

Introduction to Transcribing Documents

Why do we transcribe documents? On one level, transcribing something allows you to become intimately familiar with the contents of a document. Furthermore, transcribing makes physical documents, especially handwritten documents, **machine readable**. In other words, it allows computers to recognize text, which gives researchers more potential to analyze and use that text for digital scholarship. Think about it -- scanned letters are pictures of text, which is something a machine cannot translate. Even though there are programs that can attempt to read handwriting (like Transkribus) and even typed text, computers aren't smart enough to read things humans are!

Transcribing something bridges the gap between reading a document, and storing a picture of one.

In this module, you will learn to:

- Transform an archival, primary source document into a machine-readable file with a transcribing software (Transkribus)
- Add metadata tags to a primary source document, and understand why doing so enriches the file, enhances discoverability, and creates opportunities for further research
- Use a tool that allows for group project management, like Asana or Microsoft Teams, by tracking progress and communicating with group members synchronously and asynchronously
- Create different file types from transcribing and tagging a letter and understand their purposes (.txt, .docx, .xlsx, and .pdf)

Lessons

1. Transcribing Guidelines
2. Text Regions in Transkribus
3. Handwritten Text Recognition
4. Metadata Tags
5. Exporting Files in Transkribus

Before you begin:

Every person has preconceived beliefs that affect how they think and act - this is true for documents created by people a long time ago and for us, reading those documents. Those beliefs, or “bias,” can sometimes contribute to the misinterpretation of information, and can skew characterizations of the content. In reading these documents, be aware of potential biases that exist, and your own bias that might come into play as you transcribe and annotate the document with tags. Try to remain objective when assigning tags by following classification standard schedules, and when there's room for interpretation, to remain as objective as possible to preserve the integrity of the original materials being described.

Recommended Reading

- More reading on transcribing
 - Transkribus Transcribing Conventions, 2021. Read Co-op, <https://readcoop.eu/transkribus/howto/transkribus-transcription-conventions/>.
 - Smithsonian, "General Instructions for Transcription and Review," <https://transcription.si.edu/instructions>
 - Transkribus How-tos and videos: <https://readcoop.eu/transkribus/resources/video>
 - Hillary Richardson, YouTube video of transcribing a letter from beginning to end: <https://youtu.be/-cDD9P0rnLw> (good for watching while transcribing)
- More reading on project management
 - [How to Create a Work Breakdown Structure](#)
 - [Critical Path Method](#)

Installations

- Transkribus: <https://readcoop.eu/transkribus/?sc=Transkribus> (Installation Instructions here. Be sure you have Java 8 installed before installing Transkribus. This is in the instructions, but not in the first step. If you run into an issue, email your instructor!) Transkribus Lite is the browser-based version of this, but as of this writing is still under development. We recommend downloading the free desktop version of this software.
- Project management: There are several browser-based options to keep track of your progress and communicate with team members the documents that you are working with. These tools can help keep multiple moving parts of a project in sync like keeping track of supplies, correspondence, and tasks needed to make it to completion:
 - [Asana](#) - for organizing and assigning tasks as well as mapping them out in a timeline to track progress and milestones.
 - [Trello](#) - allows you to visualize the tasks and focus needed, through lists and cards, to complete the project
 - [Google Sheets](#) template for transcribing (make a copy of this template to edit it) - similar to Asana and Trello in managing tasks, but in a spreadsheet format that allows for collaborate editing

1. Transcribing Guidelines

In order to maintain the integrity of your original document while you are translating it for the computer, consider a few things while you're transcribing.

Stay as true to the document as you can!

- Include all text as it is written. This includes errors in spelling and grammar, and words you might not be familiar with. **Transcribing is NOT editing!**
- If there are images in the text, describe them in brackets (e.g. "[doodle of a curly-haired smiling face]").

- Keep it simple! You want people to be able to read the transcription without looking at the original letter. Your transcriptions will be plain text files (e.g. files with no formatting), so, without editorializing, keep the text readable as possible.

If you have trouble reading the text...

- Take a break and come back. Fresh eyes make a difference!
- Try to read the words in context. The word might look like "fhe," but it doesn't make sense to read "And then fhe called her sister." You would realize the word is "she."
- Ask for help. Send a screenshot to your classmates in your project management group or post it to a social media account for a second opinion.
- If you've tried all these things, and it's just not possible to figure out, write [illegible] in brackets and tag that text with the "gap" tag (more on tags in section IV!).

Take notes while you are transcribing







- Make note of things you don't understand, and Google them to see if you can make sense of the reference.
- You might find something interesting that you want to come back to! Take notes of the letters you transcribe, and what you find interesting about them.

While it is long, we recommend watching [this video](#) of transcribing a letter from start to finish. Feel free to follow along in corresponding sections!

Remember you're doing this as a team


- Use your project management tools to mark your progress and note what you need to go back to later

✓ Mark complete




mcj-dp018-19410507-smith-sonnyboy

Assignee


 Hillary Richardson

Recently assigned ▾

Due date

 No due date

Projects

 Digitized Metadata and Transcriptions

Letters added Summer 2021 ▾


Description

Add more detail to this task...

Subtasks

✓ Metadata in Google Sheet


✓ Review Metadata for QC

✓ Transcribe 

✓ Add Tags



✓ Review Transcription and Tags for QC

+ Add subtask


 Hillary Richardson created this task. Jun 15

- Communicate with your team what you've done and what you need help with through notes and comments. Tag or notify people if something else needs addressing.



Subtasks



- ✓ Metadata in google sheet
- ✓ Review metadata for quality control
- ✓ Transcribe 
- ✓ Tag 
- ✓ Review transcription and tags for quality control



+ Add subtask



 **Hillary Richardson** created this task. 29 days ago

Hillary Richardson added to [Digitized Metadata and Transcriptions](#). 29 days ago

 **Hillary Richardson** 22 days ago 
 Hey again, @Stephanie Salvaterra - this also looks like it's missing a page. Can you check this one too when you get a chance?

 **Stephanie Salvaterra** 21 days ago  1
 That one needs to be re-scanned anyway, I'll do that today

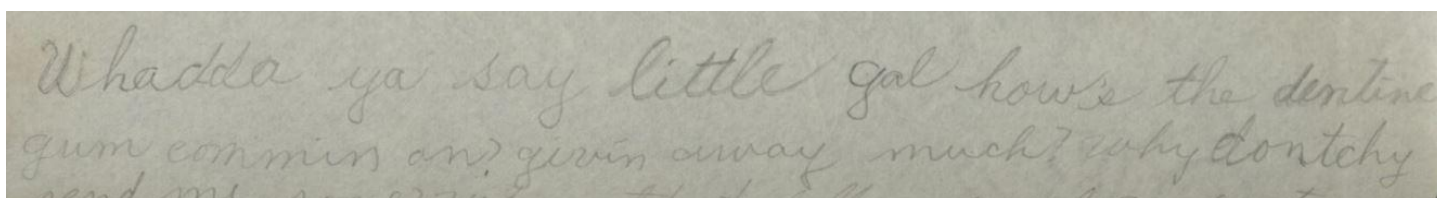
 **Hillary Richardson** 21 days ago 
 Thank you!

