Approach for Named Entity Extraction for OCR scanned Handwritten Slave Trade Volumes

Sushma Akoju
Independent Study, University of Colorado Boulder.
sushma.akoju@colorado.edu

1 Abstract Goal

As part of Independent study on OCR Scanned Handwritten Slave Trade Volumes from 19th century, I worked on understanding problem definition, understating prior research work, understanding current research works towards OCR scanned handwritten text recognition over structured documents, extracting text from OCR scanned handwritten documents, Legal documents and Named entity Recognition and finally connecting the problem with solution approaches from existing tools available in the market.

This is an Independent Study under supervision of Dr. Henry B Lovejoy, *Digital Slavery Research Lab, University of Colorado Boulder* and Kartikay Chadha, Doctoral Candidate, *School of Information Studies, McGill University & CEO, Walk With Web Inc*. The overseeing professor for this Independent Study course is Dr. Jane Wall, Faculty Director, University of Colorado Boulder.

The specific goal of this study is to bookmark pages with event matches and extract the event information from the recovered collections from volumes from African diaspora. The Slave Trade Volumes are OCR scanned Handwritten documents. Thus this requires an approach is multi-disciplinary intersection.

2 Background and Motivation

The novel goal to provide insights into Slave Trade history, to recognize people for their identity and dignity from the historical biographies and events has been the important motive to most researchers and historians from Digital Humanities over study of people in Slavery Lovejoy and Chadha (2021). Recent efforts in digitizing the content from historical documents to represent using modern Ontological techniques and make them available and searchable online are a common motivation. The papers' summaries provided insights into various aspects of research on this common motivation. The Lovejoy (2020) provided detailed accounts of efforts, with timelines for manual transcribing efforts and digitizing the historical documents along with motivations and contributions. The Bell and Ranade (2015) provided various challenges in the text corpuses from Historical digital documents and the techniques used to address the problems from historical data analysis. The Schindling (2020) provided special focus to normalizing basewords and spelling variants for text analysis. The Lovejoy and Chadha (2021) provided influential works from the project on Gustavus Vassa, which identifies individuals who were part of struggle in Slavery, as people with identities and giving prominence to their contributions, biographical details by providing search space to recognize dignity and personality to the individuals. The papers summarised in this initial report provides perspectives together with strong motivation towards exploring solutions to make these documents searchable for various historians, interested researchers on the subject. It seems that the novel goal is that of not only emancipating Africans from historical events and providing valuable insights, but also seems to indicate a shift to emancipate from the present day impressions of the past.

To achieve this goal, we initially completed a First report from May 2022. https://drive.google.com/file/d/1PacH5--jOHW64n5-HLtrmB8ClqRlvXEU/view?usp=sharing. The following paper consists of work from after the First report to until end of Summer term, i.e. Aug3rd 2022.

3 Understanding the Slave Trade Volumes towards Goal of Named Entity Recognition

I worked on understanding what is the general document structure, the Volumes, collections and documents. There is also a handwritten tabular section known as List of Papers which serves as an index. The goal to develop an index such that we have search space that learns the document title within each collection to its corresponding page number was the first task. I first explored, as per suggested solution suggest by Dr. Henry B Lovejoy and Kartikay Chadha using Transkribus Kahle et al. (2017).

3.1 Document Structure, Layout Recognition and Layout Segmentation as an Image Processing Task

Using Transkribus, it is possible to annotate OCR scanned Handwritten PDF documents for tabular segments.

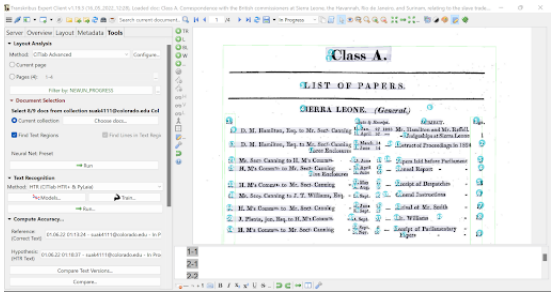

Figure 1

Figure 2: Slave Trade Volume 10: Custom Annotations using Transkribus Kahle et al. (2017)

I learnt that for Document Segmentation, following aspects hold true:

1. Annotating regions of interest into blocks
2. Annotating the Sequence of blocks already annotated
3. Making sure sequence of blocks is either vertically oriented (such as title pages that are top to bottom) or horizontally oriented (left to right such as text documents, tabular indices)
4. Layout Segmentation, Layout recognition are additional two different subtasks.
5. There is a Layout Parser Shen et al. (2021) that can parse a layout which includes segmenting all possible layouts and finally recognizing the layouts of interest from segmentations.
6. There are several novel approaches implemented towards document structure recognition involving Layout parsing and segmentation. It is a more of a research intensive problem with an intersection of Image Processing and Layout Recognition from Regions of Interest (ROI).
7. For Document structure recognition task, we consider segments of document that consist of text within well-defined blocks, that has boundaries, outlines. Examples of types of Layout segments are, Title, Paragraph, subheading - each of these options form a layout segment by themselves.

I worked on understanding the pipeline Toledo et al. (2019) that used KALDI OCR text extraction software using Convolutional Neural Network for detecting a text in a single line from multi-level knowledge extraction using Bidirectional Long-short term memory networks:

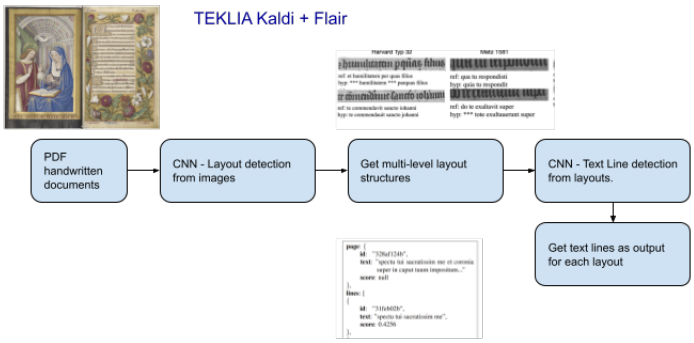

Figure 3

Figure 4: OCR scanned Historical Handwritten Baptism registers: Pipeline for multi-level Knowledge extraction using Bi-LSTM and CNN Toledo et al. (2019)

I worked on replicating Custom Layout Parser using Detectron2 which is used for Object Detection and Segmentation Wu et al. (2019). The results from Shen et al. (2021) are as follows for List of Papers Indices:

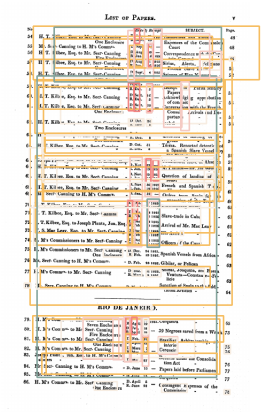

Figure 5: With faster rcnn R 50 FPN 3x

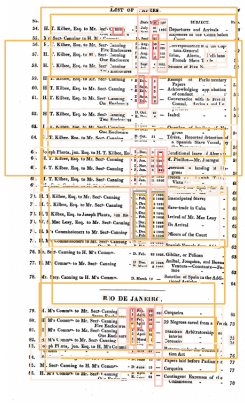

Figure 6: With mask rcnn R 50 FPN 3x

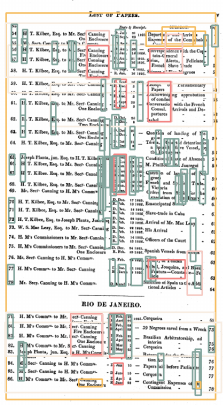

Figure 7: With retinanet R 50 FPN 3x

Figure 8: Class A Collection from Slave Trade Volume 10

The aforementioned results suggest, that block-level annotations, segment annotations, sequence of each of blocks and text segments together with significant Data augmentation such as rotated images can better provide sufficient information from handwritten table recognition from images. Thus this task of running this example suggests this task requires advanced Computing resources. Existing/current research works towards OCR scanned handwritted documents and Text extraction suggests Rodriquez et al. (2012) re-scaling the OCR scanned handwritten images to atleast 4 times that of original size using anti-aliasing filter can enhance the OCR recognition towards word level accuracy. (The notebook referring to Rodriquez et al. (2012) `https://github.com/ocropus/ocropus4/blob/main/demo.ipynb` is an example implementaion using a tool similar to that of Transkribus Kahle et al. (2017). The OCR noise at word-level accuracy has been a common problem and has been generally addressed using spell correction techniques. Toledo et al. (2019).

