if you have any
to help.



Name	Size	Last commit	Message
README.md	1.49 KB	2023-02-06	README.md edited online with Bitbucket

README.md

Transcribo - tool for transcription of text

Further information concerning the application can be found at the TCDH-Homepage.

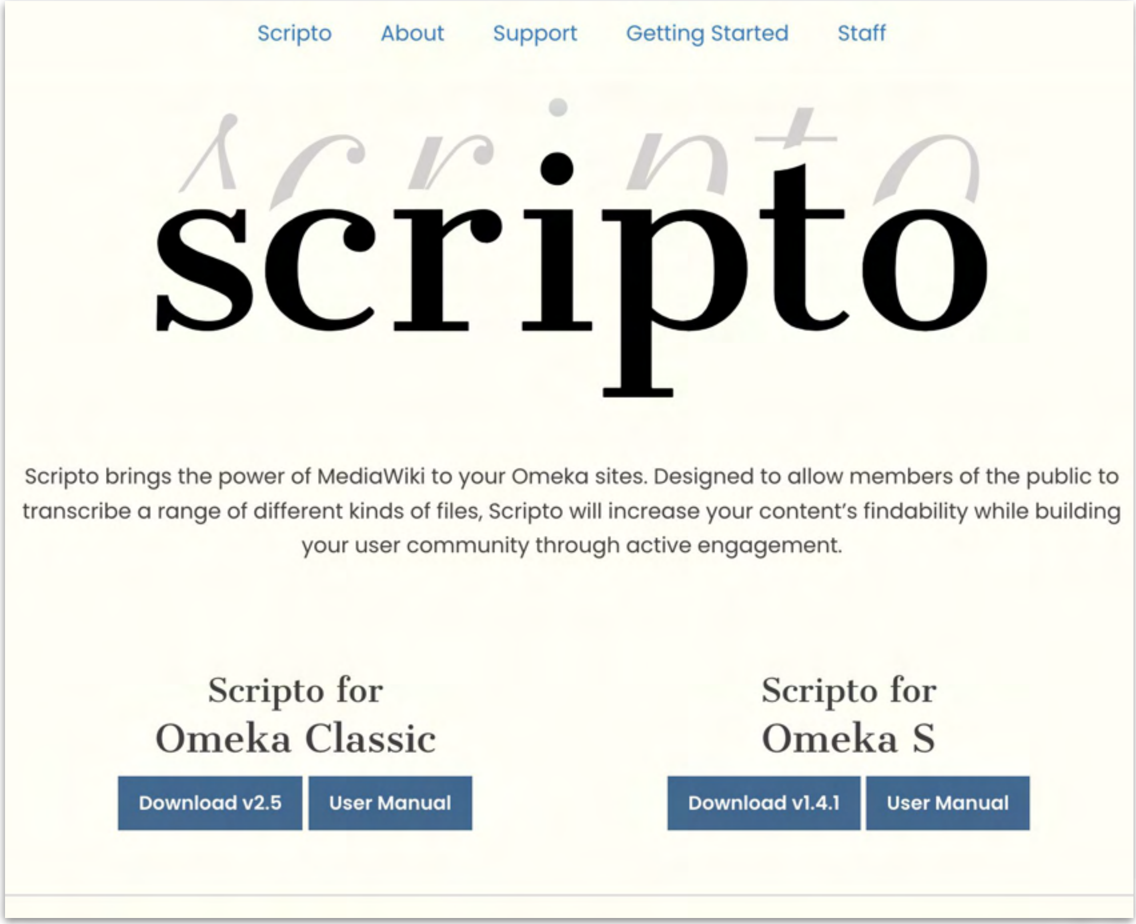
Transcribo is a *Rich Client Application* developed in the programming language *Java* by means of the Integrated Development Environment *Eclipse*.

We currently offer one version of Transcribo for the operating systems Windows and another for macOS. These can be found in the download area of the repository. The versions are packed as a zip archive and each contain its own Java Runtime Environment, so no Java environment have to be installed on the target device.

The zip archive must be unpacked into a folder for which the Transcribo user has write access. This is important because the application creates a workspace on disk when it starts.

After unpacking the archive, you will find the executable file for starting Transcribo in the *./eclipse/Transcribo.exe* directory for Windows systems and in the *./MacOS/Transcribo* directory for macOS.

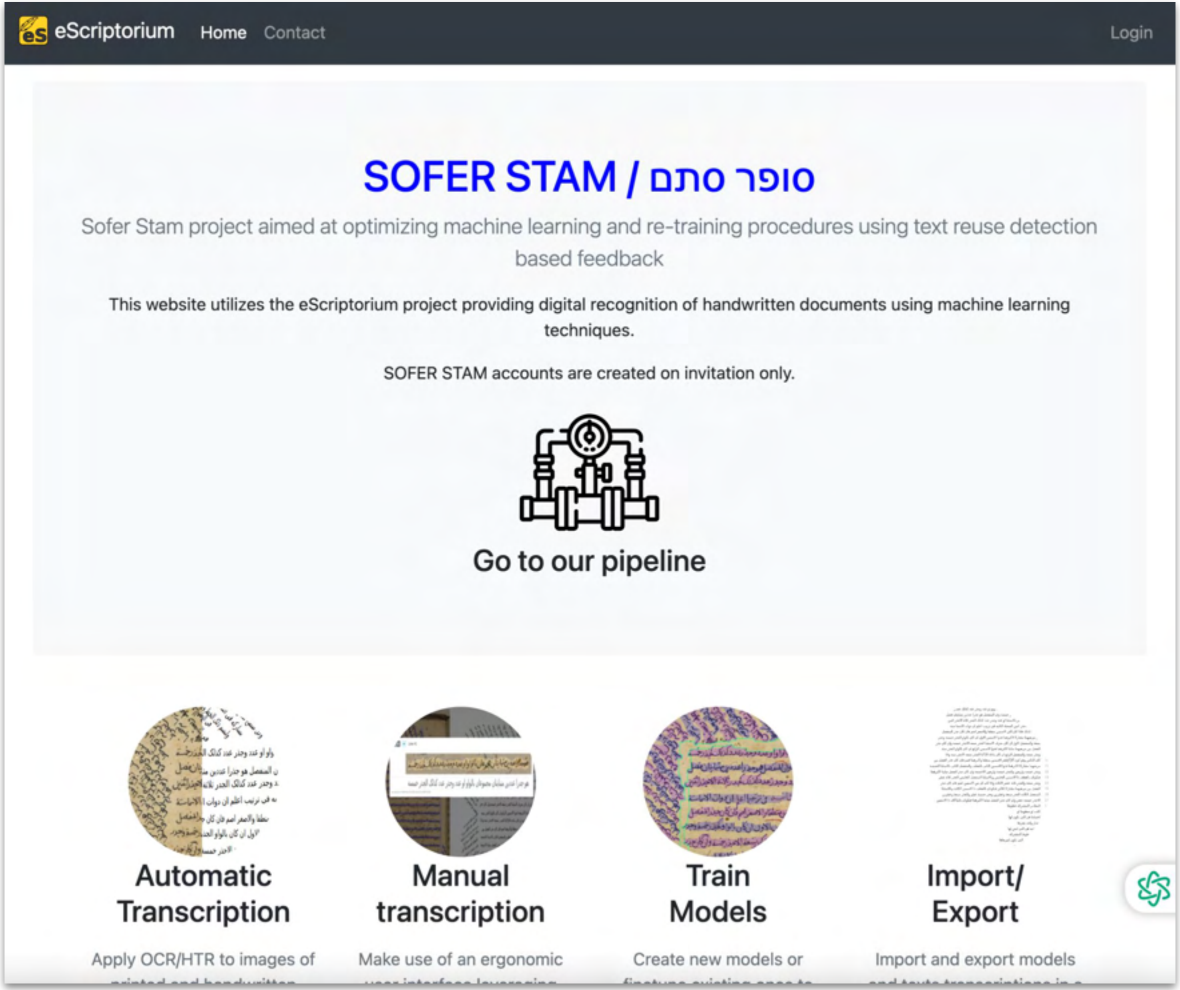
Note: There may be problems starting the application under macOS, because Transcribo is not verified for the macOS Gatekeeper. The following pages can help to fix this problem: https://support.apple.com/de-de/guide/mac-help/mh40616/mac or https://lucidgen.com/en/how-to-disable-gatekeeper-sip-on-mac/



Scripto brings the power of MediaWiki to your Omeka sites. Designed to allow members of the public to transcribe a range of different kinds of files, Scripto will increase your content's findability while building your user community through active engagement.

Scripto for Omeka Classic

Scripto for Omeka S



SOFER STAM / סופר סתם

Sofer Stam project aimed at optimizing machine learning and re-training procedures using text reuse detection based feedback

This website utilizes the eScriptorium project providing digital recognition of handwritten documents using machine learning techniques.

SOFER STAM accounts are created on invitation only.

Go to our pipeline

Automatic Transcription

Apply OCR/HTR to images of printed and handwritten

Manual transcription

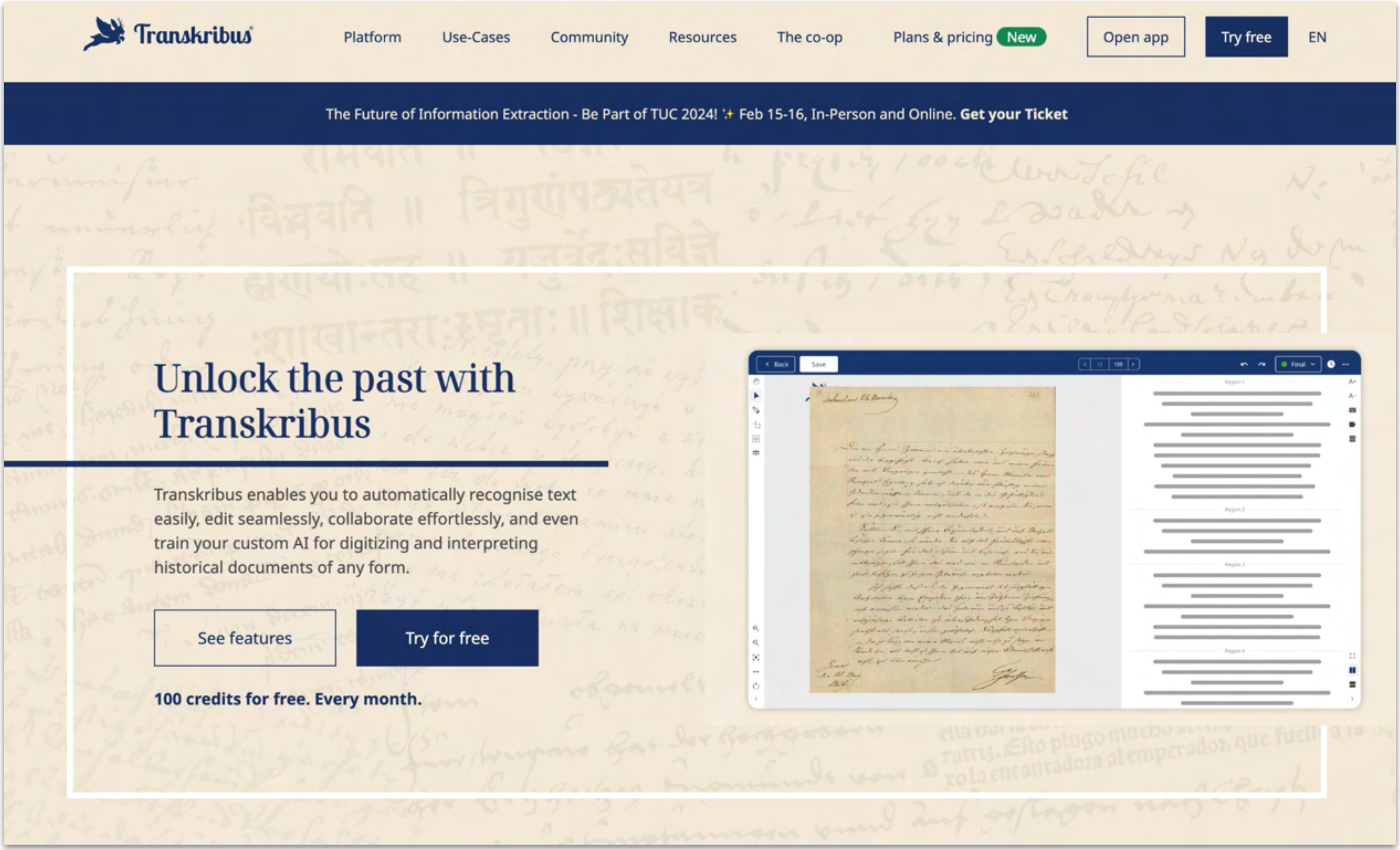
Make use of an ergonomic user interface leveraging

Train Models

Create new models or finetune existing ones to

Import/ Export

Import and export models and texts transcriptions in a



The Future of Information Extraction - Be Part of TUC 2024! ✨ Feb 15-16, In-Person and Online. **Get your Ticket**

Unlock the past with Transkribus

Transkribus enables you to automatically recognise text easily, edit seamlessly, collaborate effortlessly, and even train your custom AI for digitizing and interpreting historical documents of any form.

100 credits for free. Every month.

Characteristics, Differences, etc.

- download vs. web-based
- register
- pricing
- functions
- import formats
- export formats
- collaboration

List

- FromThePage
- Transcribo
- Transcribe Bentham: Transcription Desk
- Scripto
- eScriptorium
- https://github.com/anguelos/frat
- T Pen
- Tropy

Possible Exercise

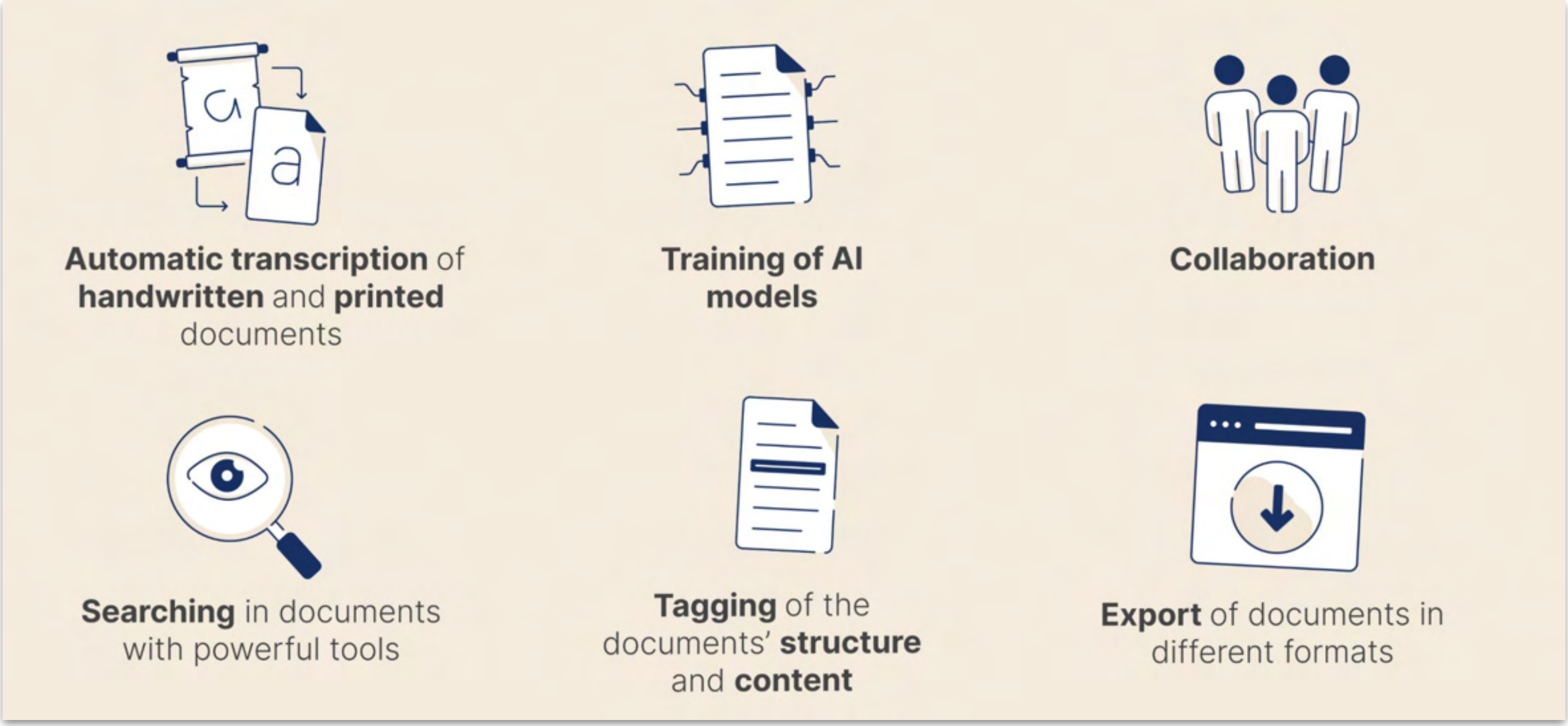
Have a look at one tool or more.