 **Stephanie Salvaterra** 17 days ago 
 Re-scanned!

- To track your progress, you can create an account and join your group's channel, or you can copy this project management template in [Google Sheets](#) for your own use.

Check-in

Practice transcribing a few lines, staying true to the original text with the sample excerpt:



Compare transcriptions and decide what the ideal transcript would look like.

Fill out your project management tool (using either the template or a chosen app) with the items you have to transcribe, and share the list with your groupmates.

2. Text Regions in Transkribus

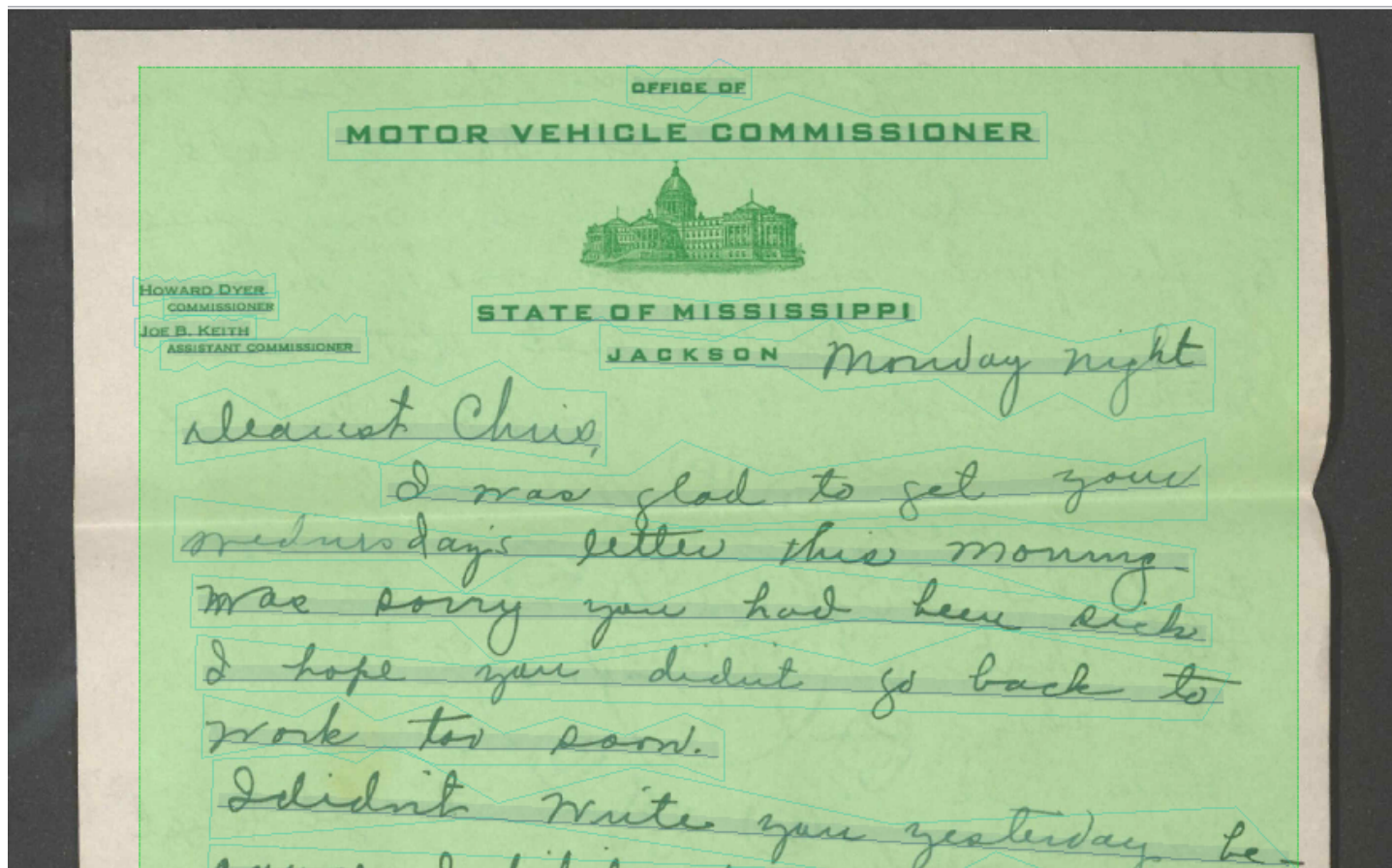
Before we get started actually transcribing, we have to tell Transkribus what parts of the page have text on them, and what order to read that text! (Remember, computers aren't smart enough to read like humans!) The following is a basic step-by-step for the most common kinds of errors you'll correct.

A full documentation of the Transkribus layout tools is at:

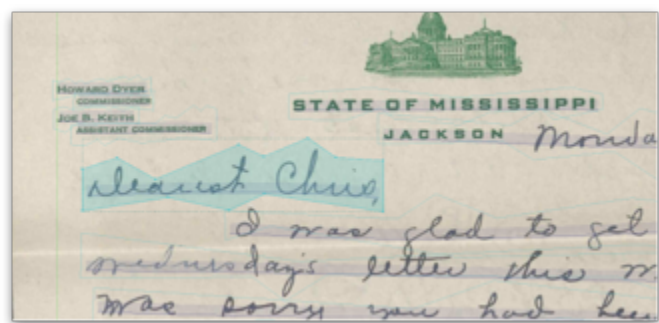
https://readcoop.eu/transkribus/howto/how-to-transcribe-documents-with-transkribus-introduction/#elementor-toc_heading-anchor-3.

The computer identifies areas of the page that have text on them. They are text regions, lines, and baselines.

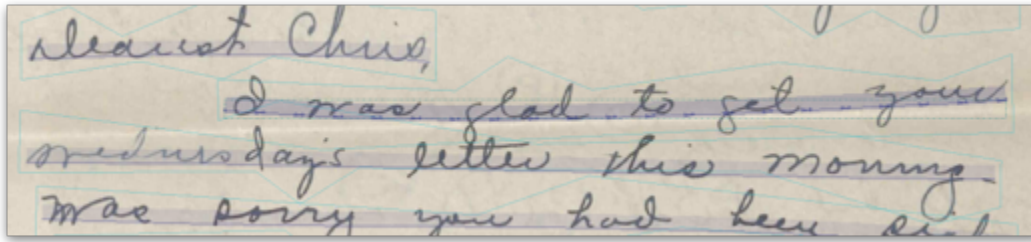
The page (or "**text region**"), where text is:



The line of text, or **line region**, reads as a left-to-right:



And the **baseline**, the most important reference point for text recognition, runs along the bottom of the handwritten text line. The letters should sit on the baseline as it underlines the words themselves:



If you make changes on lines, it is important to always do it on the baselines. This is important to know because for every line in your document, there is also a line region in the background.