Learning from this approach: It is possible to combine Layout Detection, Layout Segmentation, Multi-level structured Segmentation and Text Extraction together to extract text as output without loss of structure at all levels from a given a PDF document page as an image input.

Implementation of Layout Parser: `https://drive.google.com/file/d/1TmAo51JnU2oCKTulZAgS_P-_oh5NSF2V/view?usp=sharing`

3.2 Examining current research works on Named Entity Extraction for Historical OCR scanned documents

To study and understand ongoing research works on Named Entity Extraction for Historical OCR Scanned documents, I studied following research works which seemed relevant to Slave Trade Document analysis:

1. Menini et al. (2018) and Palmero Aprosio et al. (2017) suggest Named Entity Recognition is the starting point of constructing Knowledge representations over Historical documents. The suggested works used Coreference Resolution along side word sense disambiguation for finding context-sensitive semantic links between entities. Their research work used BIO (Begin, Inside, Outside) tags for Chunking towards Named Entity Recognition (NER) tasks.

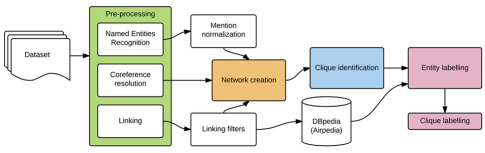

Figure 9: An pipeline of Entity linking projects

2. Ehrmann et al. (2020) which is primarily a Named Entity Processing on Historical Newspapers called as Impresso Project. The research work uses one-hot encoding for Named entity annotations. Additionally they used Semantic Segmentation for Text data to *select* a paragraph or section of text from a document, for the searched Named Entity over an OCR scanned document. Some of the key takeaways from this research work, which seemed to have higher relevance are:

   (a) *Semantic Segmentation that connects the Named Entity to the paragraph from text and the section form Image for knowledge representation* towards the complex task of OCR scanned Handwritten documents.

   (b) They use metonymy (substituting a words such as "figure of speech") as well as replacing words with custom base words (which is a technique similar to the one used in Schindling (2020)).

   (c) They used Entity Linking with two variations: strict boundary matching versus fuzzy boundary matching (which is similar to Bell and Ranade (2015)).

   (d) This work created custom Word Embeddings using `https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_9_TRAINING_LM_EMBEDDINGS.md` but by using context of 400-600 characters, the custom FLAIR embeddings can be created. *This helps with better Named Entity Recognition with Custom text corpus such as Slave Trade Volumes.*

   (e) This project's work included challenges w.r.t Historical knowledge to annotate the data. Thus *this helps to identify and define skills required to "annotate" the data correctly.*

   (f) This project's research also conducted a correlation between article publication date and performance of the Language Models and Pipeline for the NER and Entity Linking tasks but found no correlation between the two.

   (g) This project found good performance at medium OCR noise, while the extreme noise levels (low vs high noises) in text data did not seem to perform as expected. The low noise in Text data is after manual annotations and corrections were conducted over text extracted from OCR text extraction tools.

   (h) *The project's research work concludes that the Semantic segmentation, annotation guidelines, custom NERC, Custom Entity Linking tasks are a step towards efficient Semantic indexing of Historical material.*

   The pipeline from Impresso project is as follows:

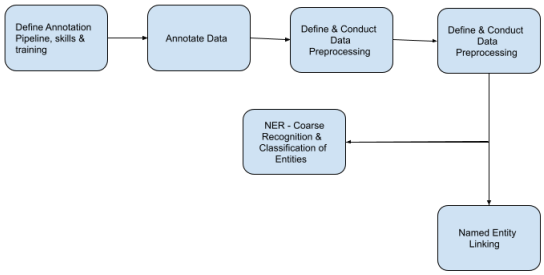

Figure 10: An abstract pipeline of Impresso project

3.3 Analysis from reading a collection from Slave Trade Volume 10

As a next natural step after studying about ongoing research projects on OCR scanned Historical papers, I read the Class A collection documents from Slave Trade Volume 10. The findings from this reading are as follows:

1. They each are court proceedings, custom document structure and follow Legal terminology.

2. The nomenclature is similar between the two categories.

3. Court case parties, judges,names of courts, case numbers, references to laws - these broader concepts remain common between British courts irrespective of the changes in formats of documents from the 1800s vs 2000s.

