Preparing "Letters as Data"

Hillary A. H. Richardson
http://orcid.org/0000-0002-7522-7918
@hillaryAHR
Mississippi University for Women

Artifact type: Assignment
Developed for: Lower-level undergraduate seminar, specifically developed for students enrolled in LIB 201: Introduction to Digital Studies at Mississippi University for Women. The assignment could be made into a workshop for any student or staff or faculty member preparing a handwritten text, or archives volunteers.
Audience: Non-majors or majors in introductory humanities courses
Time required: This is one part of a 3 part project. Transcription with this process is designed to take about 1 class session per letter transcribed, plus outside class time for reading and written responses. Depending on the size of the collection, this could take a few class sessions. The entire project will require a full semester, and is intended to be ongoing.
Method and tool: Text encoding with Transkribus

Description:

The objective of this project is to create image and text files for handwritten documents (mostly letters) to prepare a collection of computation-ready files that students (including the ones in this class) and researchers can access and manipulate.

The project is divided into 3 parts. Each of these parts corresponds to different weeks of "LIB 201: Introduction to Digital Studies," at Mississippi University for Women, and will be completed in class (required reading and writing assignments will take place outside of class). This assignment focuses on Part II. A brief description of parts I-III follows.

I. *Part I: Create a digital facsimile and its metadata*
   a. Scan all parts of the letter(s) using a high-resolution scanner, ideally one that is specifically for archival quality documents, to create an image file. For this project, we are creating TIFF files for the archives, PDF files (1 file per letter) for the image, and TXT files for the transcription. In creating PDF and TXT files through Transkribus, we will 1) sync them together for a machine-readable, high quality image, and 2) train the software through its Handwritten Text Recognition (HTR) Model to recognize the handwriting.
   b. Create metadata for the files. Using selected metadata standards, an agreed-upon data management plan, and shared servers, students will generate metadata for each file

II. *Transcribe the letters for computation (described in depth in the included materials)*
   a. Create transcriptions
   b. Add metadata tags to the transcriptions
   c. Training the HTR model in Transkribus,
   d. Add files from this process into the University's online repository.

III. Create a computational artifact

a. For the final part of the class, the PDFs, TXTs, XMLs, and CSV files generated from the process will be available for computational analysis. The first half of the semester will be preparing the documents for a collection as data, and the second half will be using that data.

b. The prepared letters will be examined in conjunction with other files available within MUW's subscriptions to digital archives. This project will be determined by the class after some discussion and examination of particular collections, but will aim to digitally display this collection to the public alongside its contemporary artifacts. Examples could include a comparative reading, a network analysis, a geospatial analysis, or a digital exhibit.

Supporting materials:
Assignment guidelines ("Part II: "Transcribing letters for computation"), transcription peer review worksheet, and training model assessment response questions.

Transcription Peer Review Worksheet
LIB 201
Due: TBA

Complete each section by explaining how the transcriptions either do or do not meet the criteria in the rubric. Yes/no answers should be explained in detail. If the transcription is missing something, give a suggestion for how it could be improved.

I. Layout Analysis

Criteria	Notes
Are the handwritten markings all contained within the text regions?	
Is the order of the baselines in a way that reads the words in correct order, including marginal or superscript words?	

II. Transcription

Criteria	Notes
Do the spelling, punctuation, and markings in the transcription accurately reflect the handwritten document?	
If there are added notes in the transcription, do they increase the clarity? Are they tagged appropriately so the HTR doesn't associate them with the document?	
Is the transcription complete? Are all sections with text represented in the transcription?	

III. Tagging

Criteria	Notes
Are the tags complete and consistent throughout the document?	
Are the tags appropriate for the text highlighted (e.g. "place" indicates a location, and not just a name of a state)	
Are the categories and properties for the tags appropriately fleshed out?	

HTR Training Model Assessment
LIB 201
Due: TBA

Respond to each of the questions with 50-100 words each on the training of the HTR model in Transkribus.

1. Look at the **test set**. Justify why these were chosen to explain what, in these letters, made them representative of the collection.