Transkribus: Introduction

Transkribus: General

- platform primarily for AI-supported layout and text recognition of printed or handwritten documents
- but also enables - with limits - the annotation of structure and content
- The platform goes back to the READ project launched in 2016 and is provided and continuously developed by the European cooperative READ-COOP SCE, which was founded in 2019
- The cooperative now has more than 130 members (institutions and private individuals) in 30 countries (as of April 2023)

Transkribus: Functions



Automatic transcription of **handwritten** and **printed** documents

Training of AI models

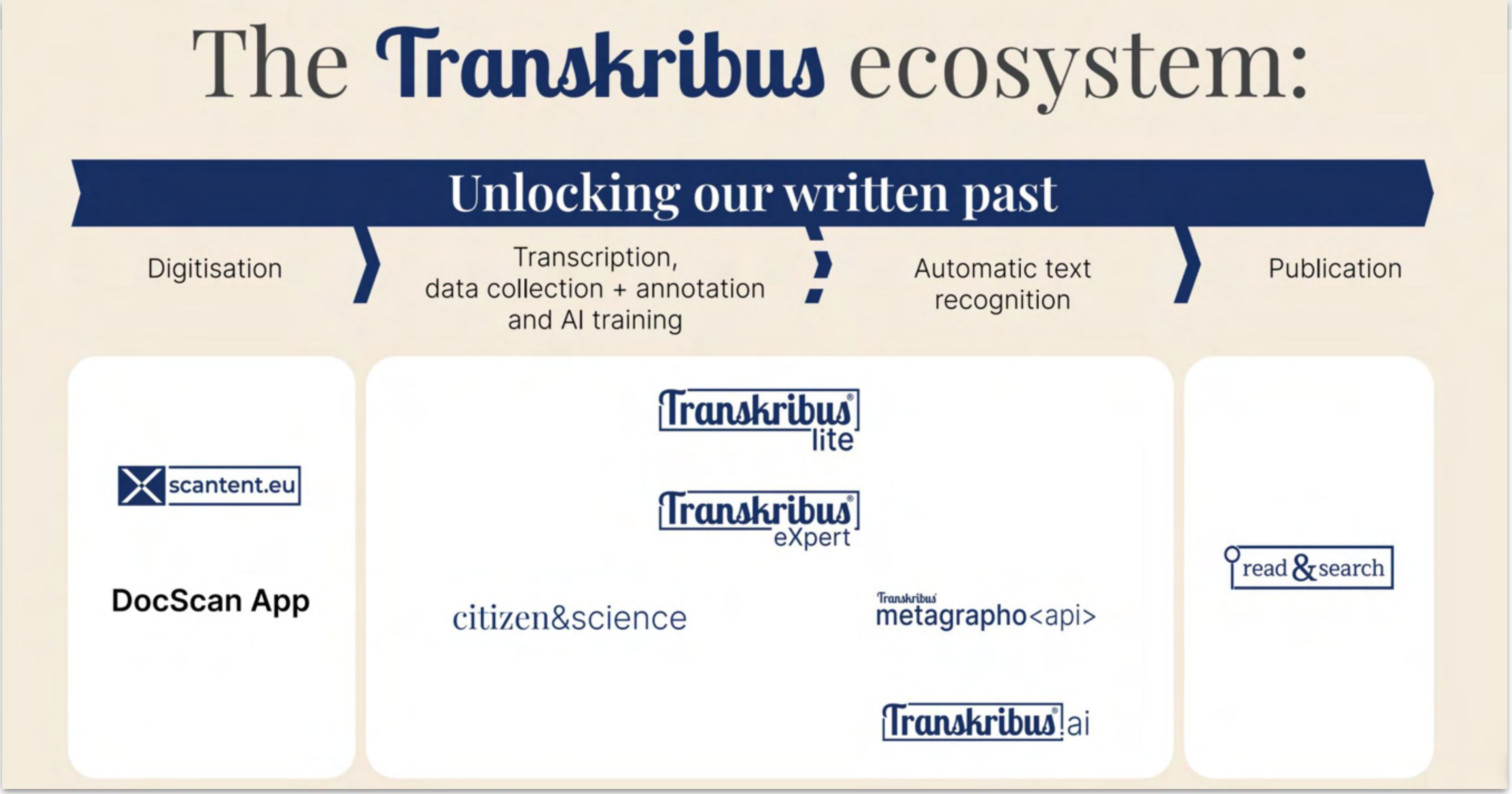
Collaboration

Searching in documents with powerful tools

Tagging of the documents' **structure** and **content**

Export of documents in different formats

Source: Slides from Transkribus Webinar 2022



The Transkribus ecosystem:

Unlocking our written past

Digitisation

Transcription, data collection + annotation and AI training

Automatic text recognition

Publication

scantent.eu

DocScan App

Transkribus lite

Transkribus eXpert

citizen&science

Transkribus metagrapho<api>

Transkribus ai

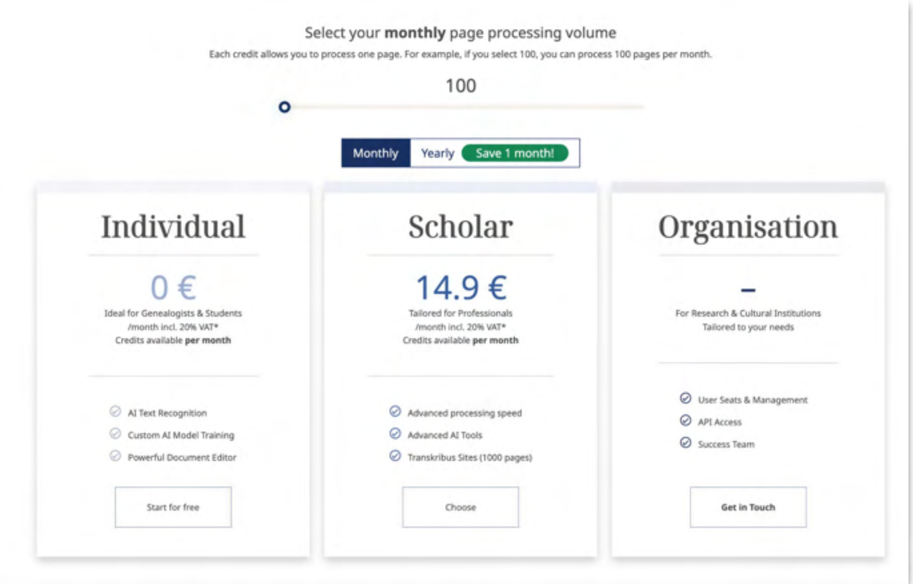
read & search

Source: Slides from Transkribus Webinar 2022

Transkribus: Plan and Prices

https://www.transkribus.org/plans

- 100 credits free per month
- 1 page processing = 1 credit
- only with scholar and organisation plans:
  - collaboration tools
  - export in TEI
  - Transkribus Sites
  - Use of super models (text recognition)
  - Table recognition
  - …



Select your **monthly** page processing volume

Each credit allows you to process one page. For example, if you select 100, you can process 100 pages per month.

100

Individual

0 €

Ideal for Genealogists & Students
/month incl. 20% VAT*
Credits available **per month**

- AI Text Recognition
- Custom AI Model Training
- Powerful Document Editor

Scholar

14.9 €

Tailored for Professionals
/month incl. 20% VAT*
Credits available **per month**

- Advanced processing speed
- Advanced AI Tools
- Transkribus Sites (1000 pages)

Organisation

–

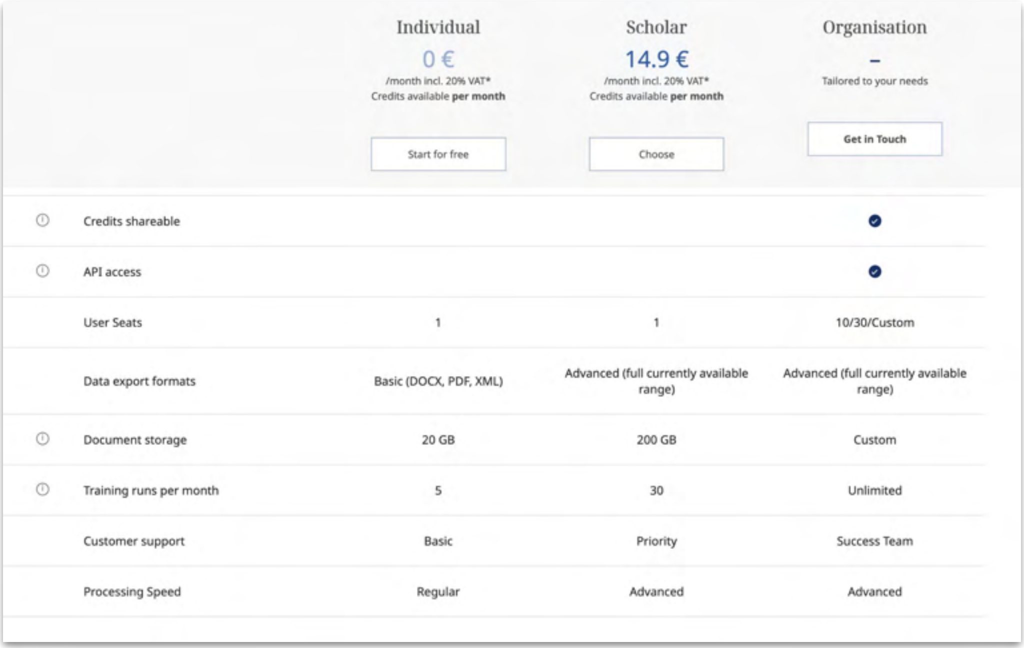
For Research & Cultural Institutions
Tailored to your needs

- User Seats & Management
- API Access
- Success Team

Transkribus: Plan and Prices

https://www.transkribus.org/plans

- also differences in document storage,how many trainings you can run per month, processing speed, etc.



	Individual	Scholar	Organisation
	0 € /month incl. 20% VAT* Credits available **per month**	14.9 € /month incl. 20% VAT* Credits available **per month**	– Tailored to your needs
Credits shareable			✓
API access			✓
User Seats	1	1	10/30/Custom
Data export formats	Basic (DOCX, PDF, XML)	Advanced (full currently available range)	Advanced (full currently available range)
Document storage	20 GB	200 GB	Custom
Training runs per month	5	30	Unlimited
Customer support	Basic	Priority	Success Team
Processing Speed	Regular	Advanced	Advanced

Transkribus eXpert vs. Transkribus Web-App

- Transkribus eXpert (Standalone-Version, Java-basiert)
- Transkribus Lite (Webversion)

- Due to the great acceptance of the web version, only this version will be developed further
- Transkribus eXpert will continue to be available, but no new features will be added
- All documents that are uploaded to Transkribus Lite are also available in Transkribus eXpert because they are stored on the READ COOP SCE servers.
- Transkribus Lite is constantly being expanded with new functionalities

Transkribus: WebApp

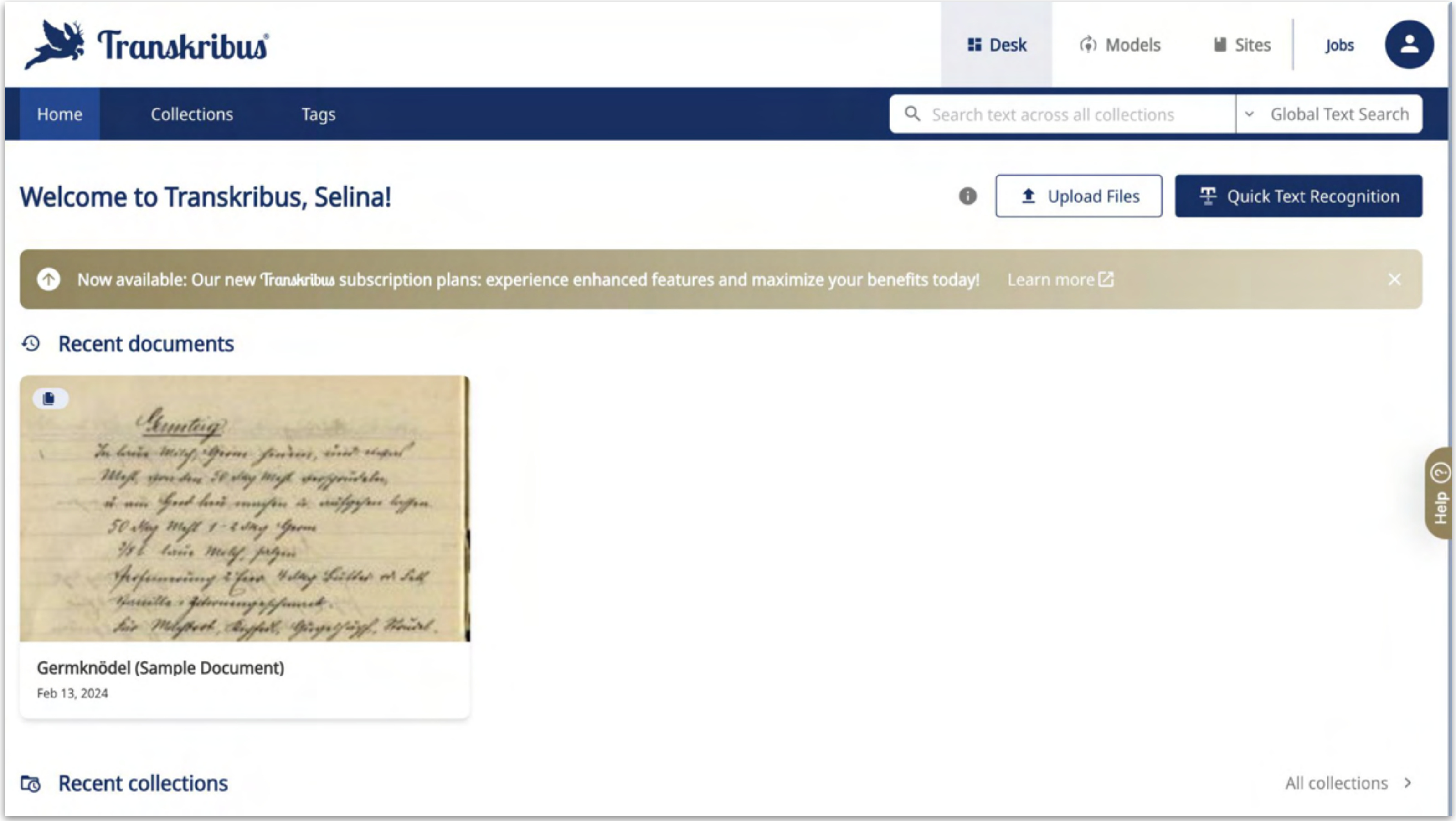
(former: Transkribus Lite)

Transkribus: WebApp

- Log in on the Transkribus Website
  - https://readcoop.eu/de/transkribus/
  - https://www.transkribus.org/
- **Start the App (right corner)**

Exploring the “Workdesk”

Workdesk: Home



Welcome to Transkribus, Selina!