Run a Layout Analysis

Before we can ask Transkribus to read the handwriting, we have to ask it to identify the text regions for us. The quickest way to do this is to run a layout analysis on the document, then correct any mistakes that Transkribus makes. Here's how you run the analysis:

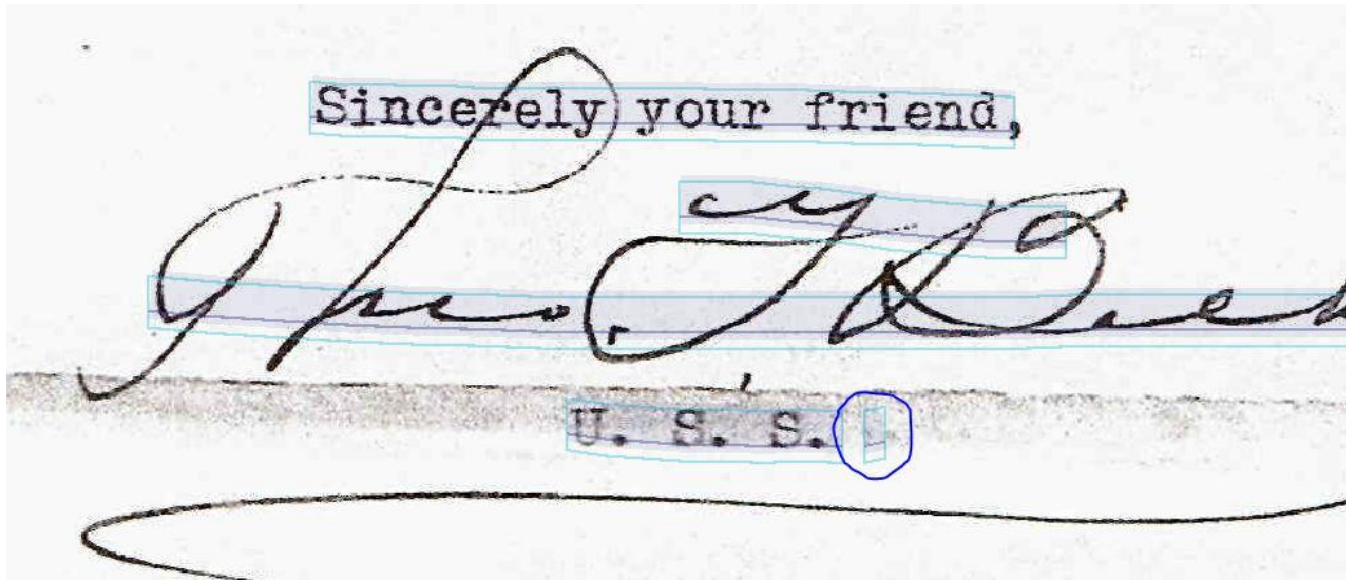
1. Double click the document you've uploaded, and click on the Tools tab.
2. Under Layout Analysis, make sure you've selected all pages.
3. Click Run

A screenshot of the Transkribus web interface, specifically the 'Tools' tab. The 'Layout Analysis' section is active. The 'Method' is set to 'CITlab Advanced'. The 'Pages (8)' dropdown is set to '1-8'. The 'Find Text Regions' and 'Find Lines in Text Regions' checkboxes are both checked. A green button labeled 'Run' is visible. Below this, the 'Text Recognition' section is visible, with the 'Method' set to 'HTR (CITlab HTR+ & PyLaia)'. There are buttons for 'Models...', 'Train...', and 'Run...'. The 'Tools' tab is circled in the top navigation bar, and the 'Run' button is also circled.

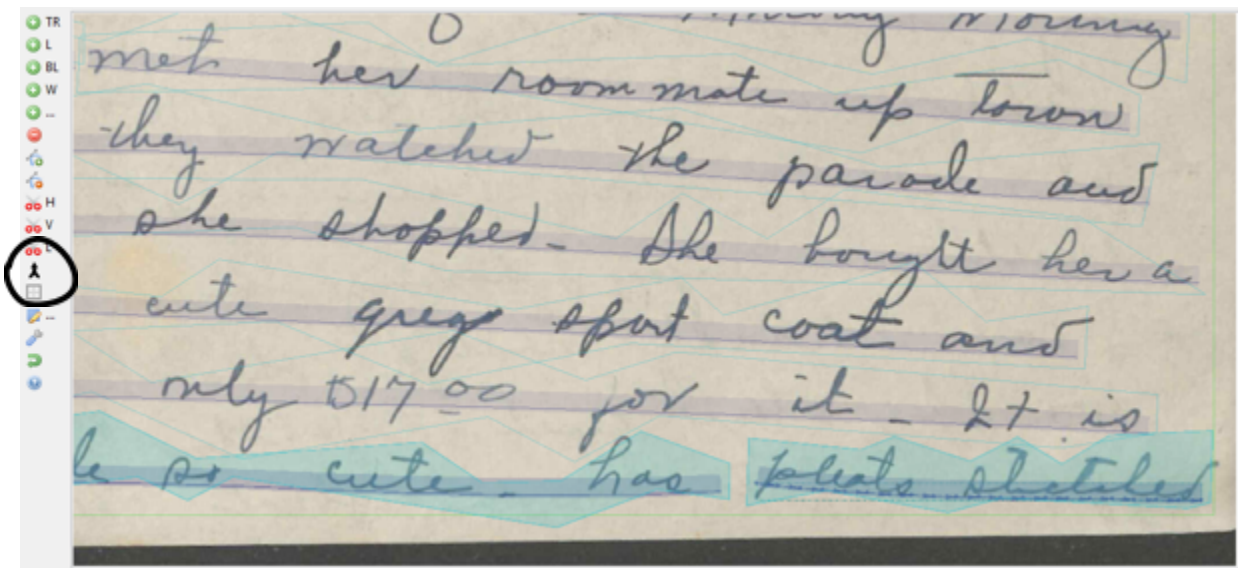
4. A notification that the layout analysis is in progress will pop up. After a few moments (between 3 and 10 seconds), the layout (green squares, blue rectangles, and dark blue lines) will appear over the letter's text.

5. Correct errors in the layout (e.g. delete text regions or lines that don't cover text, resize regions that cut off text, reorder line numbers, or merge lines that have been split). Screenshots of example corrections:

- Delete text regions that don't actually contain any transcribe-able text.