2. Look at the **learning curve** of the HTR model. The CER (Character Error Rate) should make a downward slope toward a number below 10% indicating the efficiency of the automated transcription. What could be done to improve this efficiency? (Note: if the percent efficiency is already well below 10%, explain what you thought made it so efficient)

3. **Test the model**. If it is below 10%, upload another document, and run the model to see if it accurately reads the new document. If it is between 10%-30%, try keyword spotting. If it's above 30%, run a new model using a different test set. Post a screenshot(s) of your test.

4. **Reflect.** Think about the work it took to get you to this point. Was it worth it? What are the implications of this work for future researchers?

Part II: Transcribing letters for computation

Required reading:
Smithsonian, Instructions for the Transcription Center, https://transcription.si.edu/instructions
Watch How to Use Transkribus in 10 Steps or Less: https://youtu.be/8Ei0a7WIlTI

Supplemental reading (not required, but useful for troubleshooting!):
Transkribus Wiki Page : https://transkribus.eu/wiki/index.php/Main_Page
"How to transcribe documents with Transkribus,"
https://transkribus.eu/wiki/images/5/50/How_To_Transcribe_Documents_with_Transkribus.pdf
"How To Train A Handwritten Text Recognition Model In Transkribus,"
https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf

Required software:
Adobe reader
Text editor (recommended: Notepad++)
Transkribus

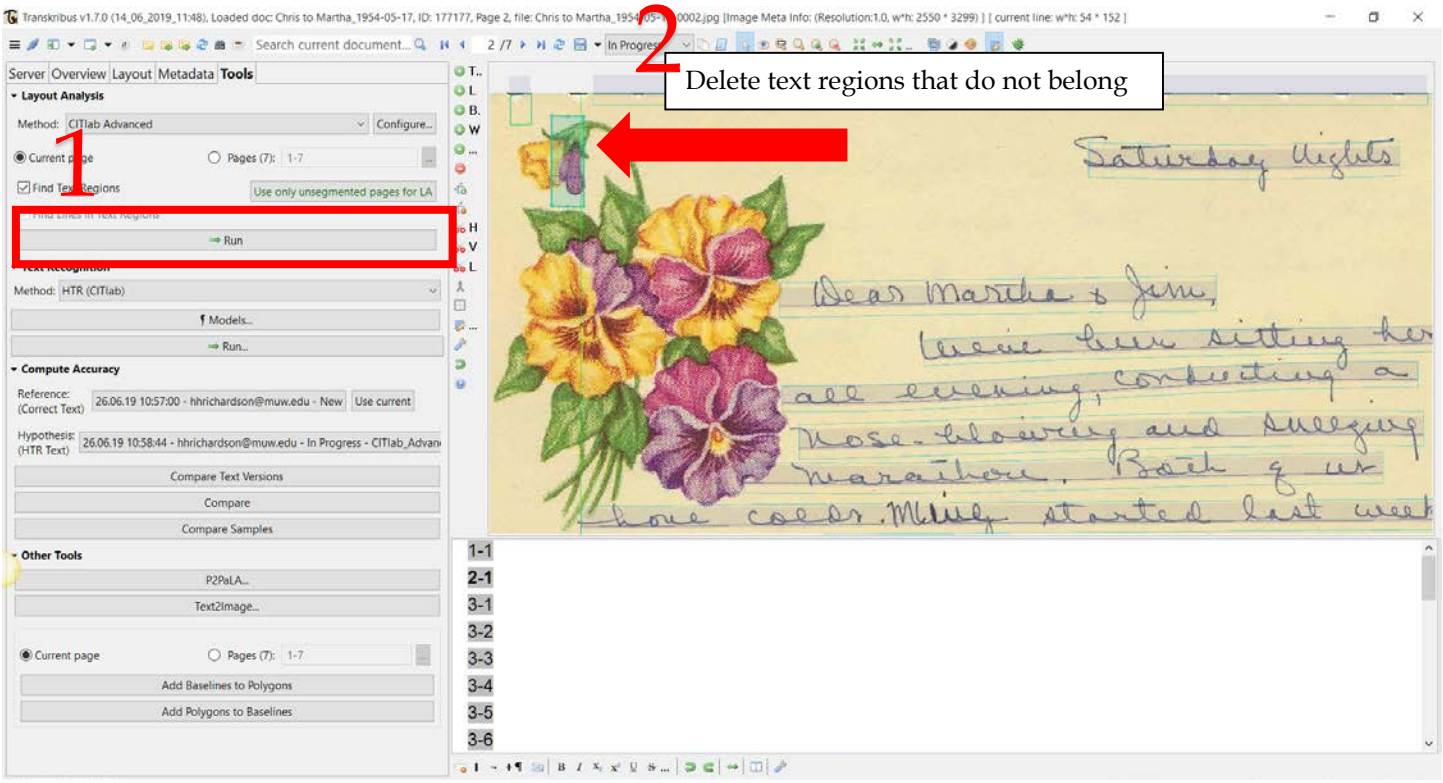
Assignment objectives:
This assignment will require you to think about the important aspects of an historical document, like a letter, and not just what's said within the message, but who is sending it, when, and where. In order to identify and highlight these important pieces, you will:

1. Transcribe the letter(s) – 50%
2. Create tags that serve as additional metadata for the contents of the letter(s) – 15%
3. Review transcripts for accuracy and editorial choices – 30%
4. Train the HTR model - 5%

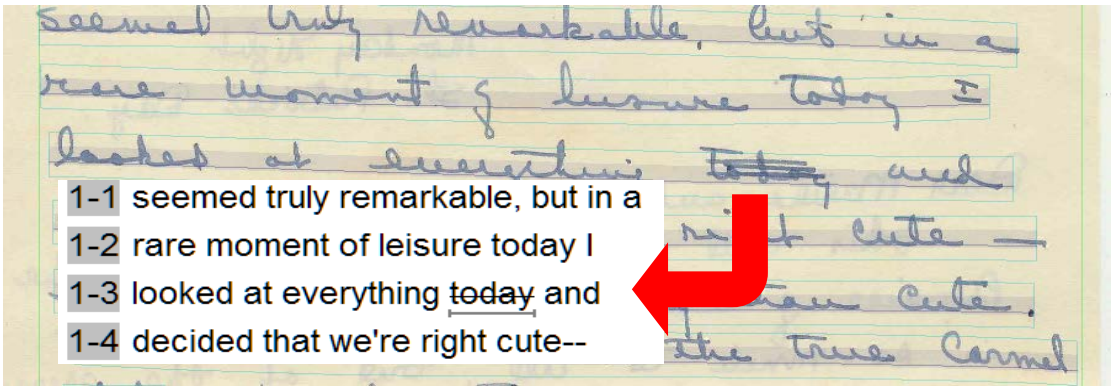
To submit the assignment: Export the document as both PDF + txt, using the filename template for the collection, and add to the course page/institutional repository. Upload your peer review worksheet to the canvas course.

1- Transcribe the Letter(s)

1. Read the transcription instructions from the Smithsonian and watch the How to Use Transkribus video (linked above)
2. Download Transkribus (Instructions for LIB 201 in the Canvas course page, and available in the Transkribus Wiki: https://transkribus.eu/wiki/index.php/Download_and_Installation)
    a. Upload your assigned letter(s) to the server for the collection, supplied by the instructor.
    b. For each page:
        i. Run the layout analysis, correcting any errors from the automation, i.e. omitted handwriting, objects that aren't handwriting selected, etc.