Now available: Our new Transkribus subscription plans: experience enhanced features and maximize your benefits today! Learn more

Recent documents

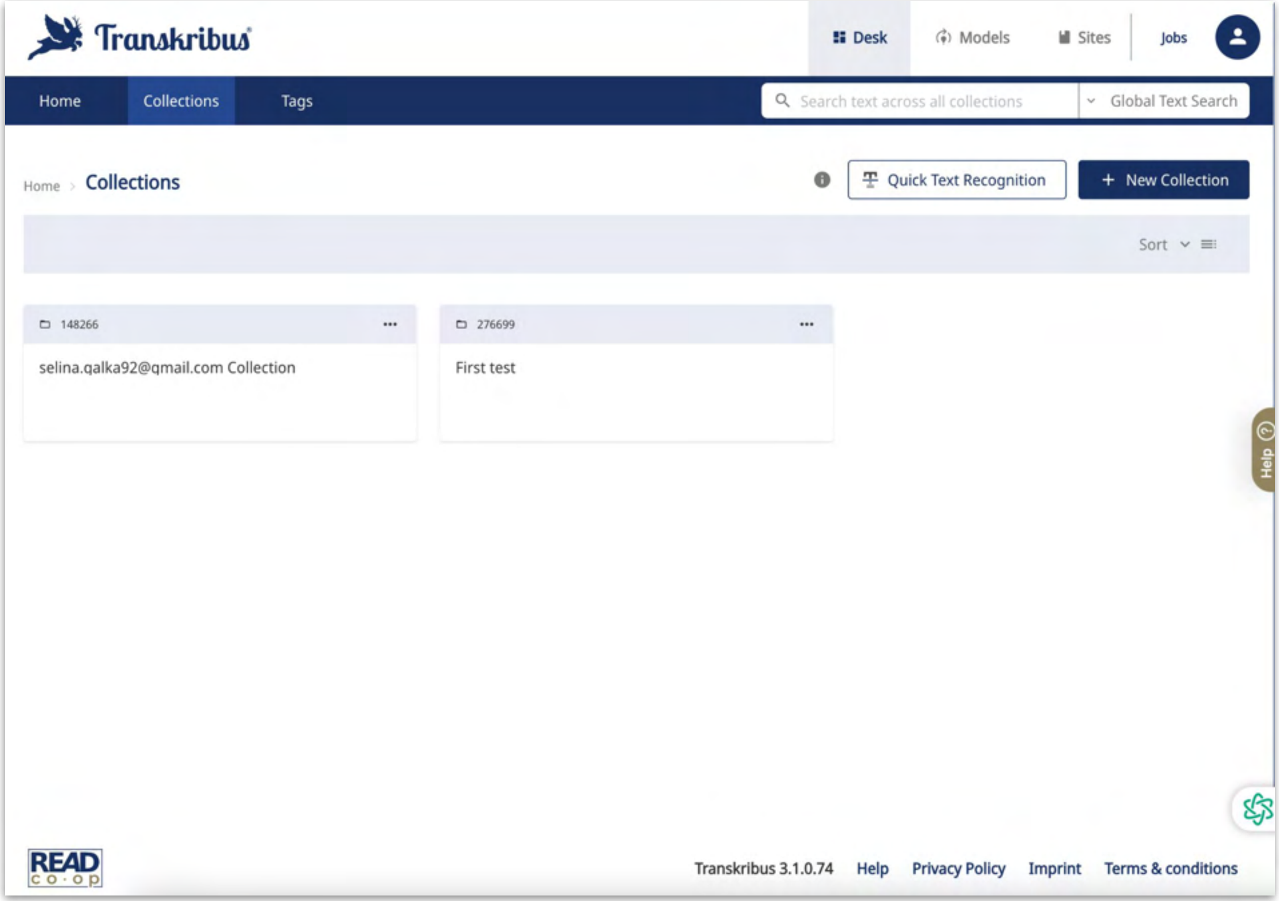
Germknödel (Sample Document)

Feb 13, 2024

Recent collections

All collections

Workdesk: Collection View



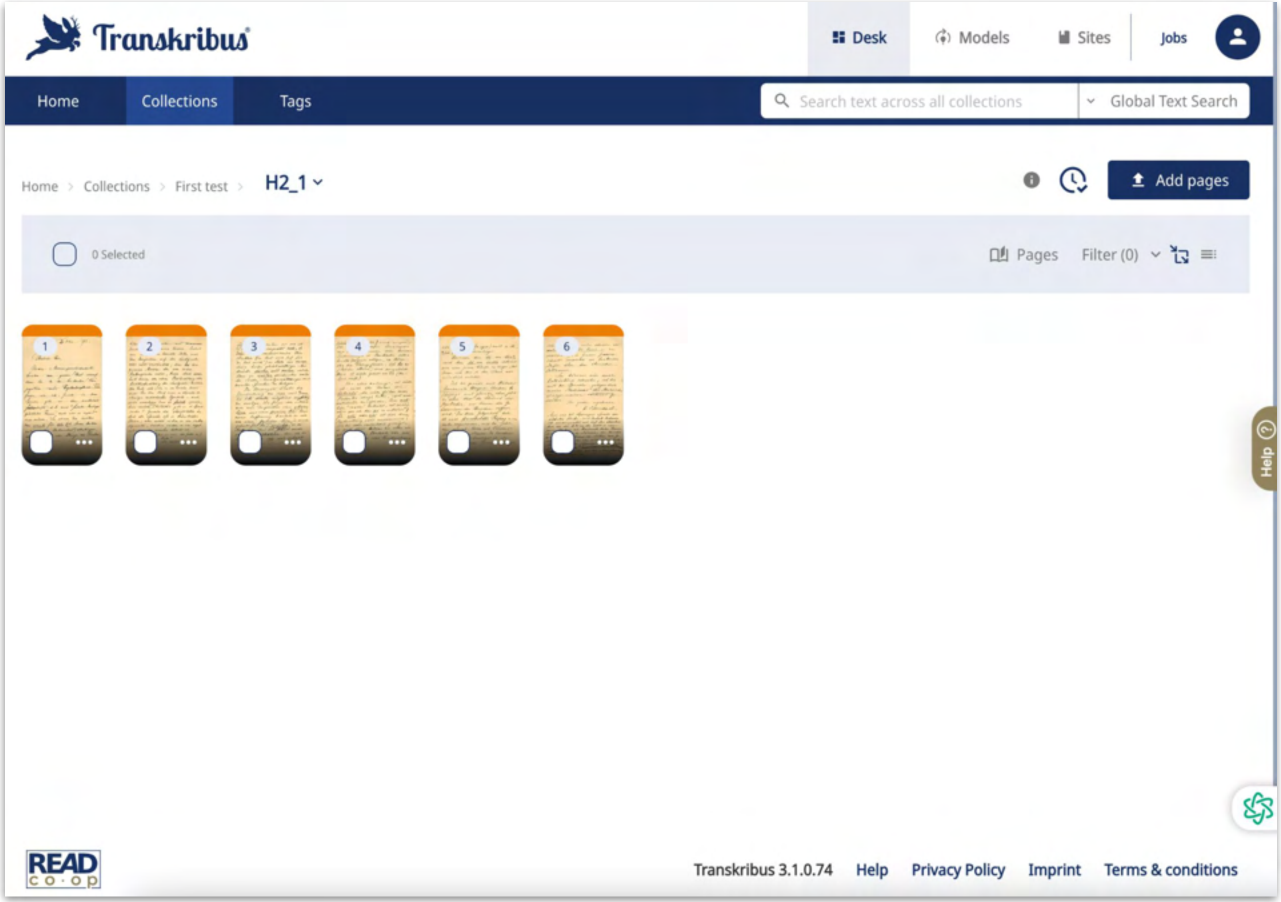
148266

selina.qalka92@qmail.com Collection

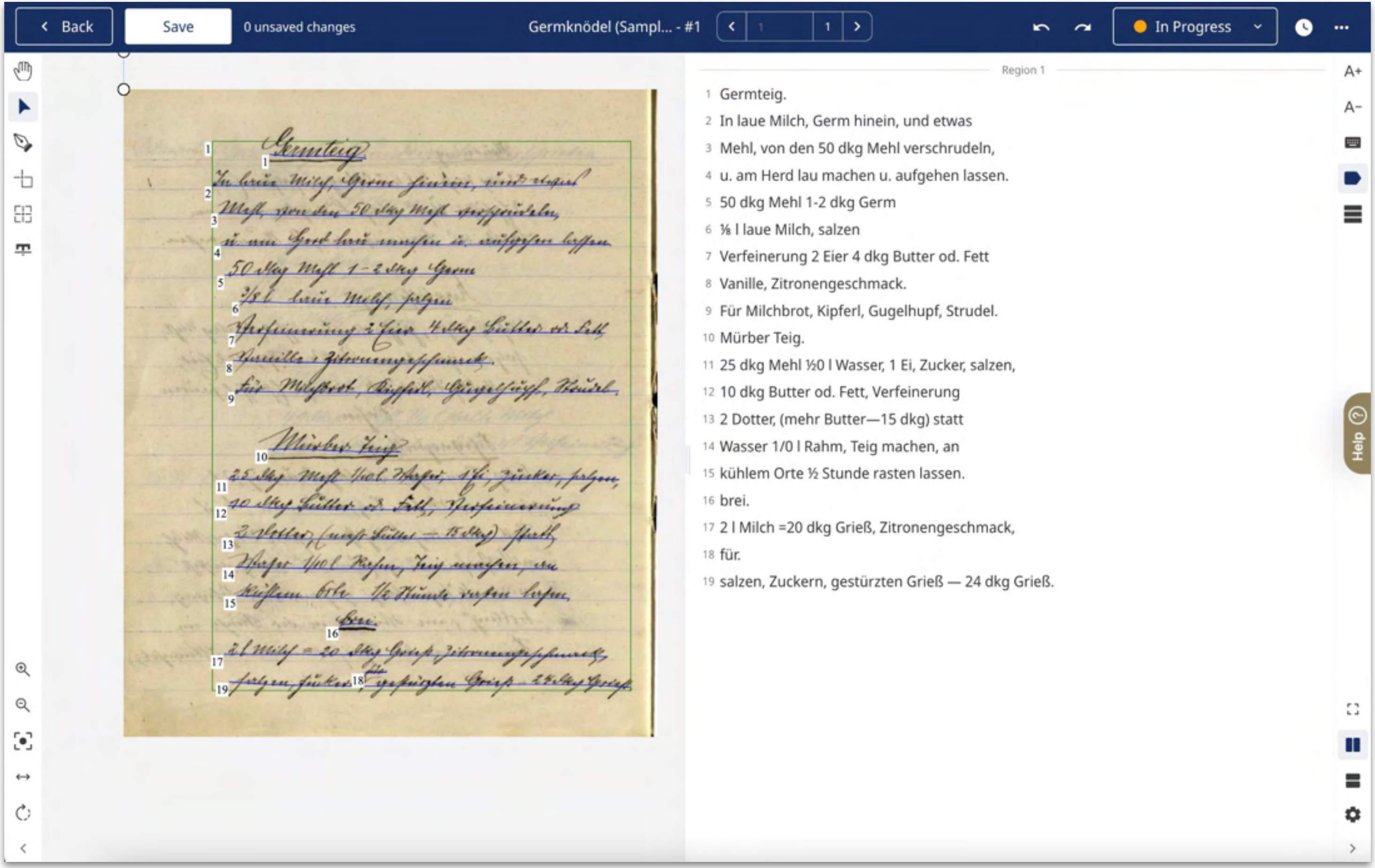
276699

First test

Workdesk: Document View

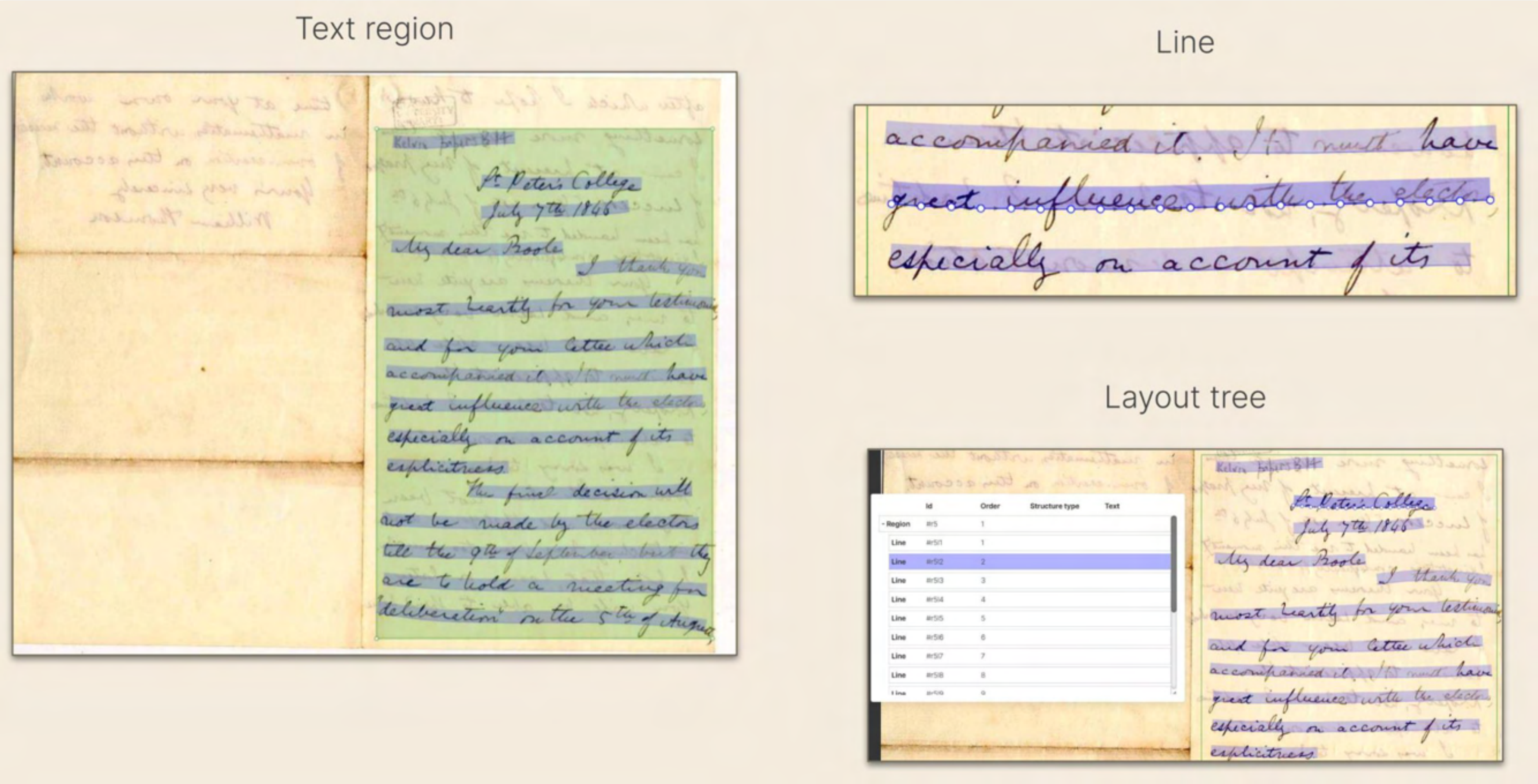


Workdesk: Page View



Germteig.
In laue Milch, Germ hinein, und etwas
Mehl, von den 50 dkg Mehl verschrudeln,
u. am Herd lau machen u. aufgehen lassen.
50 dkg Mehl 1-2 dkg Germ
⅛ l laue Milch, salzen
Verfeinerung 2 Eier 4 dkg Butter od. Fett
Vanille, Zitronengeschmack.
Für Milchbrot, Kipferl, Gugelhupf, Strudel.
Mürber Teig.
25 dkg Mehl ½0 l Wasser, 1 Ei, Zucker, salzen,
10 dkg Butter od. Fett, Verfeinerung
2 Dotter, (mehr Butter—15 dkg) statt
Wasser 1/0 l Rahm, Teig machen, an
kühlem Orte ½ Stunde rasten lassen.
brei.
2 l Milch =20 dkg Grieß, Zitronengeschmack,
für.
salzen, Zuckern, gestürzten Grieß — 24 dkg Grieß.