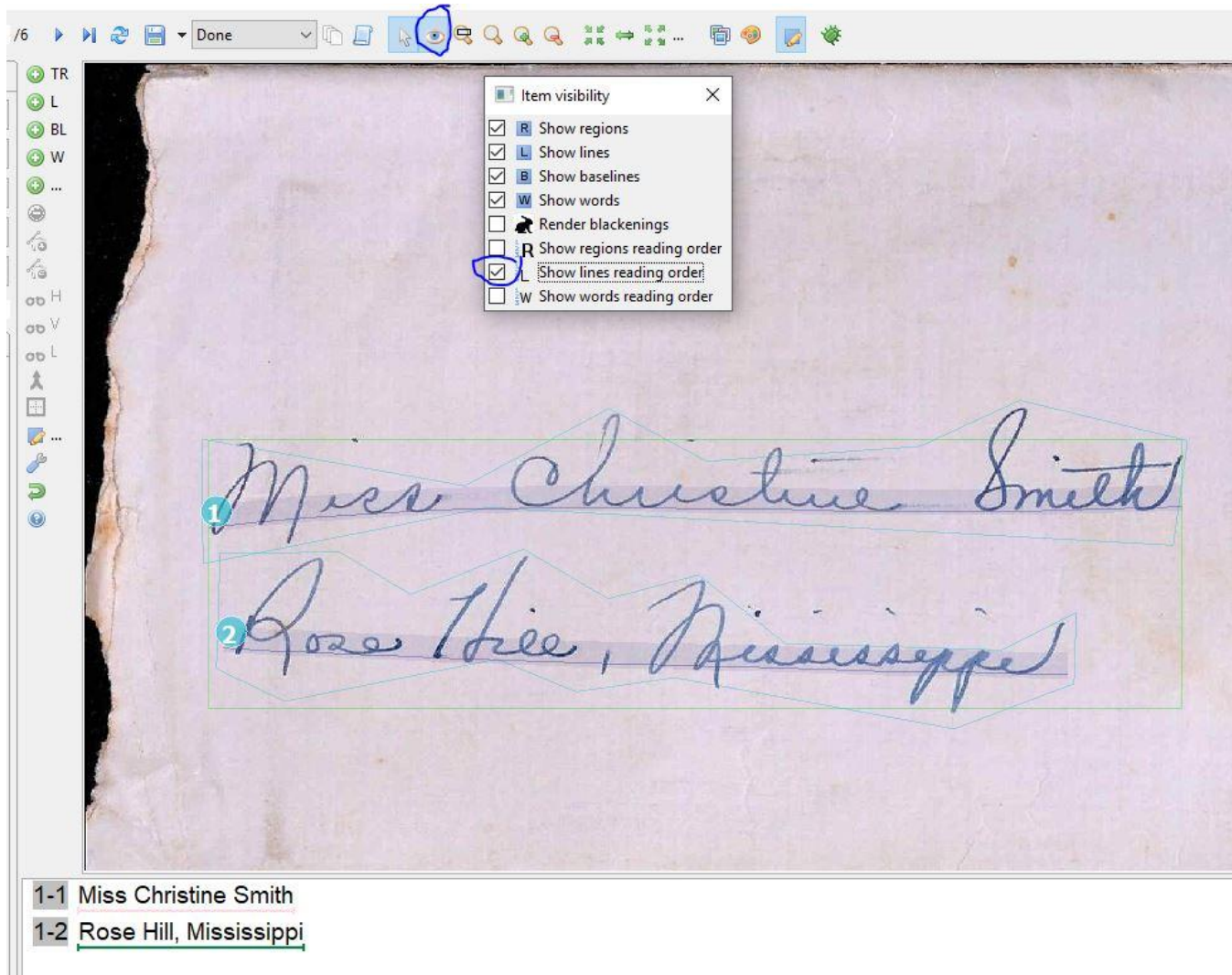


- This layout analysis split a line in two, and needs merging. (Sometimes, though less frequently, it will merge something that should be two lines. For this, use the scissor tools above the merge tool.) Holding control, click each region that needs joining (usually the baseline and the line) and then click merge.



- Sometimes the software will number things in a different order than we'd actually read them. Click on the eye in the top toolbar, and select show lines reading order. If the

lines are out of order, click on the numbers to change them.



Check-in

What do you do if Transkribus has created text regions where there isn't any text?

What do you do if Transkribus has split one line into two (or more) lines?

What button do you click on if Transkribus has ordered the lines in a different way than they're actually read?

Discuss: what is the purpose of layout recognition? Why does Transkribus need to know where text appears and what order lines are read in any way?

3. Handwritten Text Recognition

We are using the Transkribus software because it allows you to create Handwritten Text Recognition (HTR) models. This tool works with mixed results, depending on the quality of your letters and the style of handwriting. In this lesson, you will see the use of an HTR model that has had 2 years of

training and different iterations, and is based on ~100 transcriptions. While there are models for different time periods and nationalities already in Transkribus, you may find that the algorithm does not work with your dataset as well as it should. We've provided a sample letter to use with the "Pauline 2.0" HTR model, but you may choose a different HTR model depending on what you need to transcribe. A full step-by-step guide for training and using HTR models is [linked on Transkribus' website](#).

Using a model on letters gives a more complete transcription to correct, rather than having to transcribe the letter from scratch, though transcribing from scratch is still an option.

1. (To transcribe manually, skip to step 4.) To transcribe with the HTR model, click on the Tools tab, and under Text Recognition, select the model called "Pauline Smith 2.0." You might have to navigate to the next page to find the Pauline Smith model.

The screenshot shows the 'Tools' tab in the Transkribus interface. It is divided into two main sections: 'Layout Analysis' and 'Text Recognition'.

Layout Analysis:

- Method:** A dropdown menu set to 'CITlab Advanced' with a 'Configure...' button to its right.
- Options:** Two radio buttons: 'Current page' (unselected) and 'Pages (8):' (selected). The 'Pages (8):' option has a text input field containing '1-8' and a three-dot menu button to its right.
- Checkboxes:** Two checked checkboxes: 'Find Text Regions' and 'Find Lines in Text Regions'.
- Buttons:** A green button labeled 'Only use unsegmented pages' and a large 'Run' button with a green arrow icon.

Text Recognition:

- Method:** A dropdown menu set to 'HTR (CITlab HTR+ & PyLaia)'.
- Buttons:** A 'Models...' button with a puzzle piece icon, a 'Train...' button with a play icon, and a large 'Run...' button with a green arrow icon.

Hand-drawn arrows point to the 'Models...' button and the 'Run...' button. Below the 'Text Recognition' section, there are links for 'Compute Accuracy...' and 'Other Tools'.