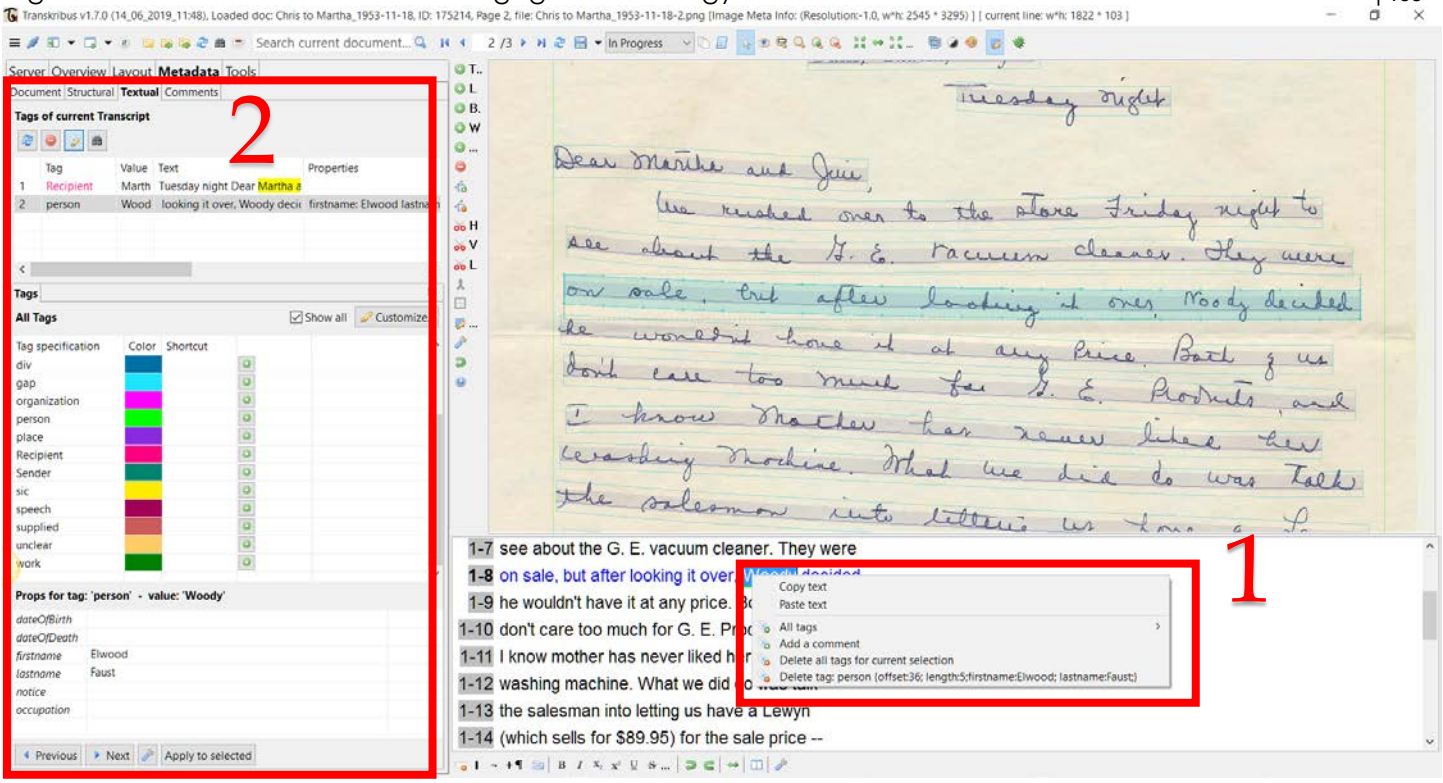


ii. Transcribe the page *exactly* as you see it, using the special characters and tags to indicate differences in text. For more examples of anomalies and inconsistencies, see the class' cheat sheet in the Canvas course page.