Layout Tree



Kelvin Papers B14

St Peter's College
July 7th 1846

My dear Boole
I thank you
most heartily for your testimon
and for your letter which
accompanied it. It must have
great influence with the electors
especially on account of its
explicitness.
The final decision will
not be made by the electors
till the 9th of September, but they
are to hold a meeting for
deliberation on the 5th of August

	Id	Order	Structure type	Text
- Region	#r5	1		
Line	#r5l1	1		
Line	#r5l2	2		
Line	#r5l3	3		
Line	#r5l4	4		
Line	#r5l5	5		
Line	#r5l6	6		
Line	#r5l7	7		
Line	#r5l8	8		
Line	#r5l9	9		

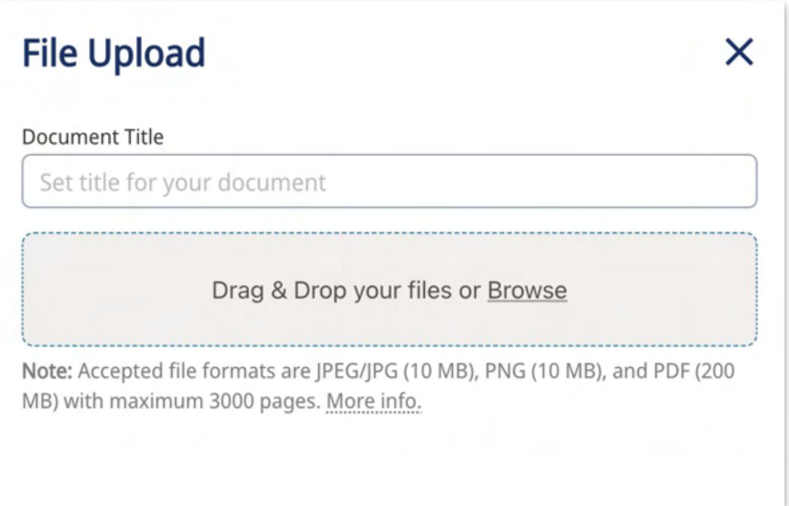
Source: Slides from Transkribus Webinar 2022

Upload

- Create a collection
- click into the collection
- “Upload files”

Accepted formats:

- JPEG
- PNG
- PDF



File Upload

Document Title

Set title for your document

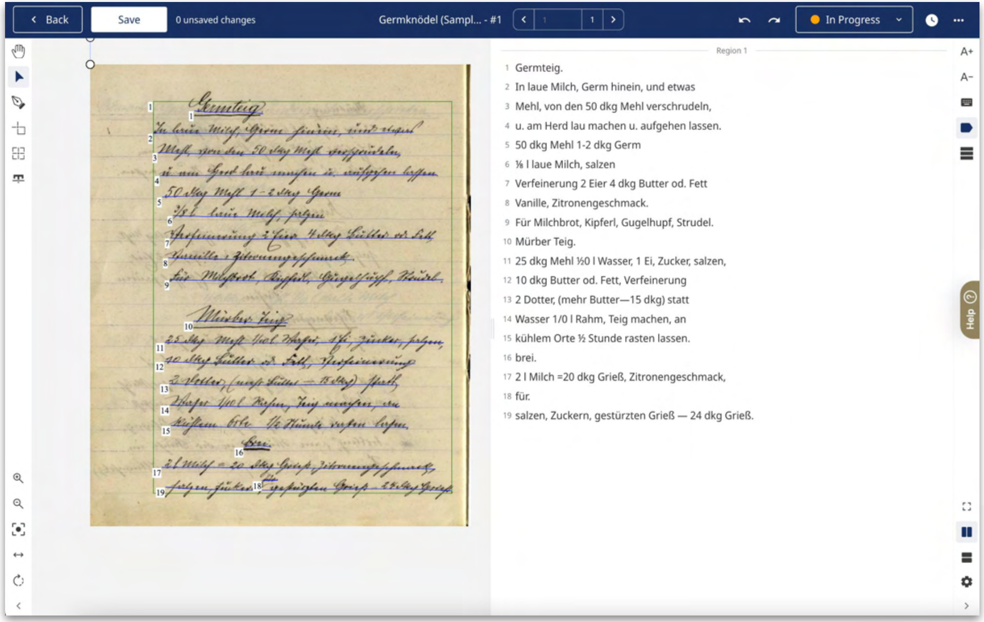
Drag & Drop your files or Browse

Note: Accepted file formats are JPEG/JPG (10 MB), PNG (10 MB), and PDF (200 MB) with maximum 3000 pages. More info.

All files uploaded together are regarded as a single document, each individual image or page of a PDF becomes a page of the document.

Hands on

- Explore the Workdesk, especially the page view! (image + transcription)



Germteig.
In laue Milch, Germ hinein, und etwas
Mehl, von den 50 dkg Mehl verschrudeln,
u. am Herd lau machen u. aufgehen lassen.
50 dkg Mehl 1-2 dkg Germ
⅛ l laue Milch, salzen
Verfeinerung 2 Eier 4 dkg Butter od. Fett
Vanille, Zitronengeschmack.
Für Milchbrot, Kipferl, Gugelhupf, Strudel.
Mürber Teig.
25 dkg Mehl ½0 l Wasser, 1 Ei, Zucker, salzen,
10 dkg Butter od. Fett, Verfeinerung
2 Dotter, (mehr Butter—15 dkg) statt
Wasser 1/0 l Rahm, Teig machen, an
kühlem Orte ½ Stunde rasten lassen.
brei.
2 l Milch =20 dkg Grieß, Zitronengeschmack,
für.
salzen, Zuckern, gestürzten Grieß — 24 dkg Grieß.

Hands on

- Upload the letters provided in the Google Drive as documents into a collection (a document per letter)
- Letters:
    - Hugo Schuchardt Archiv
    - german linguist (1842-1927)
- Information about letter 1: https://gams.uni-graz.at/o:hsa.letter.3828/sdef:TEI/get
- Information about letter 2: https://gams.uni-graz.at/o:hsa.letter.3830/sdef:TEI/get

The hands on material is based on the tutorial for Transkribus by DigEdTnT: https://digedtnt.github.io/transkribus/.

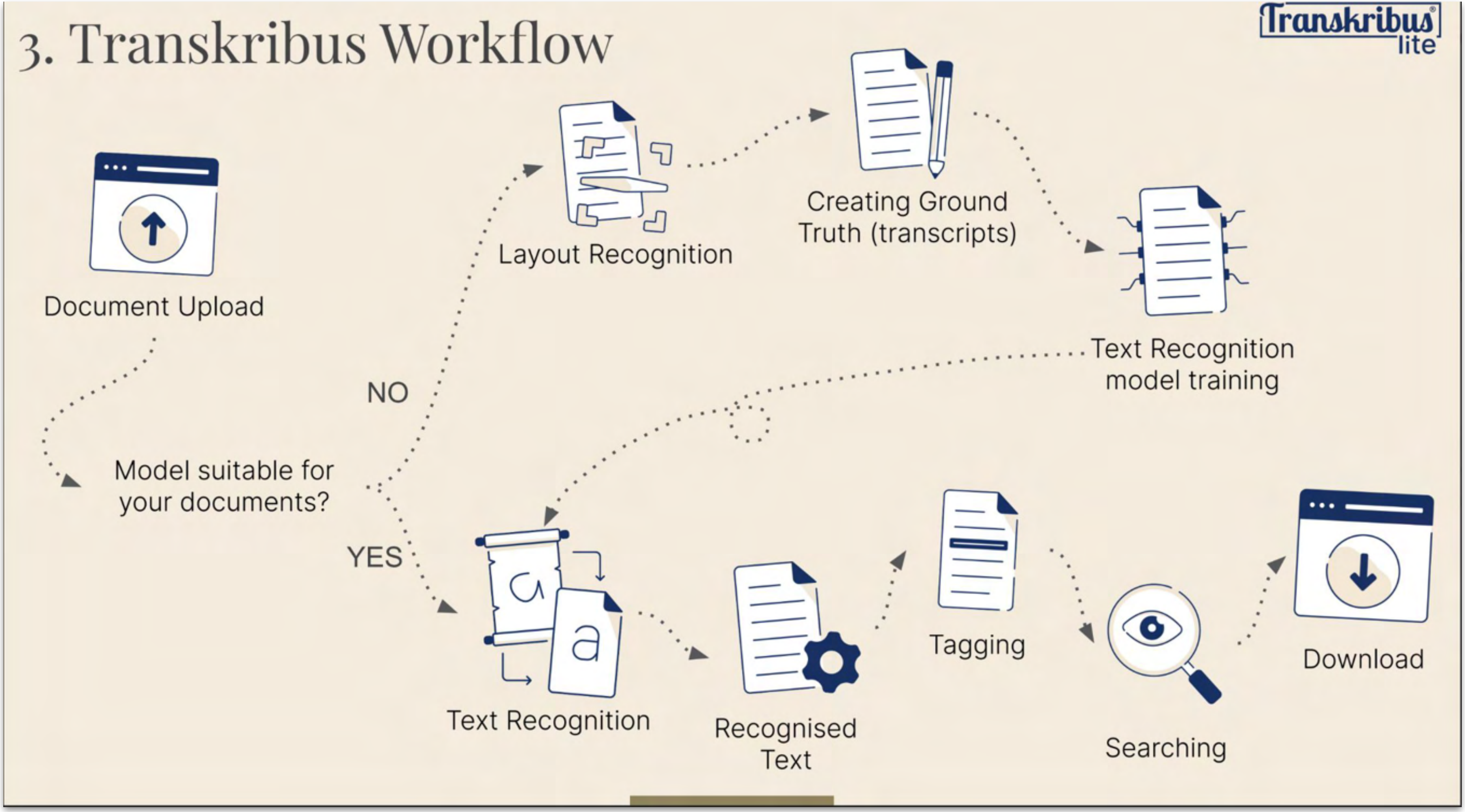
Applying a HTR model

HTR / Automatic Text Recognition

HTR is able to recognise handwritten documents and historical prints (books and newspapers) BUT there is no general model for all scripts/languages and epochs.

2 options:

1. selecting a public model that has already been trained on similar scripts by the Transkribus community
2. training a custom model for recognising a specific handwriting/font by showing it a certain amount of images and their transcriptions



3. Transkribus Workflow

Document Upload

Model suitable for your documents?

NO

Layout Recognition

Creating Ground Truth (transcripts)

Text Recognition model training

YES

Text Recognition

Recognised Text

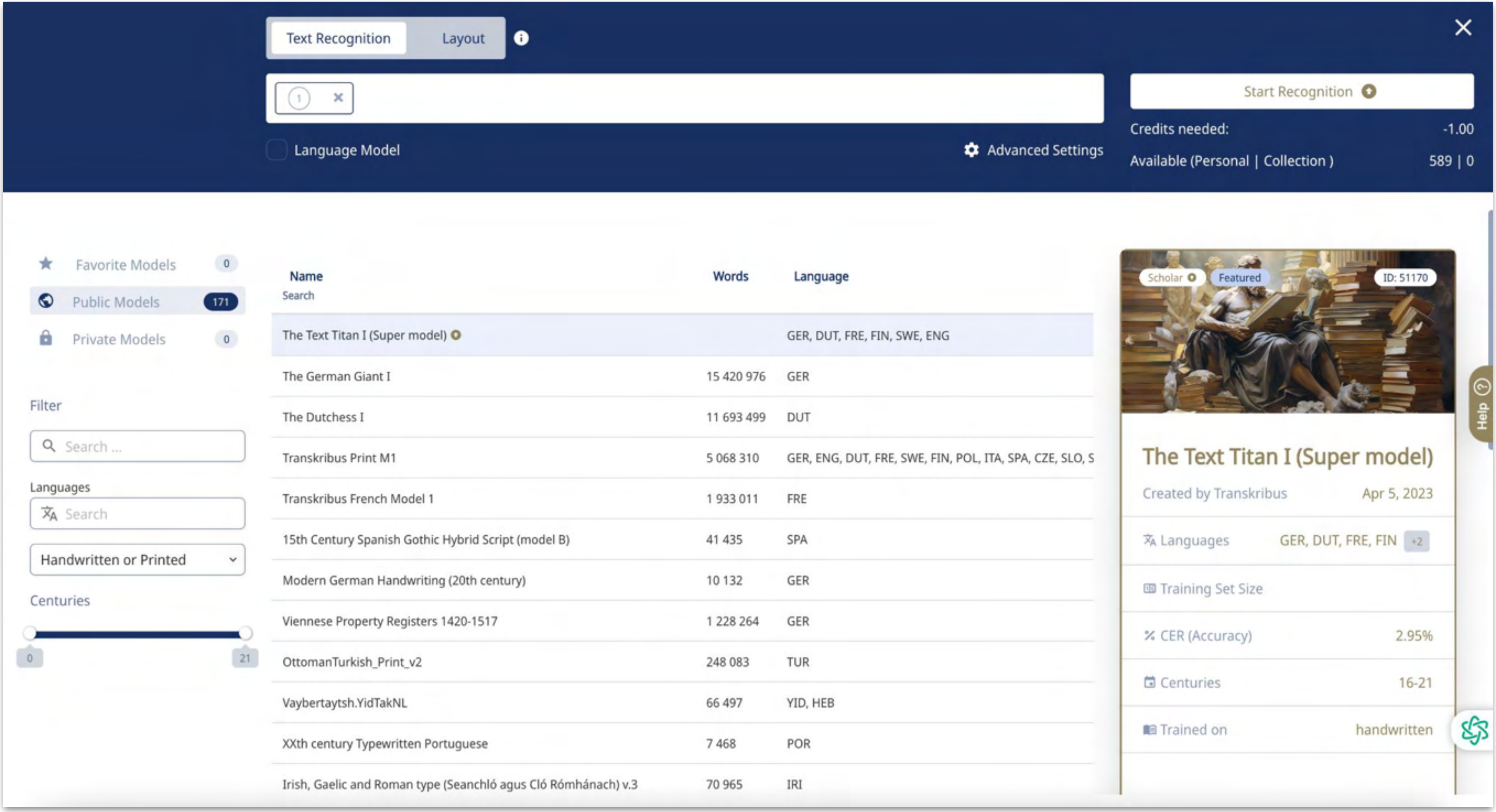
Tagging

Searching

Download

Source: Slides from Transkribus Webinar 2022

Text Recognition



□ Language Model

Credits needed: -1.00

Available (Personal | Collection) 589 | 0

Name Search	Words	Language
The Text Titan I (Super model)		GER, DUT, FRE, FIN, SWE, ENG
The German Giant I	15 420 976	GER
The Dutchess I	11 693 499	DUT
Transkribus Print M1	5 068 310	GER, ENG, DUT, FRE, SWE, FIN, POL, ITA, SPA, CZE, SLO, S
Transkribus French Model 1	1 933 011	FRE
15th Century Spanish Gothic Hybrid Script (model B)	41 435	SPA
Modern German Handwriting (20th century)	10 132	GER
Viennese Property Registers 1420-1517	1 228 264	GER
OttomanTurkish_Print_v2	248 083	TUR
Vaybertaytsh.YidTakNL	66 497	YID, HEB
XXth century Typewritten Portuguese	7 468	POR
Irish, Gaelic and Roman type (Seanchló agus Cló Rómhánach) v.3	70 965	IRI

Scholar Featured ID: 51170

The Text Titan I (Super model)

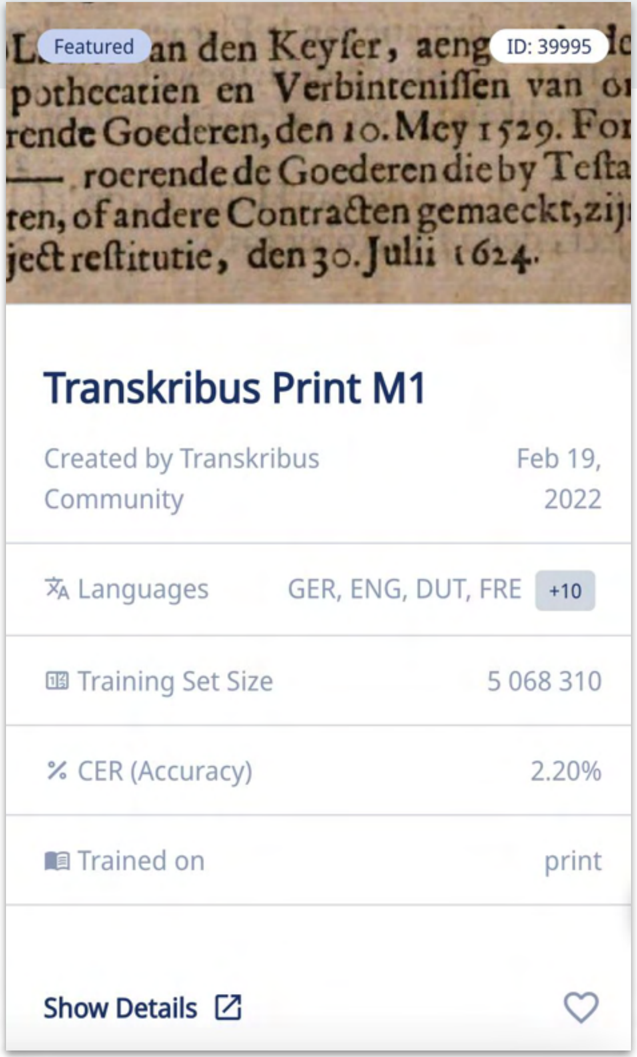
Created by Transkribus Apr 5, 2023

Languages	GER, DUT, FRE, FIN +2
Training Set Size	
CER (Accuracy)	2.95%
Centuries	16-21
Trained on	handwritten

Apply a model in Transkribus

How to choose the best public model for your documents:

- Material
- Language
- Character Error Rate (CER) = the percentage of incorrect characters out of 100 characters automatically transcribed by the AI (desired CER: below 10%)



Featured ID: 39995

L an den Keyſer, aeng
pothecatien en Verbinteniſſen van o
rende Goederen, den 10. Mey 1529. For
roerende de Goederen die by Teſta
ten, of andere Contracten gemaeckt, zij
jeƈt reſtitutie, den 30. Julii 1624.