Choose a model

All

All engines

26-50 / 71

1

2

3

Name	Language	Curator	Technology	Created	ID	nrOfWords
EvenkiRussian-RychkovArchive-v...	Evenki, Russian	sarkipo@yandex.ru	CITlab HT...	30.03.20	22629	40561
Medieval Protocolbook 's-Herto...	Latin and Dutch	geertui.vansyngh...	CITlab HT...	26.03.20	22563	46638
Neolatin_Ravenstein_1643-1772	Latin	info@caromein.nl	CITlab HT...	21.03.20	22408	64435
Pauline Smith	English	hhrichardson@m...	CITlab HT...	04.03.20	22024	11476
Dutch Mountains (18th Century)	Dutch	jirsireinders1989@...	CITlab HT...	20.02.20	21683	1384893
Danish Fraktur SB 19th century v...	Danish	poul.steen@gmail...	CITlab HT...	20.02.20	21669	30983
Dutch manuscript poetry 1603-1...	Dutch	b.j.m.caers@hum...	CITlab HT...	04.02.20	21187	51788
Gjentofte 1881-1913 Denmark 10...	Danish	spam2@busk.dk	CITlab HT...	01.02.20	21104	154388
Usberg	Dutch	vincent.noppe@n...	CITlab HT...	04.01.20	20209	1544683
Charter Scripts XIII-XV_M1	German, Latin,...	tobias.hodel@uzh...	CITlab HT...	22.12.19	19872	938325
German_Kurrent_XIX_M1	German	tobias.hodel@uzh...	CITlab HT...	21.12.19	19829	3037966
English writing M2	English	guenter	CITlab HT...	20.12.19	19809	1234795
German_Kurrent_XVI-XVIII_M1	German	tobias.hodel@uzh...	CITlab HT...	15.12.19	19584	1579208
Dutch_Romantype_Print	Dutch	info@caromein.nl	CITlab HT...	11.12.19	19423	88105
Gothic_Book_Scripts_XIII-XV_M4	Latin,German	tobias.hodel@uzh...	CITlab HT...	08.12.19	19285	875728
Français ANOM	French	maxime.gohier@...	CITlab HT...	06.12.19	19244	128175
French_18thC_Print	French	info@caromein.nl	CITlab HT...	05.12.19	19166	38487
Danish 1870-1950	Danish	ajam@aarhus.dk	CITlab HT...	29.11.19	18968	348716
Dutch_Gothic_Print	Dutch (16th, 1...	info@caromein.nl	CITlab HT...	27.11.19	18944	51143
LaMOP-Livre_Rouge_1	french	pierre.brochard@...	CITlab HT...	27.11.19	18909	20358
RoyalDanishLibrary_20thCentury+	Danish	jme@kb.dk	CITlab HT...	23.11.19	18805	580371
NAF Court Records M10	Swedish	guenter	CITlab HT...	07.11.19	18284	1226202
Edelfelt M13+	swedish	maria.vainio-kurta...	CITlab HT...	31.10.19	18107	438817
German Fraktur 18th Century - W...	18th ct. Austri...	Dario.Kampkaspar...	CITlab HT...	21.10.19	17785	829447
Sam and Pauline Smith Letters	English	hhrichardson@m...	CITlab HT...	11.10.19	17543	22088

25

Filter

Details

Name: Pauline Smith

Language: English

Description: 1937 correspondence

Parameters: Nr. of Epochs 200

Document Type: Handwritten

Show advanced parameters...

Nr. of Words: 11476

Nr. of Lines: 1933

Save

Show Train Set

Show Validation Set

Show Characters

Learning Curve

CER on Train Set: 0.88% CER on Validation Set: 8.44%

OK

Cancel

- Click Run. The HTR analysis may take anywhere from 20 seconds to 2 minutes. A notification that the "job" is done will pop up when analysis is complete. From there, correct the transcription.
- Transcribe each page as you see it, following [Transkribus transcription guidelines](#). You may have a few errors on a page or there may be significant errors. It depends on several things - the quality of the scan, the quality of the document, the handwriting and what utensil they used, etc.!
- To transcribe manually, use the layout analysis to begin typing each line as it's written. You can use the enter key to move to each line. Moving to the next page will prompt you to save your work.

1-1 21

1-2 SENATOR S. H. SMITH

1-3 Pittsboro, Miss.

2-1 1n

3-1 1

3-2 Miss Christine Smith

3-3 18s0 Inion It

3-4 Memphis, Serin.

4-1 OCT

4-2 18

4-3 1937

4-4 MISS.

5-1 1

5. After you've completed editing the transcription, go back over the letter, making sure you didn't miss typos, and you can simultaneously begin tagging the document, which will be explained in the next lesson! Don't forget to track your progress in the project management tool!

Check-in

What does HTR stand for?

True or false - HTR replaces our need to transcribe documents at all.

Discuss: what do you do if you haven't trained an HTR model in Transkribus?

4. Metadata Tags

Tagging, annotating, or marking-up a document is a common practice for text analysis. Why is this? Think about the things someone might talk to you about in a letter. They might mention people they talked about, places they went, or a book they read. Now think about the contents of 6,000 letters, and all of this information together! Tagging will allow us to use the information in the collection of thousands of letters to track these trends over time, explore topics, and find people.

Think of tagging the letter like highlighting categories of information with a different color highlighter. All of the people's names are blue, the places are purple, etc. Transkribus does the same thing.

Categories of metadata tags

It is tempting to tag everything, but not necessary! You only need to tag:

1. Proper nouns (e.g. people, places, or organizations)
2. When you add something for clarity that isn't there originally (e.g. "[illegible]").
3. If you want to provide further information as context.

Metadata tags, defined.

As a group, you will need to decide which tags you'll use, and what their definitions will be.. Examples of tags and definitions that already exist in Transkribus include:

1. person - when any person's name is mentioned within the body of the letter (e.g. "Daddy," Martha, Gilbert, etc.). You can also decide to tag people without names, i.e. "girl," or "baby."
2. place - when a place name is mentioned within the body of the letter. This does not usually include general places like Sunday school or hospital. Examples would be town names like Bruce or Memphis.
3. organization - a specific group, corporation, or entity (e.g. the Senate, the T.V.A., or Ole Miss), not a generic place like hospital or school.
4. sic - for an error in the original letter (i.e. spelling, grammar, etc.) that you include it in the transcript for accuracy, tag it with "sic," showing that it's "as is," and you didn't make the error in transcribing.
5. date - a specific date (e.g. January 15th, 1948). This does not include general references to days of the week or years (e.g. "Saturday nite" or "next year.")
6. unclear - when you write something in the transcription, but aren't sure of what it actually says in the letter: a guess.
7. gap - when you include something (i.e. [illegible], a note, etc.) that isn't in the original letter to give clarity, but to prevent the HTR from picking up the note.

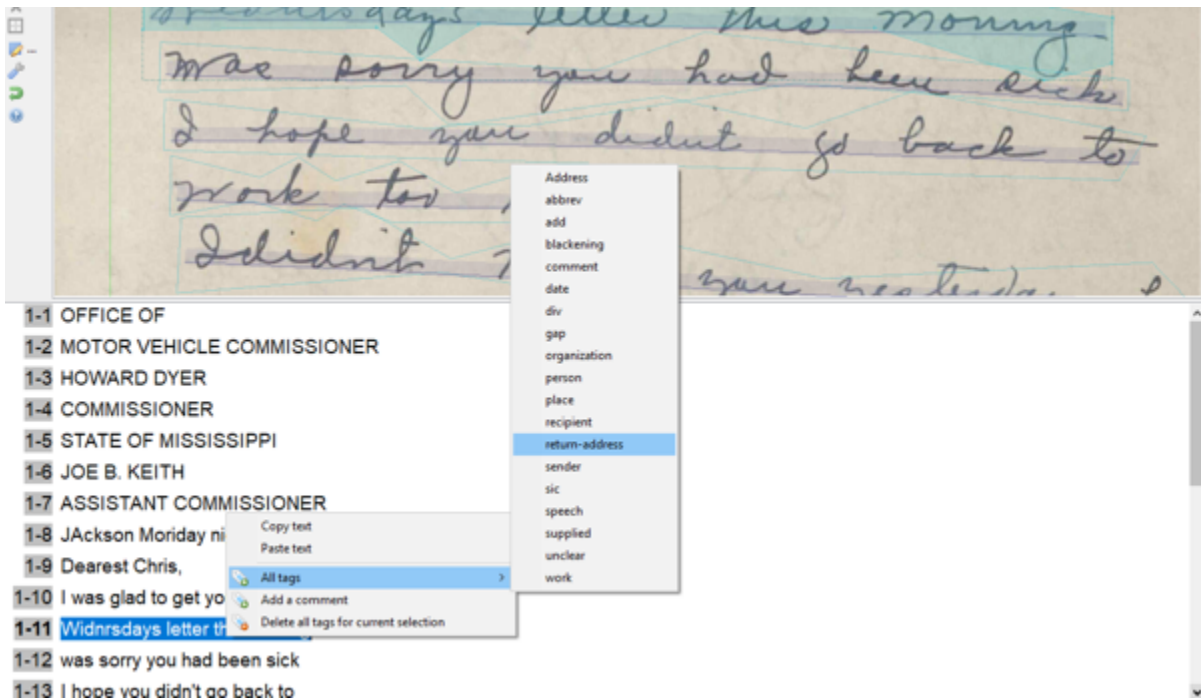
Depending on what's in your collection, you may also decide to create custom tags. For the examples in this lesson, we used a collection of correspondence, and in order to track who was sending letters, who was receiving them, where they were coming from and where they were going to, we created the following custom tags:

1. sender - The person who sent the letter, either in the return address, or in the letter's closing.

2. recipient - The person the letter is addressed to, either on the envelope, or in the salutation.
3. address - The address the letter was sent to.
4. return-address - the address the letter came from (not always included)
5. title - the title of a book, play, radio show, government decision, etc. Anything that is published, for example. Specifications may include full title, author, and publication date.

How to tag something

1. To tag something, highlight the text you want to tag, right-click, and from “All tags,” select the relevant tag.



2. If the tag isn't in the list, go to the metadata menu, then the Textual menu, and click Customize.

Server
Overview
Layout
Metadata
Tools

Document
Page
Structural
Textual
Comments

Tags of current Transcript

	Tag	Value	Text	Properties
1	textStyle	OFFICE	DAYS RETURN TO OFFICE OF MOTC	strikethrough: tr
2	textStyle	MOTOF	RETURN TO OFFICE OF MOTOR VEH	strikethrough: tr
3	textStyle	STATE C	VEHICLE COMMISSIONER STATE OF	strikethrough: tr
4	textStyle	JACKSC	OF MISSISSIPPI JACKSON Box 3 Pitt	strikethrough: tr
5	return-address	Box 3	MISSISSIPPI JACKSON Box 3 Pittsbo	
6	return-address	Pittsboi	JACKSON Box 3 Pittsboro Miss Miss	
7	return-address	Miss	Box 3 Pittsboro Miss Miss Martha Sn	
8	person	Martha	Pittsboro Miss Miss Martha Smith 1:	
9	Address	134 fiftl	Miss Martha Smith 134 fifth Ave. Mc	
10	Address	McCon	134 fifth Ave. McComb Miss	
11	textStyle	c	134 fifth Ave. McComb Miss	fontSize: 0.0 kerr
12	date	OCT	PITTSBORO OCT 2 A.M. 1939 MISS	
13	date	2	PITTSBORO OCT 2 A.M. 1939 MISS	
14	date	A.M.	PITTSBORO OCT 2 A.M. 1939 MISS	
15	date	1939	PITTSBORO OCT 2 A.M. 1939 MISS	

Tags

Tag Specification
☐ User tags
☐ Collection tags
☐ Tag specifications
 Customize...

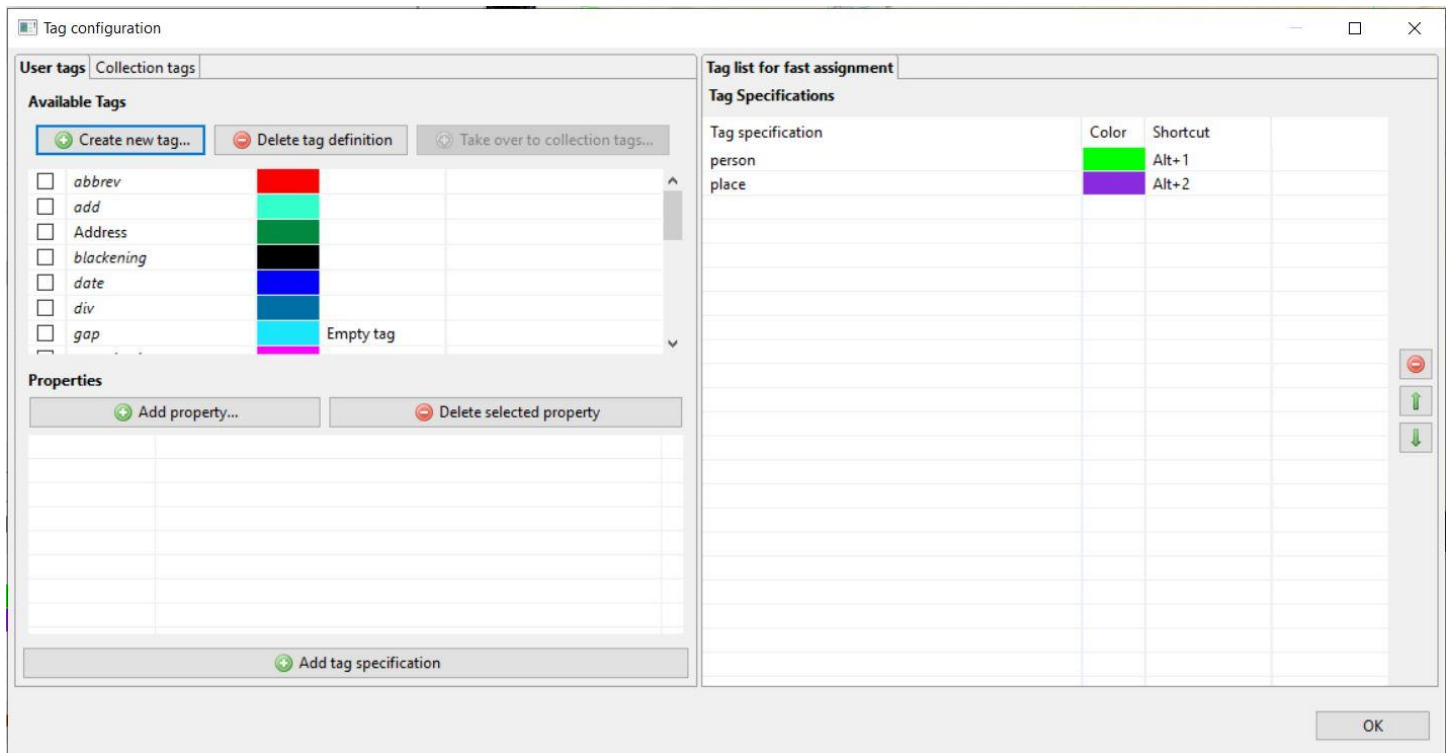
Tag specification	Color	Shortcut	
person		Alt+1	
place		Alt+2	

Props for tag: 'person' - value: 'Martha Smith'

dateOfBirth
dateOfDeath

Previous
Next
 Apply to selected

- Click "Create new tag," then type in the name of the tag, lowercase, like it's shown in the Tag definitions document.