4. Handwritten, OCR scanned Legal documents of court proceedings.

3.4 Examining Named Entity Recognition over Legal Entities

The analysis and findings from studying existing research works towards Named Entity Recognition over Legal documents is as following:

1. The Legal Entity Recognition (LER) Leitner et al. (2019) LER acknowledges documents requiring transliteration, OCR errors and spelling variations. Place names are generally complex Named Entities for NER tasks over Historical Document.

2. The study of British Parliamentary papers from 17th century to until 19th century Grover et al. (2008) suggest spell correction using Machine learning methods.

3. Rodriquez et al. (2012) compare OCR extracted text w.r.t OCR extracted and corrected text and find that it seems models need spelling correction to recognize for unseen text even though NER tasks seems to give similar performance between uncorrected versus corrected OCR extracted text.

4. (Leitner, 2019) suggest building dictionary lookup and rule-based pattern recognition as a common method for Legal NER tasks.

5. Ehrmann et al. (2021) suggest Bidirectional Long-short term memory and BERT models recognize entities from OCR text correctly without spell correction as well as better accuracy over spelling variations over spelling changes over Historical documents.

4 Approaches Implemented

I conducted three types of analysis. First analysis over manually extracted text data from Slave Trade Volumes and second analysis by extracting text using Google Document AI Cui et al. (2021). Third analysis over Cases data manually extracted which serves as Ground Truth Named Entities i.e. data manually created by Henry's team.

4.1 Analysis over Manually Extracted text data

I received manually extracted text data from Slave Trade Volumes from Walk With Web team. Following are the analysis from manually extracted text:

1. High OCR errors and noise.

2. Separating document text between known document start and end tags such as "No. XXX" was not accurate since the document structure was not preserved.

3. There are missing documents in each collection. This can be attributed to encoding in Text data between Portable Document Format (PDF) to Microsoft Word along with text encoding in "Copy" command.

4. The document order was incorrect and merged documents was not correct. Two or more documents' text was overlapped. Hence document boundaries was not preserved.

5. Running Parts-of-speech over this extracted data was very poor. No tags were detected except propositions, some Proper Names. NeuSpell (commonly used Spelling correction tool) and Word Embeddings for a small paragraph did not yield any accuracy.

6. Slave Trade Volumes consist of Legal entity words from 19th Century language context and additional OCR noise which require spelling correction.

4.2 Analysis over Extracted data using Google Document AI Cui et al. (2021)

Google Document AI Cui et al. (2021) is an Application Programming Interface (API) from Google that can detect structured text data and segments from various types of PDF documents.

I implemented following steps to extract Text data over Slave Trade Volumes 10 & 11 using Google Document AI API:

1. Split each Collection pdf from each of Slave Trade volumes into 10 page partitions.

2. For each document, authenticate Document AI token, get results from document.

3. Each document result consists of Image block with location info (x,y pairs) and paragraphs from each page indexed by page number.

The Google Document AI results are as follows:

Figure 11: An sample output from Google Document AI API

Implemented works are as follows:

1. I implemented Parts of Speech tagging using SpACy software. First I had to conduct several preprocessing steps to bring all of documents data of each collection and combines all volumes by creating separate Identifiers.
   (a) Each collection is partitioned into 17-20 10-page partitions
   (b) For each volume, there are 1) Collections and 2) List of Papers.
   (c) For each collection, there are JSON responses, Blocks & Paragraphs results
2. Some of the observations from Google Document AI to make a note of are as follows:
   (a) The sequence of block recognition is the conventional left-to-right similar to Western form of reading a book. Transkribus supports custom sequences.
   (b)
3. The full data generated, cleaned, organized, merged at various stages is available at https://drive.google.com/drive/folders/1HX_Mo_nxYsV63wZb-eiRztQBttFnOiiZ?usp=sharing
4. JSON, Paragraph-wise and block-wise data are as follows:

Figure 12: JSON response format that was saved for each partition of a collection.

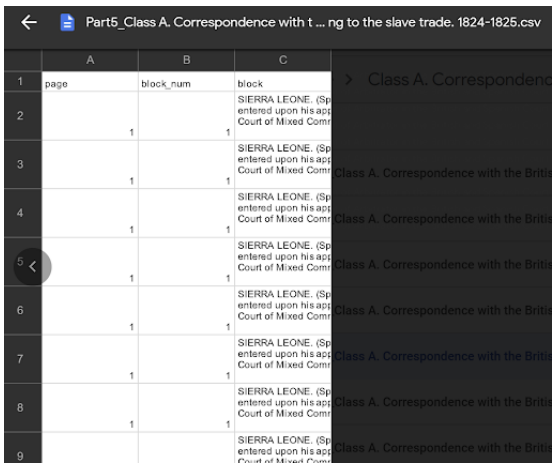

Figure 13: Block-wise response format that was saved for each partition of a collection.