Transkribus Print M1

Created by Transkribus Community — Feb 19, 2022

Languages	GER, ENG, DUT, FRE +10
Training Set Size	5 068 310
CER (Accuracy)	2.20%
Trained on	print

Show Details

Applying a model in Transkribus

- **TIPP:** Try it out on only one or a few pages first, to see, how it performs!

- **HANDS ON**: Apply a pretrained model on source material
  - provided material on google drive
    - Letter 1
    - search for a model in Transkribus
      - information on the material: handwritten letters from Hugo Schuchardt, german, 1860-1925, Information about the project: https://gams.uni-graz.at/context:hsa

Applying a model in Transkribus

- Check the results
- How good did the layout recognition work?
  - Do you need to correct lines?
- Have a look at the automatic created transcription
  - Comparison with the transcription from the archive (https://gams.uni-graz.at/o:hsa.letter.3828/sdef:TEI/get)
  - How good did it work?

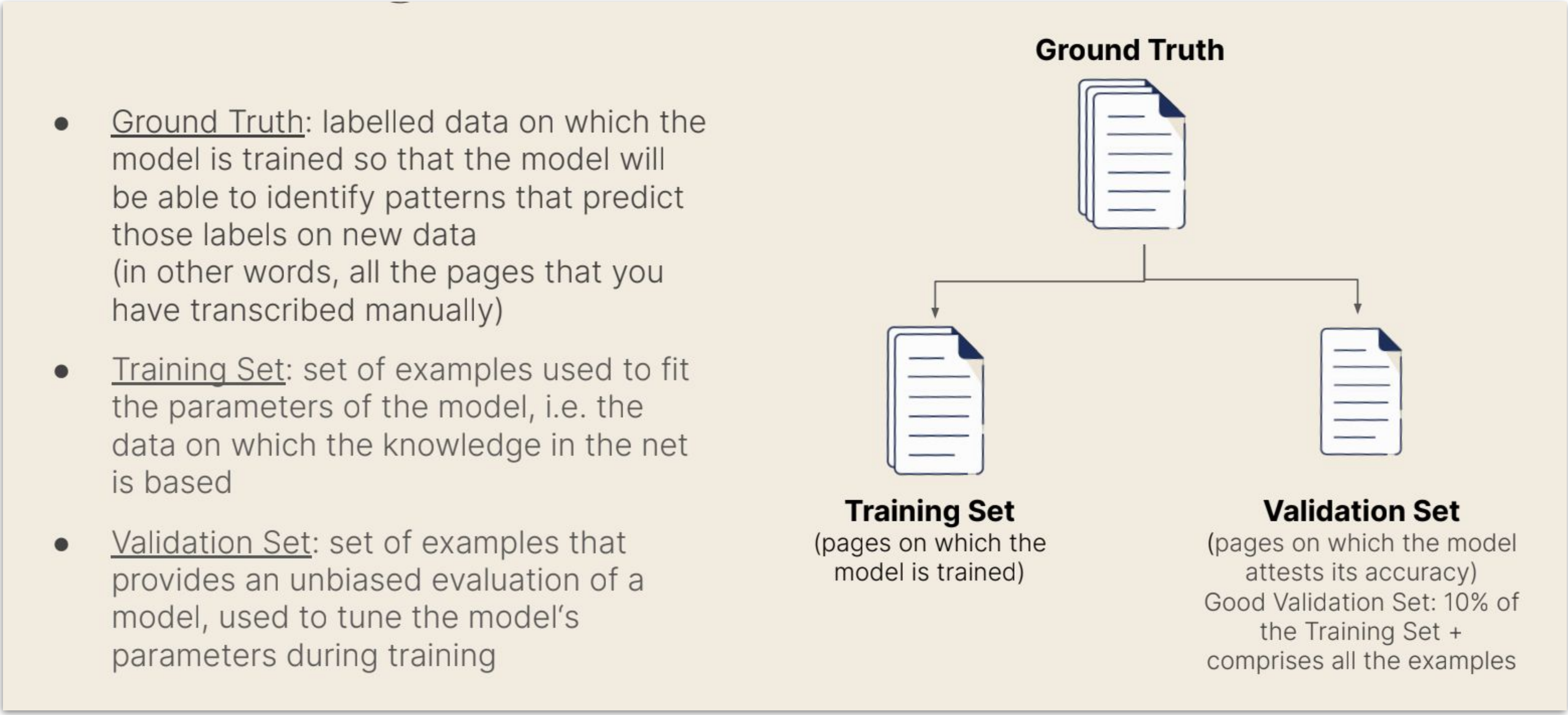
Training a HTR model

Training a HTR model

- you need data for training a model (“**Training Data**”)
- 25-75 pages (5.000-15.000 words) as training data (“**Ground Truth**”)
  - **for handwritten documents: at least 10.000 words per hand**
  - the Ground Truth should include examples of all the scripts that you want your model to be able to transcribe.
  - training data in Transkribus: Faksimiles + Transcriptions (correct and as accurate as possible!)
  - the pages to include in the Ground Truth are, therefore, important because they will affect the effectiveness of the model

Training a HTR model

- 2 options for creating the data:
    1. transcribing the page manually (Attention: at the moment it's not possible, to copy text into the transcription window!)
    2. applying a model that has been trained on similar handwritings (if available) and correcting the transcription manually



- Ground Truth: labelled data on which the model is trained so that the model will be able to identify patterns that predict those labels on new data (in other words, all the pages that you have transcribed manually)
- Training Set: set of examples used to fit the parameters of the model, i.e. the data on which the knowledge in the net is based
- Validation Set: set of examples that provides an unbiased evaluation of a model, used to tune the model's parameters during training

Ground Truth

Training Set
(pages on which the model is trained)

Validation Set
(pages on which the model attests its accuracy)
Good Validation Set: 10% of the Training Set + comprises all the examples

Source: Slides from Transkribus Webinar 2022

Workflow 1: Training a model with manually transcribed text

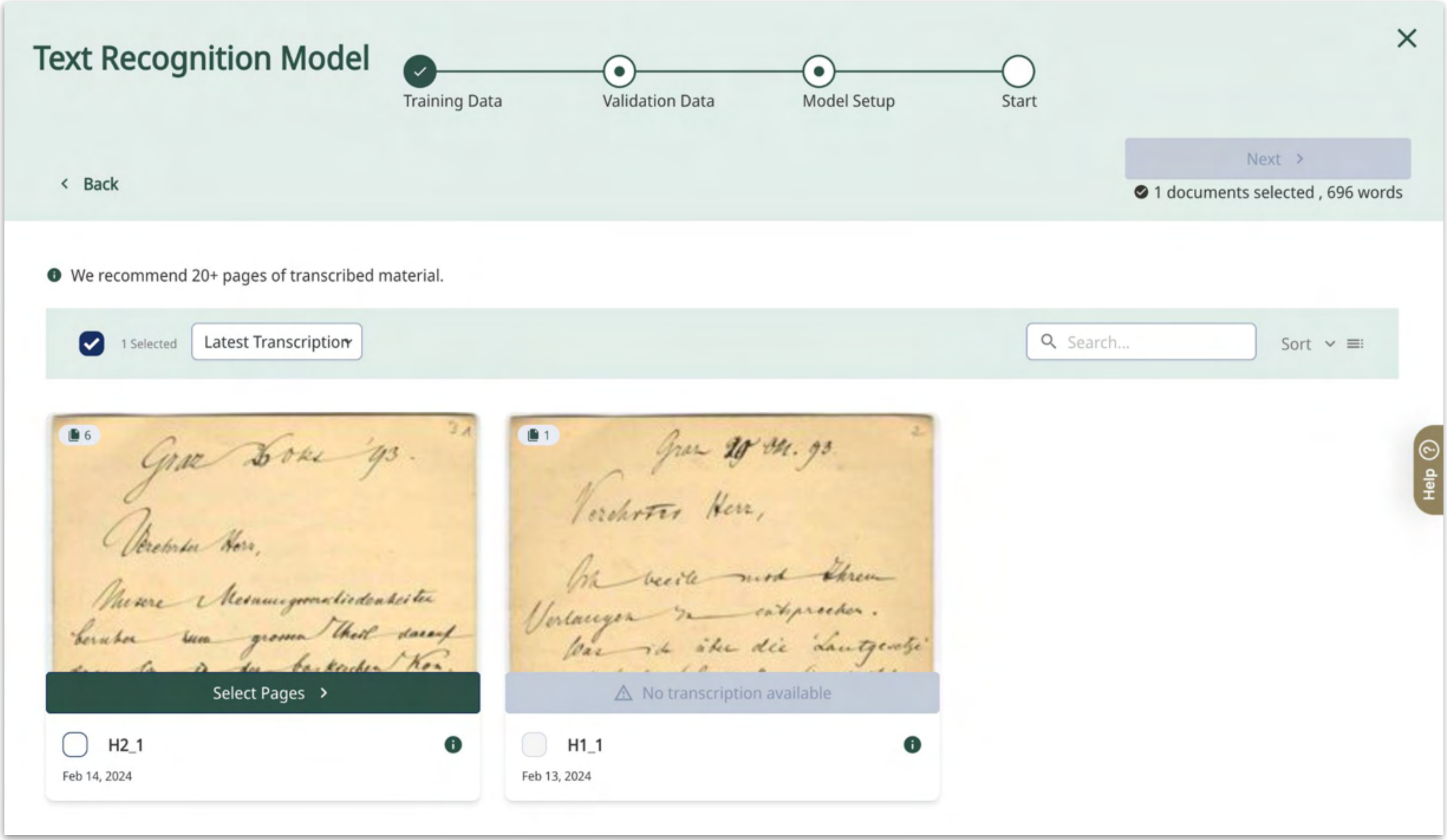
- Choose the pages to include in the ground truth
- Run the Layout Recognition
- Transcribe:
  - Transcribe what you read (including errors and punctuation)
  - Be consistent! (suggestion: write a document with your decisions)
  - Tag the words you can’t read as “uncertain” or “gap”
  - Lines left blank: aren’t considered in the training
  - Abbreviations: maintained/solved/tagged: it depends on what you expect as final output
- Save the page as “Ground Truth”

Workflow 1: Apply a model and correct the automatic transcriptions

- Choose the pages to include in the ground truth
- Run the Text Recognition
- Correct the automatic transcriptions
- Save the page as “Ground Truth”

How to train a model in Transkribus

Step 1: Select your training data



Text Recognition Model

1 documents selected , 696 words

We recommend 20+ pages of transcribed material.