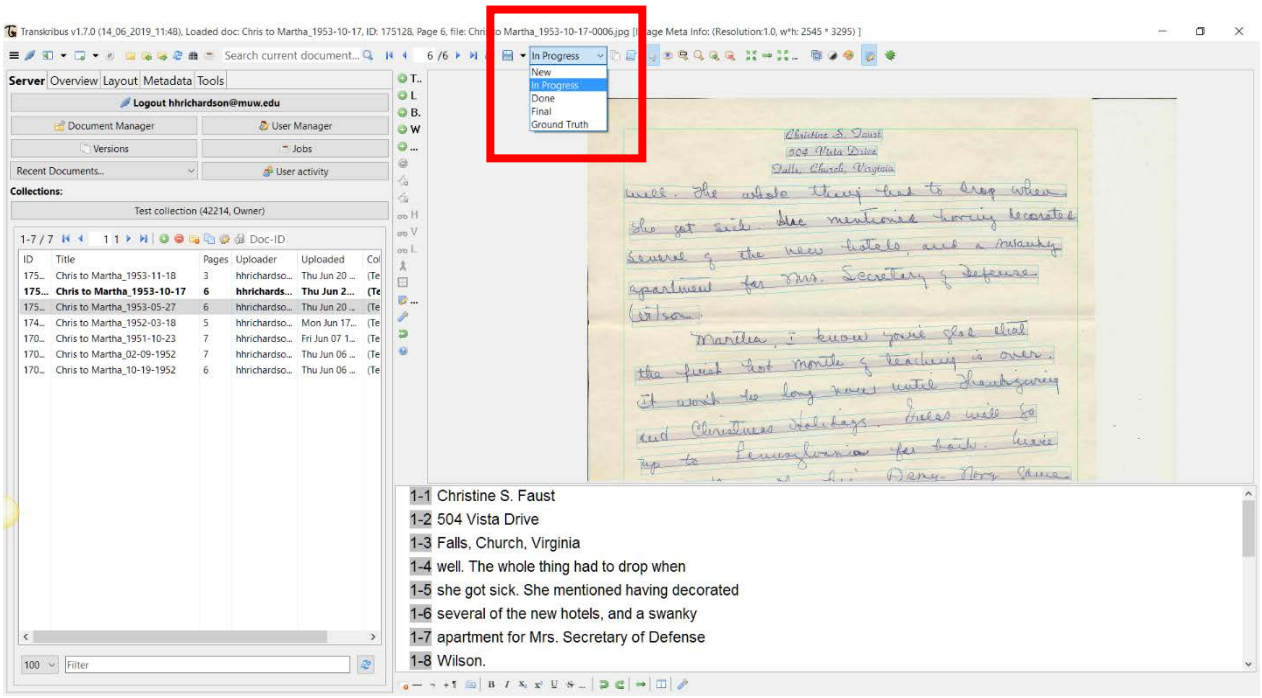
2- Create Tags

1. Using the contents of the letter, the context of the collection, and the metadata that has already been gathered at the beginning of the process, tag contents of the letter as appropriate.
2. If necessary, adjust the specifications, values, and properties of the tags.
3. Use the "supplied" tag for notes that you have added (i.e. [illegible], punctuation that isn't there, etc.) aside from what is on the page.



3- Complete a Peer Review for a classmate's transcription and tags

1. Using the rubric for peer review (appended), go over the layout, transcription, and metadata in your peer's letter(s). Pay attention to:
    a. Inclusion of handwriting
    b. Accuracy of the transcription
    c. Choices and properties of metadata tags
2. Change the document from "In Progress" to "Done"



4- Train the HTR Model

Note: since the technology for this relies heavily on volume of transcriptions, this will not be assessed by how accurately the model runs. Instead, you will assess the process you followed to train the model.

How to Train the HTR Model

1. Only the letters that have been completely transcribed and reviewed will be ready to include in the first instance of training the HTR model. Select these letters to add to the **Training Set**.
2. Within that group, choose 1-3 letters that are "representative of the documents" in the collection ("How To Train A Handwritten Text Recognition Model In Transkribus," https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf). These will be saved for the **Test Set**.
3. Run the model! This process can be the longest (a couple of days, in some cases), so it should be done at the end of class.
4. Assess the model - Because the development for this is also ongoing and experimental, so you will run the model, and then briefly discuss, in a class discussion post (appended), what made it successful or not, examining the efficiency, accuracy, and usefulness of the model.

Acknowledgements:

Developed using "The Santa Barbara Statement on Collections as Data," Version 2 (https://collectionsasdata.github.io/statement/)

With consultation from Sarah Ketchley, Digital Humanities Specialist for Gale Cengage's Digital Scholars Lab