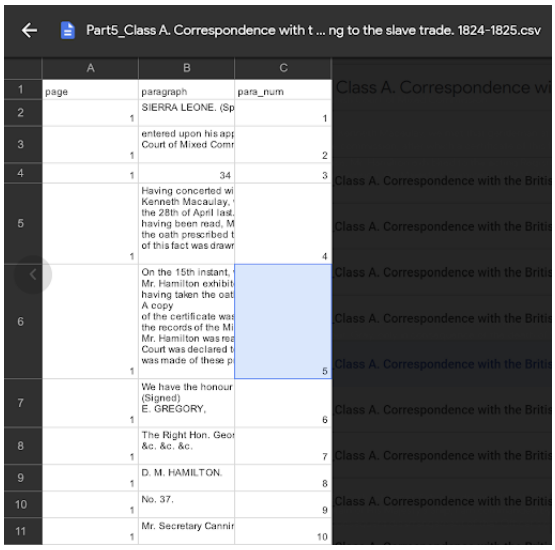

Figure 14: Paragraph-wise response format that was saved for each partition of a collection.

5. POS Tags for all volumes combined together each identified by the unique identifier as follows: volumeNum_collectionName_TruePageNumber_ParagraphNum.
6. All volumes with POS tags at word level are available at: https://drive.google.com/file/d/1Hk7xCQdWIwyZaze96W5s0ze1neUsG6FV/view?usp=sharing

	sent_id	word	tag	short_id
896934	34040	the	DT	11_Class A_1_1
896935	34040	Island	NNP	11_Class A_1_1
896936	34040	of	IN	11_Class A_1_1
896937	34040	Cuba	NNP	11_Class A_1_1
896938	34040	.	.	11_Class A_1_1

Figure 15: An sample output from Parts of speech tags with unique identifier (short_id column) for Slave Trade Volumes

For the third analysis over Ground Truth data of the entities and events data over Slave Trade Volumes, I conducted following analysis:

1. Selected columns Case_Name, Year, Arrival_Region, Arrival_Port, Court_Type, Court_Name.

2. For each column generated parts of speech tags. POS tags are as follows:

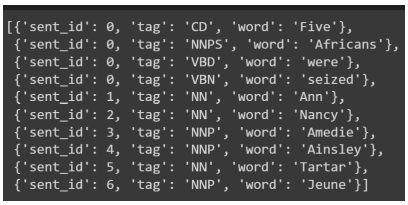

Figure 16: An sample output from Parts of speech tags for Ground Truth Entities.

3. For each of combined entity name: for example for a case name such as "Five Africans were seized", the entity POS tag sequence pattern "CB NNPS VB VBN". This will serve as a rule pattern to "chunk" the data into IOB (Inside, Outside and Begin) tags. Another example parts of speech tag for "San Joaquim" the tag sequence is "NNP NNP". Combining all of unique POS tag sequences, we could analyze the tag sequences and select rule patterns for chunking the Entities from Slave Trade Volumes.

4. Thus for the selected columns , the POS tag sequences are as follows:

   (a) Case_Name : there 192 unique POS tag sequences

   (b) Year : there 1 unique POS tag sequences i.e. "CD" i.e. Cardinal Digit.

   (c) Arrival_Region : there 6 unique POS tag sequences

   (d) Arrival_Port: there 16 unique POS tag sequences

   (e) Court_Type: there 14 unique POS tag sequences

   (f) Court_Name: there 31 unique POS tag sequences

   (g)

5. Combining all of the columns' POS tag sequences, the data is stored as follows:

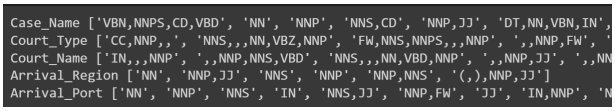

Figure 17: An sample output from Parts of speech tags for Ground Truth Entities.