☑ 1 Selected

Latest Transcription

6

□ H2_1

Feb 14, 2024

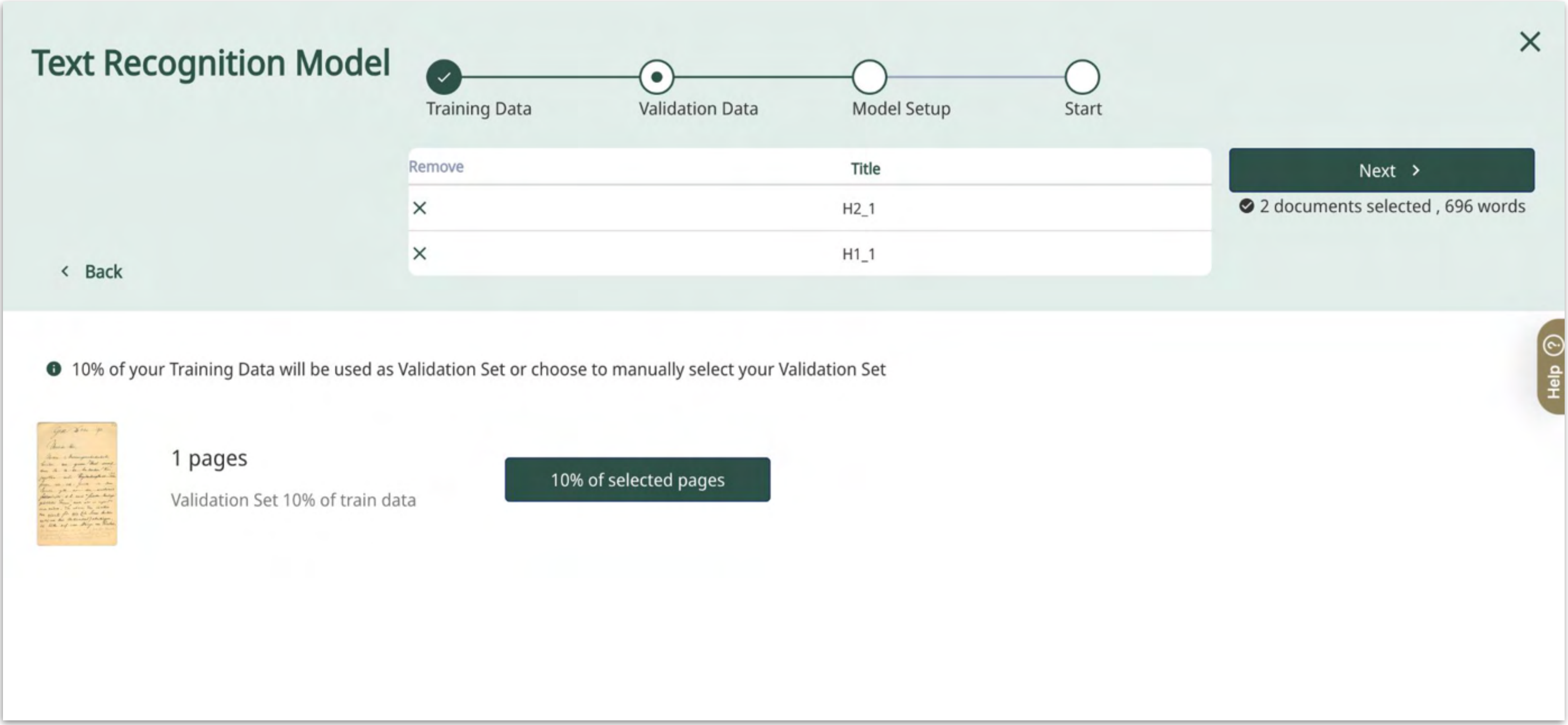
1

No transcription available

□ H1_1

Feb 13, 2024

Step 2: Select your validation data



Text Recognition Model

Remove	Title
✕	H2_1
✕	H1_1

✔ 2 documents selected , 696 words

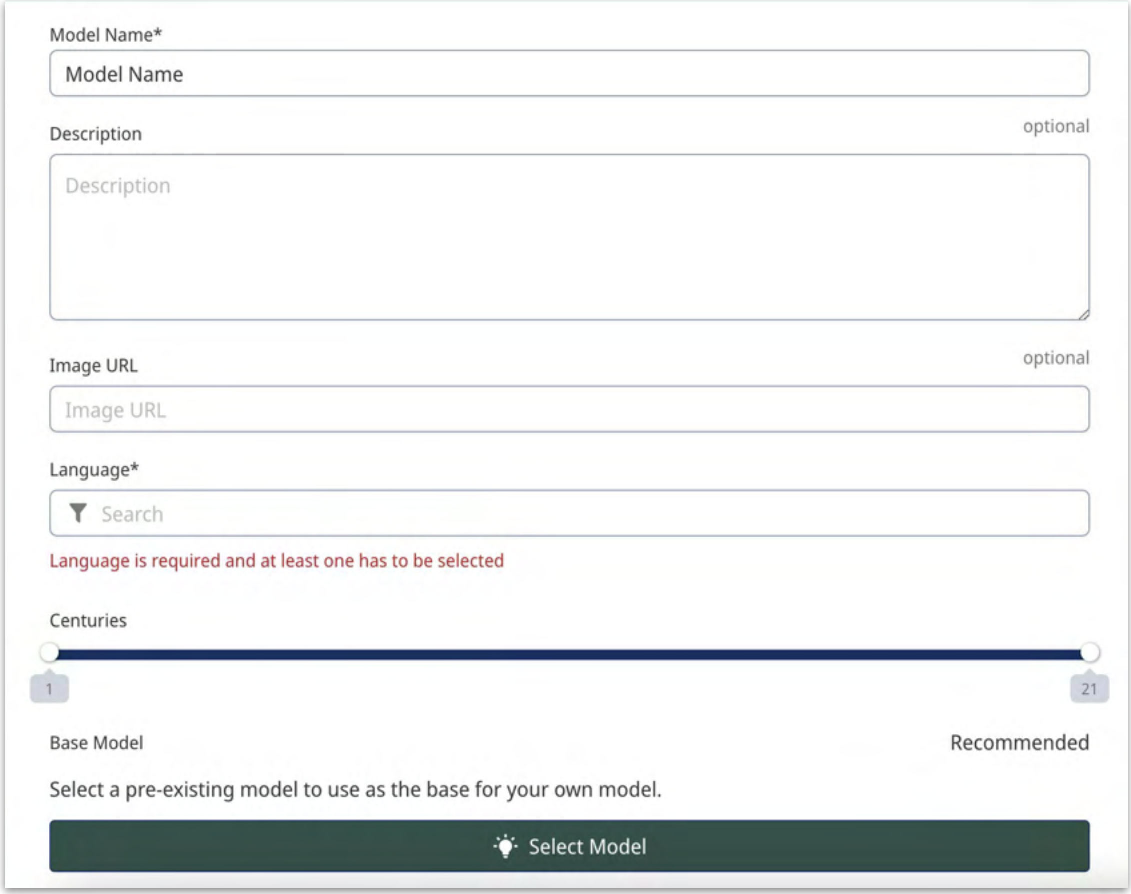
10% of your Training Data will be used as Validation Set or choose to manually select your Validation Set

1 pages

Validation Set 10% of train data

Step 3: Model Set Up

- you can enter a name of your model
- description
- Image URL
- language
- centuries
- you can already choose a **base model** (means: with a Base model, the training doesn't start from scratch but from what it has already been learnt in the base model)



Model Name*

Model Name

Description

optional

Description

Image URL

optional

Image URL

Language*

Search

Language is required and at least one has to be selected

Centuries

1

21

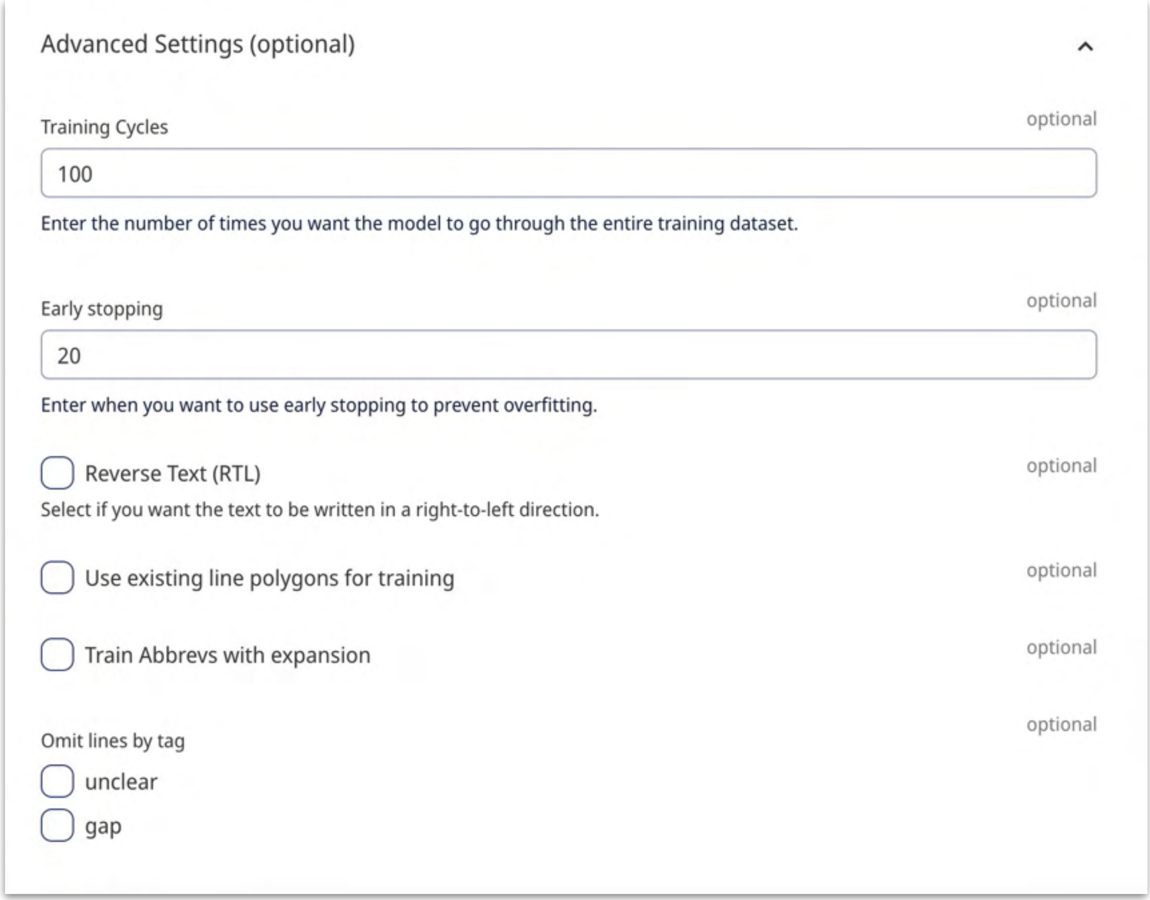
Base Model

Recommended

Select a pre-existing model to use as the base for your own model.

Step 3: Model Set Up (Advanced Settings)

- Training cycles: the number of times that the learning algorithm will go through the entire Training Data and evaluate itself on both the Training and the Validation Data
- maximum number, because the training stops automatically when the model does not improve anymore
- early stopping: minimum number of epochs for the training, meaning the model will at least run this many epochs



Advanced Settings (optional)

Training Cycles — optional

100

Enter the number of times you want the model to go through the entire training dataset.

Early stopping — optional

20

Enter when you want to use early stopping to prevent overfitting.

□ Reverse Text (RTL) — optional

Select if you want the text to be written in a right-to-left direction.

□ Use existing line polygons for training — optional

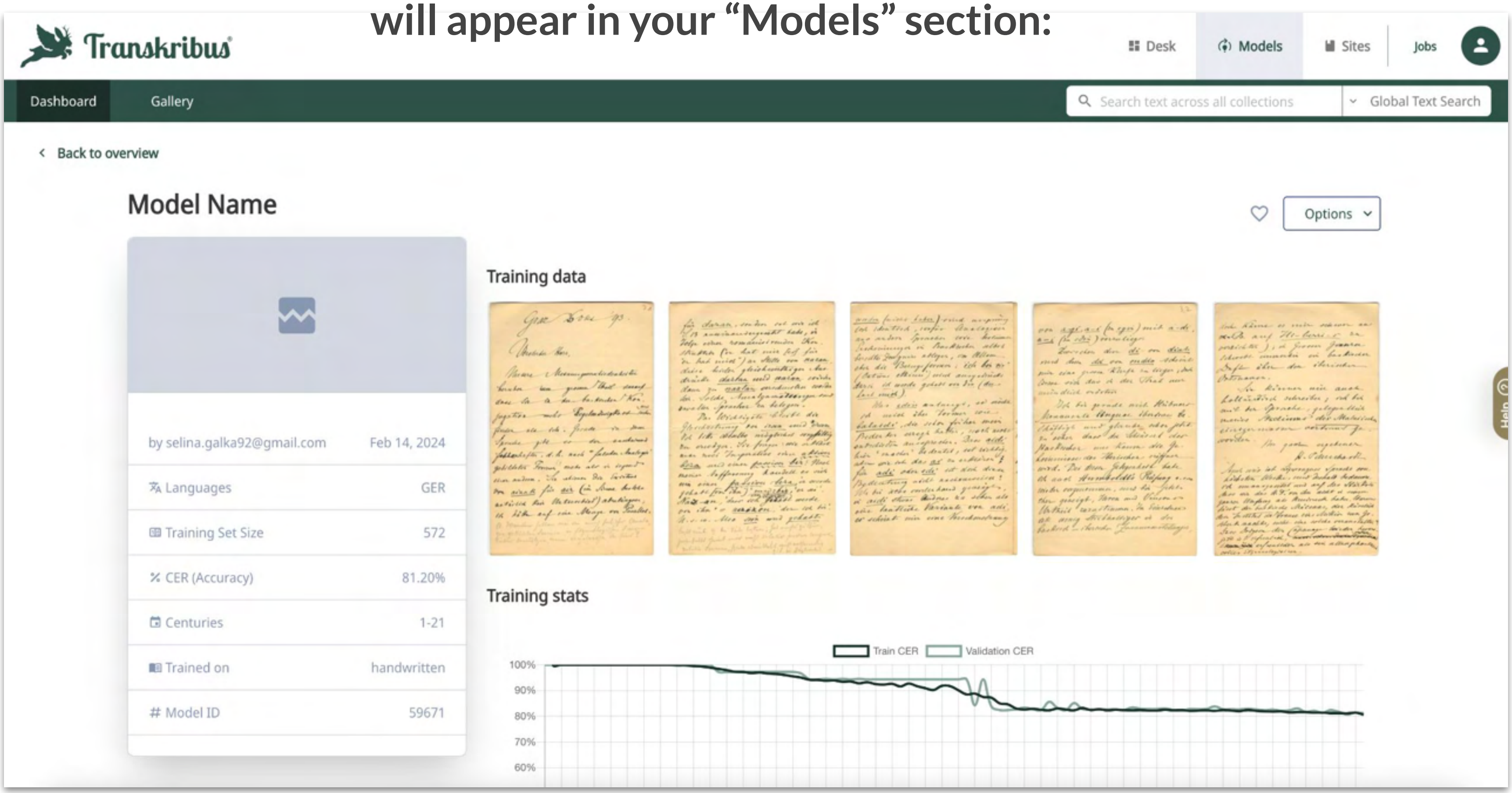
□ Train Abbrevs with expansion — optional

Omit lines by tag — optional

□ unclear

□ gap

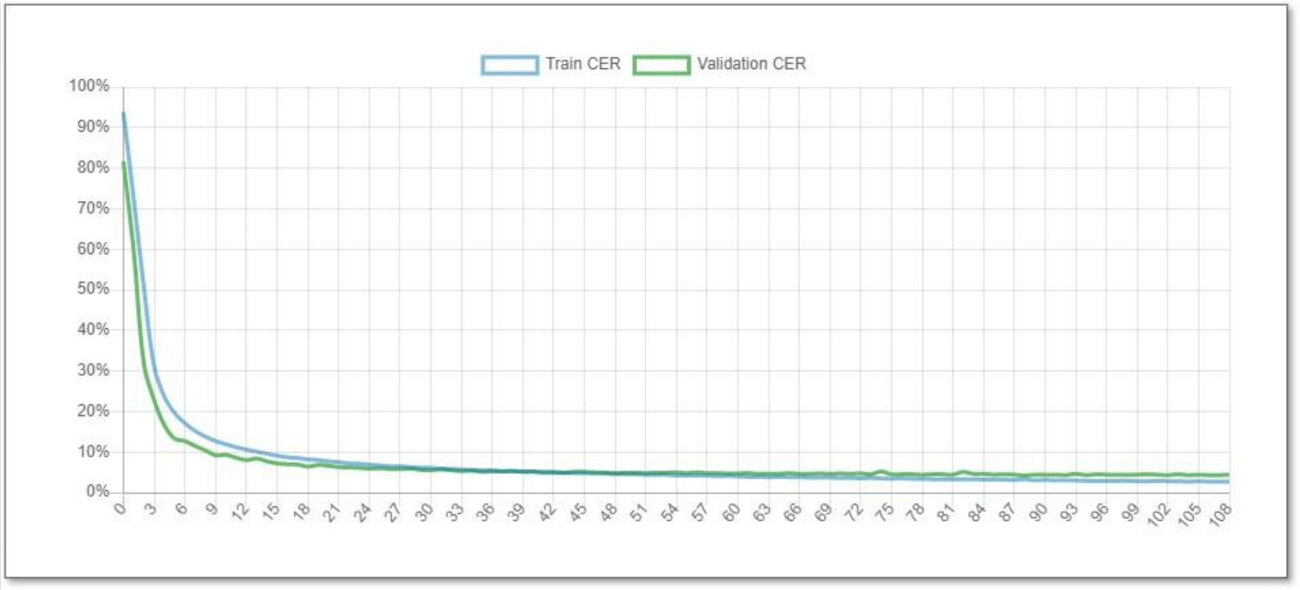
Step 4: Train the model! The model will appear in your "Models" section:



Model Name

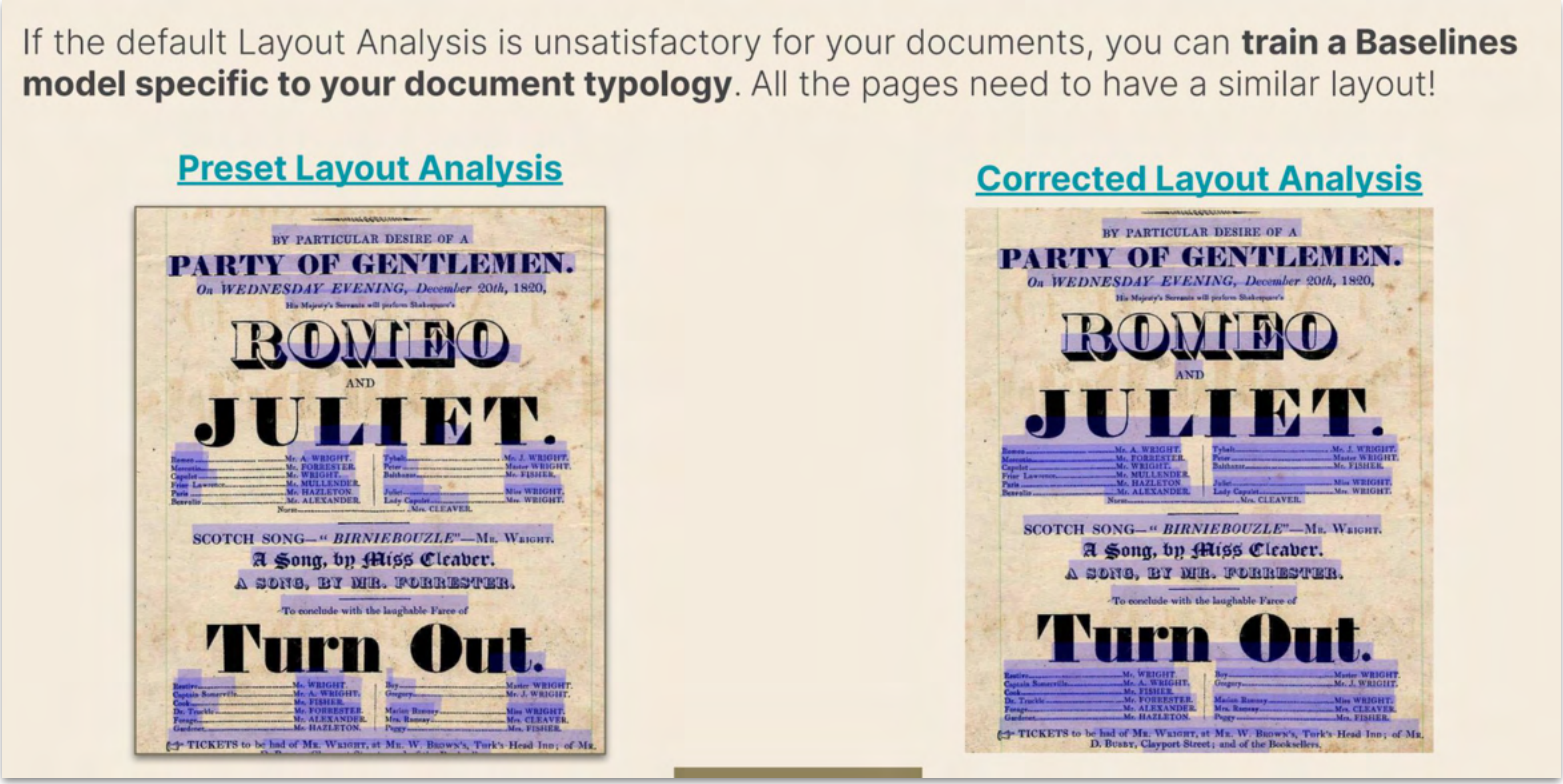
by selina.galka92@gmail.com	Feb 14, 2024
Languages	GER
Training Set Size	572
CER (Accuracy)	81.20%
Centuries	1-21
Trained on	handwritten
# Model ID	59671

Learning Curve



Training a baseline model

Training a baseline model



If the default Layout Analysis is unsatisfactory for your documents, you can **train a Baselines model specific to your document typology**. All the pages need to have a similar layout!