Check-in

Tags are tricky! Let's discuss some different scenarios.

Scenario 1: You are tagging a document that says, "We went to a football game between Louisiana State University and Texas A&M." Would you tag the two universities as places or organizations?

Scenario 2: If a document mentions a fictional person, do you tag them with "person"?

Scenario 3: What do you tag [illegible] with?

Scenario 4: What is a scenario where a person or place is mentioned, but you might *not* tag them?

Scenario 5: What's an example of a time you might use the comment tag?

Discuss: Does your collection warrant specific tags? Think about what you'd like to track or analyze, and list custom tags that would create points of connection for later analyses of the collection.

As a reminder, there is a video of transcribing a letter from beginning to end, which you can watch as a refresher: <https://youtu.be/-cDD9P0rnLw>.

5. Exporting files in Transkribus

After you have transcribed (or corrected a transcription of) a letter, tagged it, and proofread it, it has been digitally transformed, and we can now use the transcriptions and metadata to create

visualizations through text analysis, network analysis, and more! But before you can do that, you have to export the files.

Before the step-by-step, we recommend watching a 3-minute video on this process:

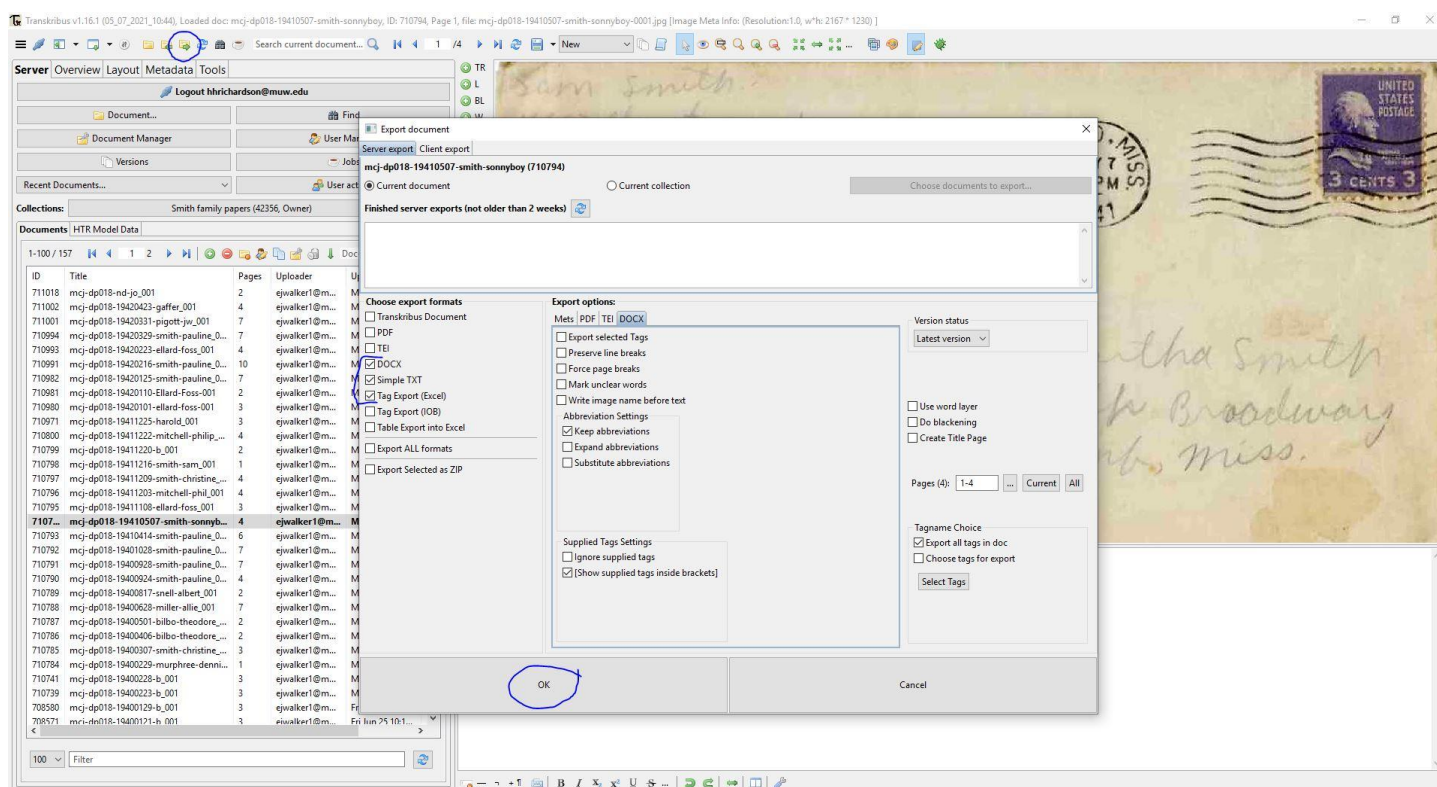
<https://youtu.be/k21t3jGL4LM>.

Export

Click the export button (a manila folder with a green, right-facing arrow), and you will choose the following files to export:

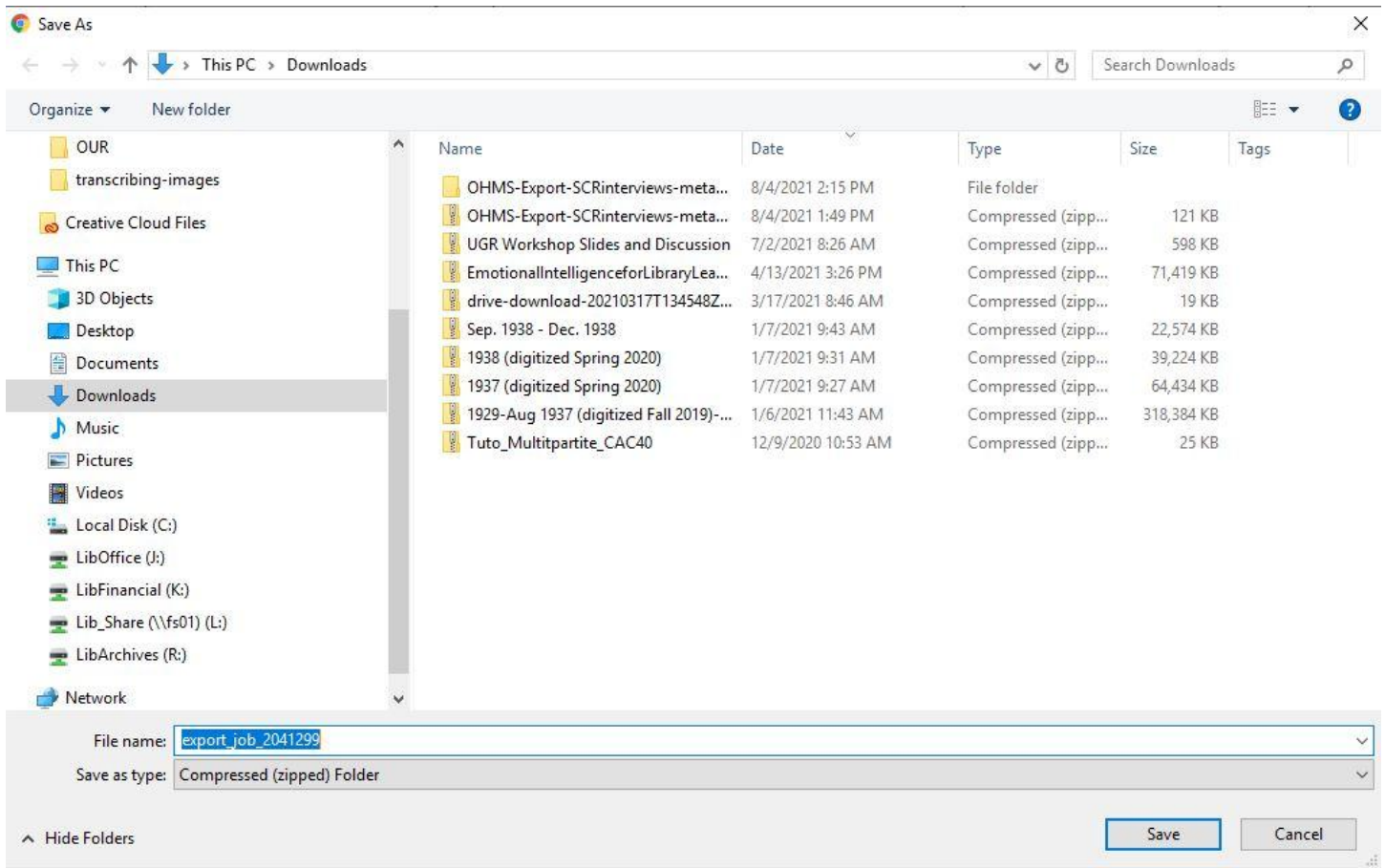
- .docx - to give you a formatted version of the transcript
- .txt - to give a non-formatted version of the transcript, and create a file that is more universally accessible, both to text readers, and to text analysis softwares.
- .xlsx (tag export) - to give you structured lists of the tags you created

If you would like to export more than one document, select the radio button that says "Current Collection," and choose which letters to export.

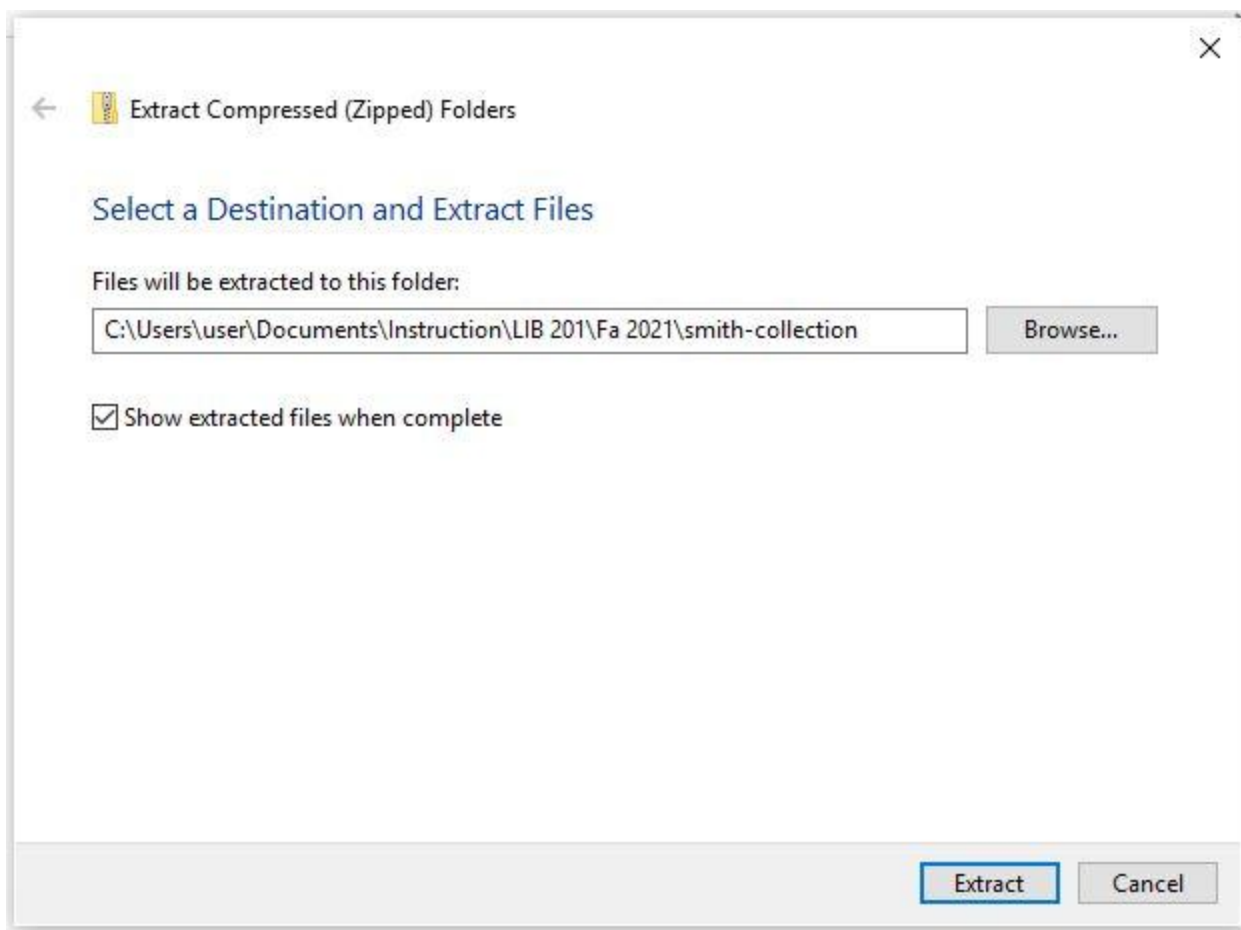


Extract zipped files

When you click OK, Transkribus emails you a temporary link where you can access the compressed files you just created. Compressing them allows them to send large amounts of data over email. When you click the link, it will ask you to save the compressed files (.zip).



Saving them as compressed files stores them to your computer, but does not let you access them until you extract them. Right-click your compressed folder and click **Extract All...** then tell your computer where you want to extract them to. I recommend saving them to a folder where you're keeping files for this class. (We will also store them on a cloud server, e.g. Google Drive or our Institutional Repository at some point.)



After you extract the files, you should notice that your folder no longer has a zipper on it, and you have 3 files that represent digital versions of the letter you prepared. You are now ready to start playing with these digital artifacts!