4.3 Solution Recommendation

Using the aforementioned information, analysis and generated data, the proposed steps are:

1. Using the POS tag sequences, filter the rule patterns.

2. Using the filtered rule patterns, generate Chunking for IOB tags.

3. Using the IOB tags, for each entity names in the Paragraphs file, match the entity names and mark start and end indices.

4. Conduct NER training on Slave Trade Volumes 10 and 11.

5. conduct quality analysis, GLUE and F1, precision classification scores over Test data results. We consider Slave Trade Volume 12 as test data.

Google Document AI Configuration guidelines are as follows:

1. Introduction: https://cloud.google.com/document-ai

2. Configuration : https://cloud.google.com/document-ai/docs/setup

3. OCR Document API details: https://cloud.google.com/document-ai/docs/processors-list#processor_doc-ocr

4. Getting started with Client libraries in Python https://cloud.google.com/document-ai/docs/process-documents-client-libraries

5. How-to guides https://cloud.google.com/document-ai/docs/how-to

5 Artifacts, Datasets & Presentations

1. All Jupyter Notebooks: https://drive.google.com/drive/folders/1ov_kCjDrUUwwfEzZ66EOscnX23rESxdm?usp=sharing

2. All datasets for Inputs and the results generated by API and processed output for future use: https://drive.google.com/drive/folders/1HX_Mo_nxYsV63wZb-eiRztQBttFnOiiZ?usp=sharing

3. All Powerpoints: https://drive.google.com/drive/folders/18BPPLDhe3ApJEkIH3YI9-2GKU6DjLVo5?usp=sharing

4. Uploaded Jupyter Notebooks to Github repository: https://github.com/walkwithweb/Independent_Project-S-Akoju

6 Additional work

(Moguillansky et al., 2019)

7 Conclusion

In conclusion, as part of this study, we could see various complex layers that existed with just OCR data and text extraction itself. This requires multi-disciplinary approach that includes Layout Segmentation, Layout Recognition, OCR text recognition, NER task pipeline together. This is unique data set with diversified, but multiple smaller and hence simpler subtasks. The insights from understanding the data itself, could help to breakdown the complexity into. With such vast amounts of research and contemporary Neural methods, it is possible to develop a proof-of-concept by first extracting data from creating new Segmentation models, then extracting text using Google document AI or Transkribus, then by automating annotation process using Ground Truth and then using annotated data to create NER model. Thus using NER model and analyzing the results from GLUE and standard benchmarks, we can analyze the quality of data and assess necessary steps to further improve this approach.

Acknowledgments

Thank you so much to Dr. Jane Wall, Dr. Henry B Lovejoy, Kartikay Chadha and Walk With Web for the opportunity to learn, study, implement as well as for the support, encouragement and guidance. https://www.colorado.edu/lab/dsrl/collaborators

References

Mark Bell and Sonia Ranade. 2015. Traces through time: a case-study of applying statistical methods to refine algorithms for linking biographical data. In *BD*, pages 24–32.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *arXiv preprint arXiv:2109.11406*.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In *CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, volume 2696, page 38, Thessaloniki, Greece. CEUR-WS.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE.

Elena Leitner. 2019. Eigennamen- und zitaterkennung in rechtstexten. Master's thesis, Universität Potsdam, Potsdam, 2.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany. Springer. 10/11 September 2019.

Henry B Lovejoy. 2020. Who did what when? acknowledging collaborative contributions in digital history projects. *Esclavages & Post-esclavages. Slaveries & Post-Slaveries*, (3).

Paul E Lovejoy and Kartikay Chadha. 2021. Equiano's world: Chronicling the life and times of gustavus vassa. *Esclavages & Post-esclavages. Slaveries & Post-Slaveries*, (4).

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Martin O Moguillansky, Antonino Rotolo, and Guillermo R Simari. 2019. Hypotheses and their dynamics in legal argumentation. *Expert Systems with Applications*, 129:37–55.

Alessio Palmero Aprosio, Sara Tonelli, Stefano Menini, and Giovanni Moretti. 2017. Using semantic linking to understand persons' networks extracted from text. *Frontiers in Digital Humanities*, 4:22.

Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw ocr text. In *Konvens*, pages 410–414.

James P Schindling. 2020. *The Spatial Historian: Creating a Spatially Aware Historical Research System*. West Virginia University.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*.

J Ignacio Toledo, Manuel Carbonell, Alicia Fornés, and Josep Lladós. 2019. Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition*, 86:27–36.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.