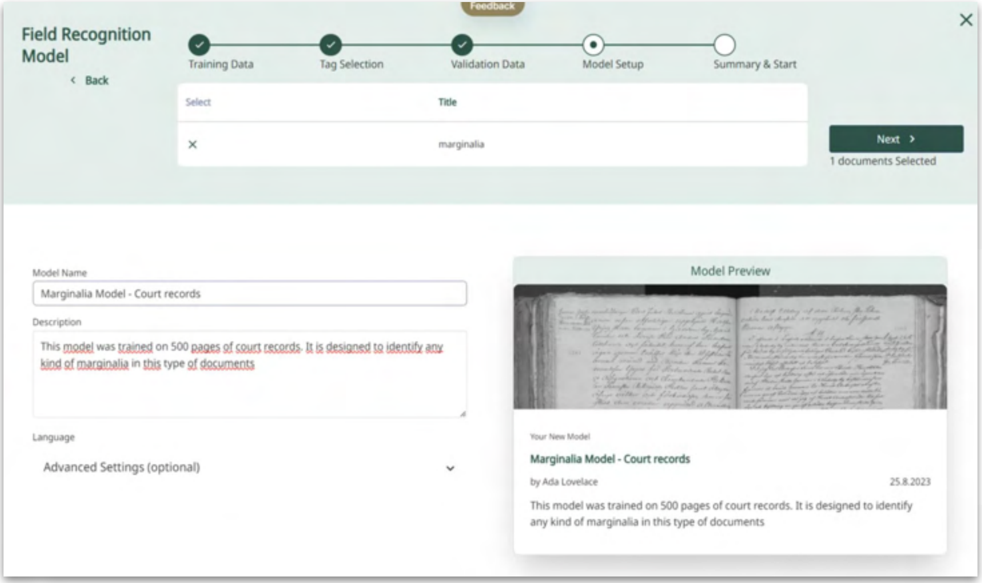
Preset Layout Analysis

Corrected Layout Analysis

BY PARTICULAR DESIRE OF A
PARTY OF GENTLEMEN.
On WEDNESDAY EVENING, December 20th, 1820,
ROMEO
AND
JULIET.
SCOTCH SONG—"BIRNIEBOUZLE"—Mr. Wright.
A Song, by Miss Cleaver.
A SONG, BY MR. FORRESTER.
To conclude with the laughable Farce of
Turn Out.

Other possibilities

- Training field models
  - enhancing the layout recognition
  - e.g. training a model for the layout recognition of marginalia
  - based on structural tags
- Table models
  - AI to identify the tabular layout of your historical documents, simplifying data extraction and export into spreadsheets



Select	Title
×	marginalia

1 documents Selected

Model Name: Marginalia Model - Court records

Description: This model was trained on 500 pages of court records. It is designed to identify any kind of marginalia in this type of documents

Language

Advanced Settings (optional)

Your New Model

Marginalia Model - Court records

by Ada Lovelace 25.8.2023

This model was trained on 500 pages of court records. It is designed to identify any kind of marginalia in this type of documents

Example from https://help.transkribus.org/field-models

Tagging

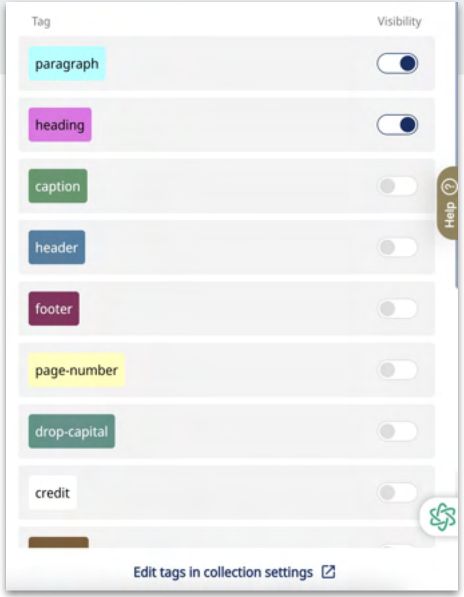
Tagging

- you can enrich your documents with tags
- can be useful for XML export

There are two different types of tags:

1. Structure tags: to mark up the structure of your documents, for instance, paragraphs, headers, marginalia or footers. They are assigned to **layout shapes** (text regions and lines) in the image.

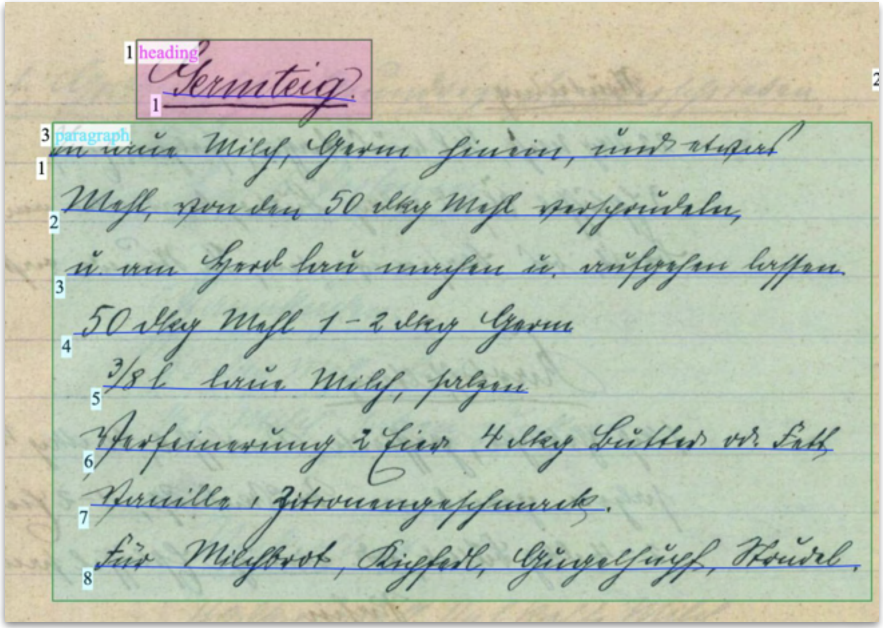
2. Textual tags: to mark up words and phrases of the transcriptions, for instance, persons, places, abbreviations, and add attributes. They are **added to words** within the transcribed text.



Tag	Visibility
paragraph	●
heading	●
caption	○
header	○
footer	○
page-number	○
drop-capital	○
credit	○

Structure tags

- can be used for **Field Models**
  - trainable models, to identify specific fields in your documents such as regions, marginalia, name fields etc.
- or when you want to **restrict the text recognition** to certain structure types instead of recognising the whole page
- they are centrally **managed at the collection level**



Germteig

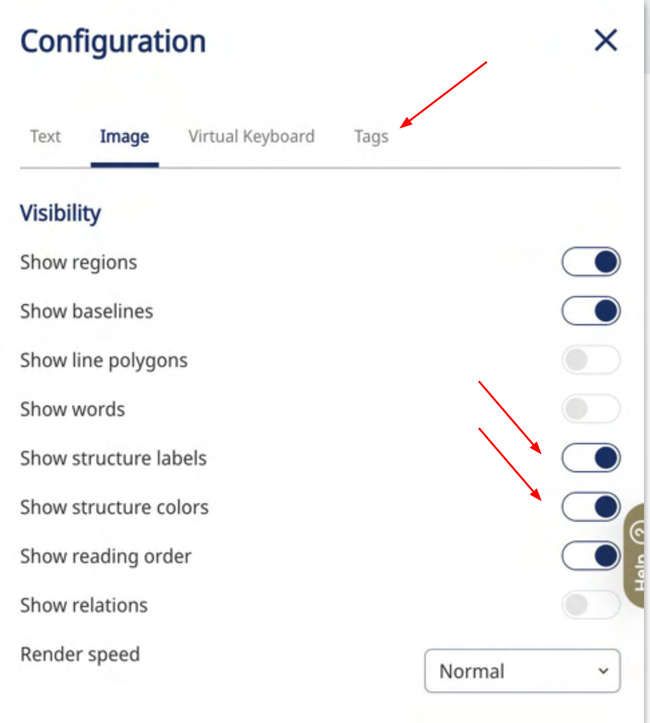
in laue Milch, Germ hinein, und etwas
Mehl, von den 50 dkg Mehl versprudeln,
u. am Herd lau machen u. aufgehen lassen.
50 dkg Mehl 1–2 dkg Germ
3/8 l laue Milch, salzen
Verfeinerung 2 Eier 4 dkg Butter od. Fett
Vanille, Zitronengeschmack.
für Milchbrot, Kipferl, Gugelhupf, Strudel

Textual tags

- Mark-up words, such as names, dates, places, and events, with textual tags to add information to your transcriptions
- Attributes provide information about the content of the tag and can be used to extract and process data from the transcription
- e.g. the date tag allows you to tag a date written in the document and add properties such as the day, month, and year in a standardised format

Work with tags

- To use them in the editor, **you need to make them visible**!
    - editor view: **“Settings” in the right corner**
    - click on “tags”
    - make tags visible, you want to see
    - you can choose between structural and textual tags
    - on the bottom: link to manage the tags for the collection
- To see tags in the image: Image - show structure labels - show structure colors!
- add a structural tag to a region: right-click
- add a textual tag: highlight the word/characters



Configuration

Text | **Image** | Virtual Keyboard | Tags

Visibility

Show regions ●

Show baselines ●

Show line polygons ○

Show words ○

Show structure labels ●

Show structure colors ●

Show reading order ●

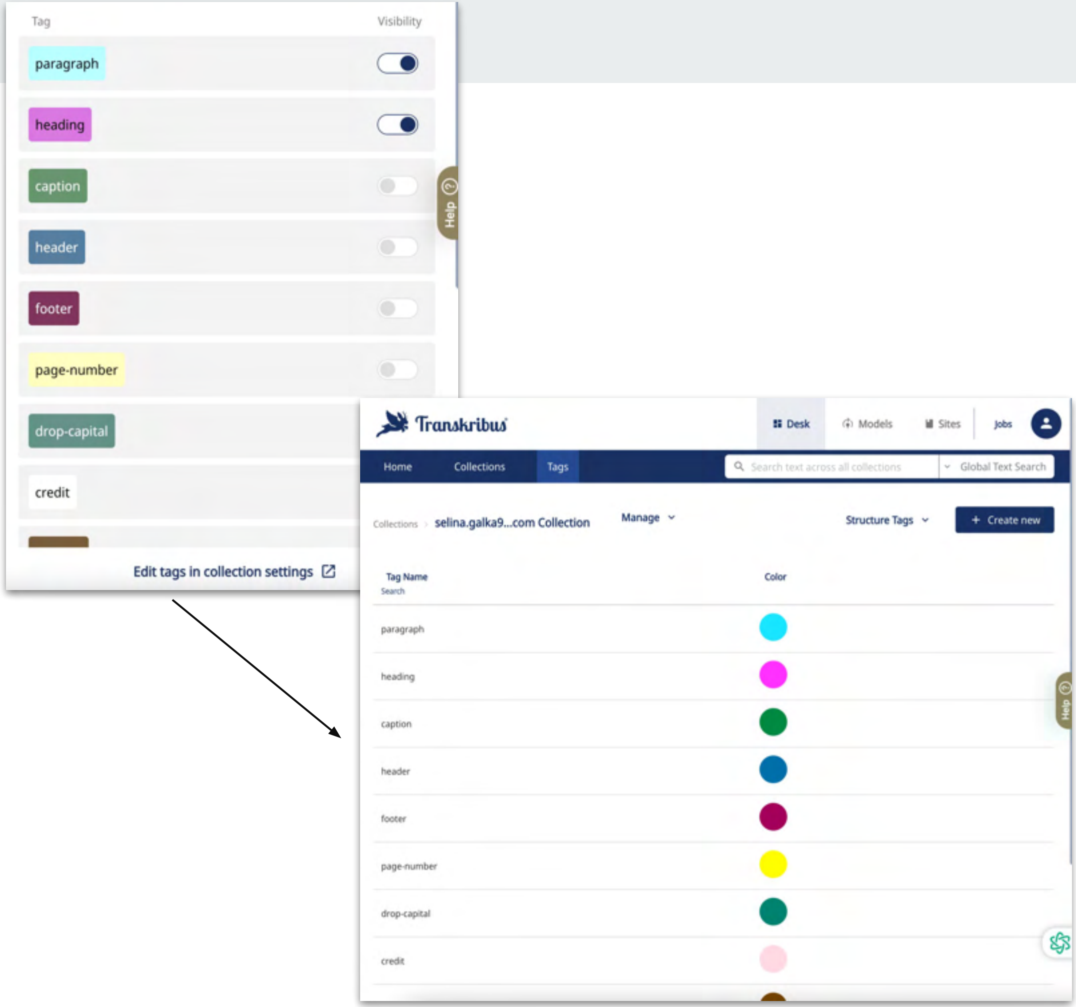
Show relations ○

Render speed: Normal

Work with tags

- Editor View → Settings → Tags → Edit tags in collection settings → this opens the window to manage the tags

- you can add new tags
- switch between structure tags and textual tags
- all the changes done here are saved only for the collection in question (the one opened in the background)



Tag	Visibility
paragraph	●
heading	●
caption	○
header	○
footer	○
page-number	○
drop-capital	
credit	

Edit tags in collection settings

Tag Name Search	Color
paragraph	●
heading	●
caption	●
header	●
footer	●
page-number	●
drop-capital	●
credit	●

Hands On

- Try to add tags to the text!
  - to the first letter from Hugo Schuchardt
- e.g. transcribe the first line
  - add tags: place, date (also attribute)
  - create new tags: opener, salute

Hands On

- create new tags: opener, salute
  - TEI <opener>: (opener) groups together dateline, byline, salutation, and similar phrases appearing as a preliminary group at the start of a division, especially of a letter
  - TEI <salute>: (salutation) contains a salutation or greeting prefixed to a foreword, dedicatory epistle, or other division of a text, or the salutation in the closing of a letter, preface, etc.

Hands On

- Try to add structural tags to the standard Germknödel recipe!
    - e.g. heading, paragraph

Searching

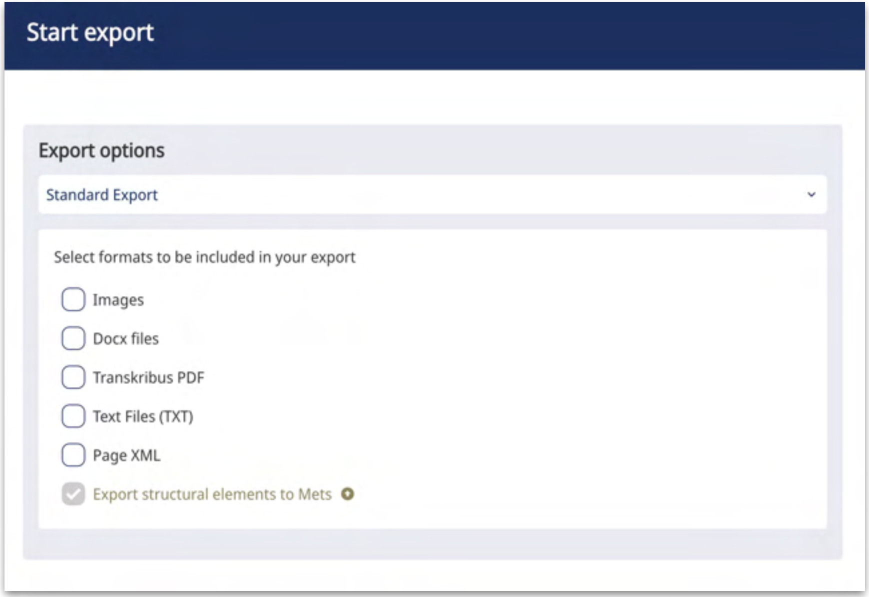
Searching

- Search tags and text
    - Full-text search: search terms can be modified using wildcards, among other things
    - Fuzzy search: Finds results that differ from the search term by one or two letters
    - Smart Search: With Smart Search, not only the automatically recognized words are saved, but also possible variants, which means that words transcribed incorrectly by the text recognition model can also be found. For this type of search to be possible, it must be selected before the text recognition is carried out. It is associated with 50% higher costs for text recognition, as it is more memory and computationally intensive

Export

Export (Free Plan)

- Images
    - choose between original or compressed jpg
- Docx-Files
    - transcriptions in Word-Files, it's also possible to export it with tags
- PDF
    - you can choose with or without tags
- TXT



Start export

Export options

Standard Export

Select formats to be included in your export

☐ Images

☐ Docx files

☐ Transkribus PDF

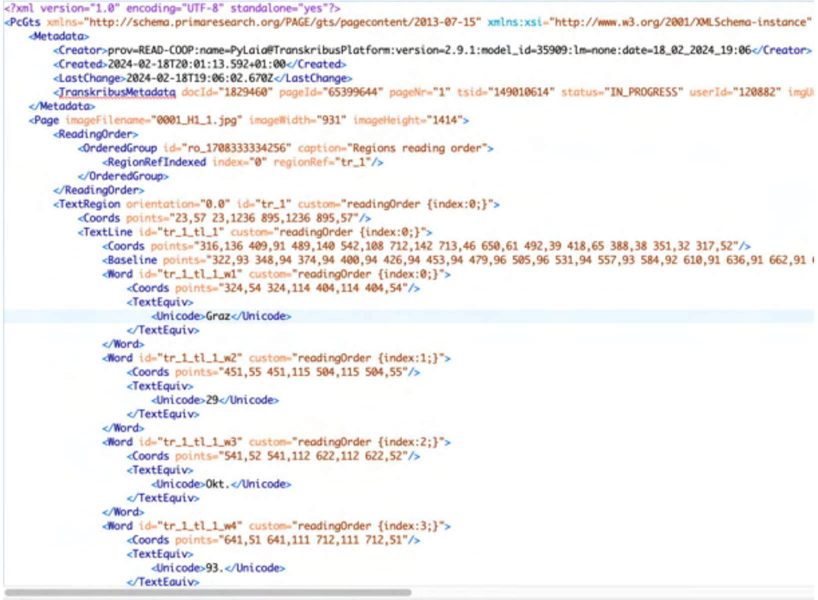
☐ Text Files (TXT)

☐ Page XML

☑ Export structural elements to Mets

Export (Free Plan)

- Page XML
  - XML-based page image representation framework that records information on image characteristics in addition to layout structure and page content
  - can e.g. be used to describe page content like regions, lines of text, words, glyphs, etc.
  - important as import- and export format for OCR/HTR Software bzw. Tools



```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    <Metadata>
        <Creator>prov=READ-COOP:name=PyLaia@TranskribusPlatform:version=2.9.1:model_id=35909:lm=none:date=18_02_2024_19:06</Creator>
        <Created>2024-02-18T20:01:13.592+01:00</Created>
        <LastChange>2024-02-18T19:06:02.670Z</LastChange>
        <TranskribusMetadata docId="1829460" pageId="65399644" pageNr="1" tsid="149010614" status="IN_PROGRESS" userId="120882" imgU
    </Metadata>
    <Page imageFilename="0001_H1_1.jpg" imageWidth="931" imageHeight="1414">
        <ReadingOrder>
            <OrderedGroup id="ro_1708333334256" caption="Regions reading order">
                <RegionRefIndexed index="0" regionRef="tr_1"/>
            </OrderedGroup>
        </ReadingOrder>
        <TextRegion orientation="0.0" id="tr_1" custom="readingOrder {index:0;}">
            <Coords points="23,57 23,1236 895,1236 895,57"/>
            <TextLine id="tr_1_tl_1" custom="readingOrder {index:0;}">
                <Coords points="316,136 409,91 489,140 542,108 712,142 713,46 650,61 492,39 418,65 388,38 351,32 317,52"/>
                <Baseline points="322,93 348,94 374,94 400,94 426,94 453,94 479,96 505,96 531,94 557,93 584,92 610,91 636,91 662,91
                <Word id="tr_1_tl_1_w1" custom="readingOrder {index:0;}">
                    <Coords points="324,54 324,114 404,114 404,54"/>
                    <TextEquiv>
                        <Unicode>Graz</Unicode>
                    </TextEquiv>
                </Word>
                <Word id="tr_1_tl_1_w2" custom="readingOrder {index:1;}">
                    <Coords points="451,55 451,115 504,115 504,55"/>
                    <TextEquiv>
                        <Unicode>29</Unicode>
                    </TextEquiv>
                </Word>
                <Word id="tr_1_tl_1_w3" custom="readingOrder {index:2;}">
                    <Coords points="541,52 541,112 622,112 622,52"/>
                    <TextEquiv>
                        <Unicode>Okt.</Unicode>
                    </TextEquiv>
                </Word>
                <Word id="tr_1_tl_1_w4" custom="readingOrder {index:3;}">
                    <Coords points="641,51 641,111 712,111 712,51"/>
                    <TextEquiv>
                        <Unicode>93.</Unicode>
                    </TextEquiv>
```

Export (Scholar Plan)

- Spreadsheet
    - Table Export
    - Tag Export
    - Page Metadata
- TEI/XML
    - not possible at the moment?
- ALTO XML
    - XML file for each page, containing the content and layout information of the page.
    - It is often used in combination with METS for the description of the whole digitized object

Hands on

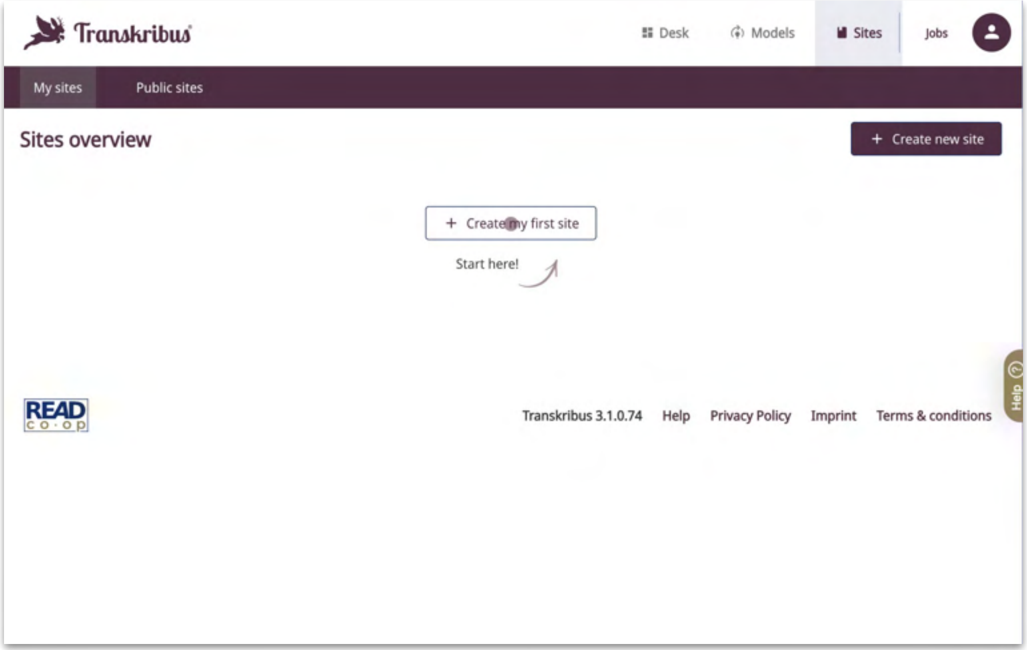
- Try out the export function!

You get the export via your registered E-Mail.

Transkribus Sites

Transkribus Sites

- one of the main pages in the Transkribus app (button on the top)
- included in the scholar plan
  - for publishing you need a paid subscription!
- platform to publish and share your digital documents online
- offering search capabilities



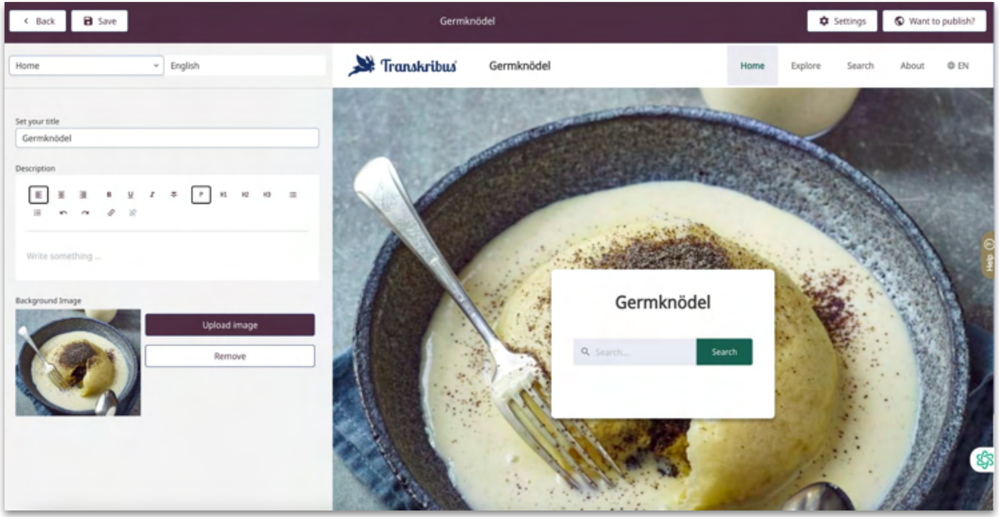
Sites overview

Transkribus 3.1.0.74 Help Privacy Policy Imprint Terms & conditions

Transkribus Sites

“flexible viewing options”

- 4 basic pages:
  - Home (Picture and Search Window)
  - About (Pictures and Description of the project)
  - Explore (List of Documents)
  - Search
- Settings (right corner): delete the site, change theme, manage users



Set your title
Germknödel

Description

Write something ...

Background Image

Hands on

- Have a look at the Sites page and try it out!

Conclusion

Conclusion: Strengths of Transkribus

- No software download or installation required - only a web browser is needed
- No hardware requirements as text and layout recognition and model training is performed on READ COOP SCE's servers in Innsbruck (Austria)
- As the data is stored on European servers
- Multilingual user interface (de, en, es, et, fi, fr, it, nl, pl, pt, sl, sv)
- hands-on training of your own models
- Suitable for collaborative work on transcriptions with other Transkribus users

Conclusion: Strengths of Transkribus

- Own structure and text tags can be defined so that conformity with the TEI guidelines can be achieved
  - Integration of normdata possible (Wikidata ID)
- Right-to-left writing direction is supported
- Smart Search (higher credit consumption): Not only a recognized word, but also alternatives are saved, so that (incorrectly) recognized words can be searched for easily

Conclusion: Challenges

- Transkribus expert client is not developed further
- replaced by the WebApp → not yet as powerful as the expert client
- Training HTR models can be time-consuming and have a high error rate with very heterogeneous manuscripts
- Although (simple) annotations can be made and custom tags can be created, the editor is not a fully-fledged replacement for a standalone annotation tool
- No internal communication tool to coordinate with other users working together on a collection (no comment function, no place where guidelines for transcription can be stored, etc.)
- Text recognition with free credits has lower priority than with purchased credits

General & Resources

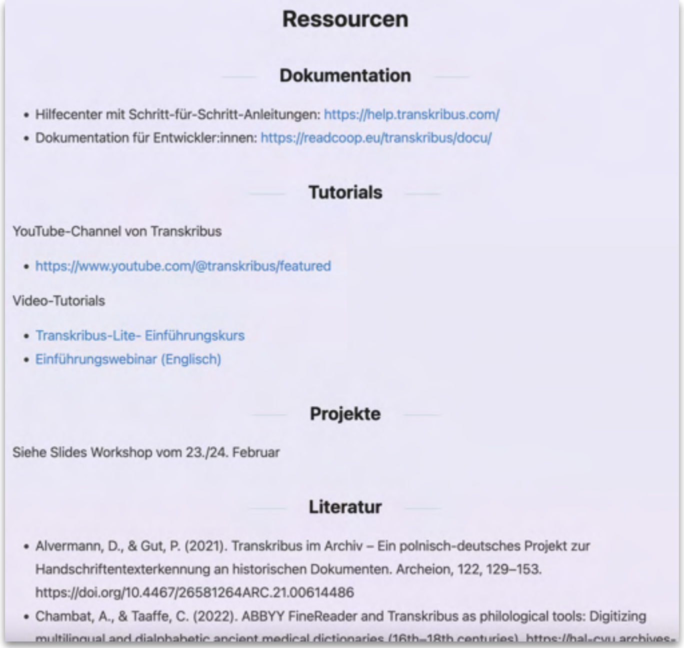
- The transcription software has established itself as an essential tool in the creation of Digital Editions

Resources

- The READ website's Resource Centre contains numerous instructions on how to use Transkribus
  - https://help.transkribus.org/
- Documentation for developers
  - https://readcoop.eu/de/transkribus/docu/
- forText also offers instructions on how to digitize manuscripts with Transkribus
  - https://fortext.net/routinen/lerneinheiten/manuskriptdigitalisierung-mit-transkribus
- There are also various field reports

Ressources: Transkribus

- Project DigEdTnT (“Digital Edition Creation Pipelines: Tools and Transitions”)
  - https://digedtnt.github.io/transkribus/#ressourcen
  - overview, description of the tool
  - in german, but at the end there is a comprehensive list of resources (**documentation, tutorials, projects, literature**)



Ressourcen

Dokumentation

- Hilfecenter mit Schritt-für-Schritt-Anleitungen: https://help.transkribus.com/
- Dokumentation für Entwickler:innen: https://readcoop.eu/transkribus/docu/

Tutorials

YouTube-Channel von Transkribus

- https://www.youtube.com/@transkribus/featured

Video-Tutorials

- Transkribus-Lite- Einführungskurs
- Einführungswebinar (Englisch)

Projekte

Siehe Slides Workshop vom 23./24. Februar

Literatur

- Alvermann, D., & Gut, P. (2021). Transkribus im Archiv – Ein polnisch-deutsches Projekt zur Handschriftentexterkennung an historischen Dokumenten. Archeion, 122, 129–153. https://doi.org/10.4467/26581264ARC.21.00614486
- Chambat, A., & Taaffe, C. (2022). ABBYY FineReader and Transkribus as philological tools: Digitizing multilingual and dialphabetic ancient medical dictionaries (16th–18th centuries). https://hal-cyu.archives-