

# First Summary Report

Sushma Akoju

Independent Study, University of Colorado Boulder.

`sushma.akoju@colorado.edu`

June 4th, 2022

## 1 Abstract Goal

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 following goals were described and defined by supervisors. We define two goals, as described per the area of research. 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. We extract events from document volumes based on event information already collected in the dataset, then bookmark the page numbers from the volumes to each of the corresponding events. We repeat this task over one year's event-volume pairs of information. This task includes two types of searches: boat name vs people's names. The other search terms could be start year, by location and various terms described in a manually generated event details dataset. Given a search term or set of search terms (is a row of event details in text format), search the terms in documents and extract and auto-fill an inventory table with explanations. We call this entire goal as Henry's problem definition.

## 2 Introduction

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 Recent efforts in digitizing the content from historical documents to represent using modern Ontological techniques and make them available and searchable online are a common motivation.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2020) provided detailed accounts of efforts, with timelines for manual transcribing efforts and digitizing the historical documents along with motivations and contributions.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.

The studies in Slavery document analysis in Digital Humanities seems to include a vast amounts of manual efforts from transcribing, gathering, analyzing data which was conducted by Historians and researchers for long from around the world. The main goal of this study is that we extract events from document volumes based on event information already collected in the dataset, then bookmark the page numbers from the volumes to each of the corresponding events. We repeat this task over one year's event-volume pairs of information. This task includes types of searches based on boat name and on people's names. The other search terms could be start year, by location and various terms described in a manually generated event details dataset by

Henry. Given a search term or set of search terms (is a row of event details in text format), search the terms in documents and extract and auto-fill an inventory table with explanations. We attempt to find solution for extract text and layout information from historical handwritten Optical Character recognition (OCR) scanned Portable Document Format (PDF) documents to first build indices while preserving document structure to extract. We then provide this as input to solution that could address the Henry’s Problem Definition.

### 3 Background

The slavery trade volumes were collected from 1800s which are British Parliamentary papers. The volumes serve as records of Africans from Havana, Sierra Leone, Great Britain, Cuba, Caribbean Islands etc. Each of the volumes is grouped by People, Events, Places and Sources and are uploaded to CRL Digital Delivery System. Collections are grouped into each of the volumes by year. Each of the collections is a scanned PDF document. Each document is associated with Person or event or place or source or a court proceeding is grouped into a collection. Each collection follows a document Structure. Pages of interest for the main goal are: Title page, Table of Contents, List of Documents/papers and content files. We describe further in Section 5 with data format and the details of each of the structure of document and volumes. Example: <http://ddsnext.crl.edu/titles/33509/items?terms=&page=0> (Lovejoy, 2020).

## 4 Paper Summaries

### 4.1 Who Did What When? Lovejoy (2020)

1969 - Philip Curtin first raised questions about 1) when and where people came from in Africa and when and 2) where people went in the transatlantic slave trade.

1970s - Herbert S. Klein, David Eltis, and others, collected data on slave trading voyages. David Eltis began compiling information systematically into SPSS and created “African Names Database”, which accounted for 68,000 Liberated Africans.

1999 - The Trans-Atlantic Slave Trade: A Database on CD-ROM which covered 27,000 voyages, worked by Eltis and Behrendt.

2002 - African Names Database expanded to 90,000 Liberated Africans, which became the basis for African Origins.

2008 - Voyages: The Trans-Atlantic Slave Trade Database with more data and released online.

2009 - Henry Lovejoy transcribed registers from Cuba and Caribbean which were updated to African Origins database.

2013 - identified and digitized by Suzanne Schwarz and Paul Lovejoy and later joined by many others which led to recording Liberated Africans disembarking at Freetown from England and Sierra Leone.

2015 - 36,000 voyages were added to the data.

2018 - Eltis's estimates updated to 12.5 million who were forced to leave the African coast. But only 10.7 million reached the Americas. Also Digital Microfilm Project for Liberated Africans.

Henry Lovejoy digitized materials located in Barbados, Brazil, Cuba, Curaçao, France, the Netherlands, Spain, Suriname, and Great Britain between 2001 and 2014 Database and Relational Schema

Names Database Schema: Names with roles (Liberated Africans). Each Name connected to an event which has a case (series of events). Each event has 1) object which is a resource and 2) place (court).

Each box in the above diagram represents a specific spreadsheet of available data obtained from HTML coded design of Liberated Africans. The core-team of historians initially had difficulty understanding and accepting this proposed concept, which deviated from the original website structure based on cases mostly involving voyages. The relational database structure emerged into four key spreadsheets:

1. **People**, which hosts definitive records for each person with a unique ID, whether registered or not. Content accommodates variations of data obtained from the Registers of Liberated Africans.
2. **Events**, which includes pivotal information about the date an activity happened to an individual. At present, these data mostly revolve around slave voyages and trial proceedings, such as embarkation, capture, disembarkation, trial, registration, emancipation, etc. In the future, pre- and post-trial events can be added to link in evidence

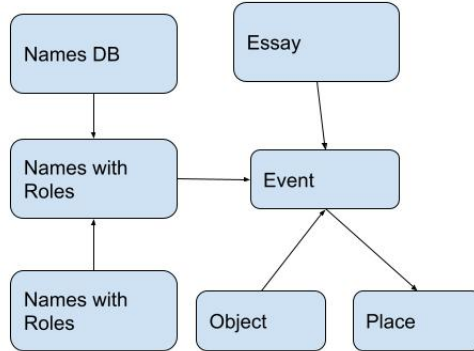


Figure 1: Database Schema

of: baptisms, military service, post-trial resettlement, education, re-enslavement, second emancipation, marriage, birth of children, death, etc.

3. **Places**, which involves geographic coordinates that are associated or connected with people and events. These data have implications for mapping, visualizations and analysis over space and time.
4. **Sources**, which reflects the digital archive of primary source, multimedia objects, such as PDFs or JPGs. All source metadata adheres to DublinCore standards and any published materials must abide by copyright law. Since all data derive from these digitized materials, each object links into the people, events and places datasets; thus, access to the primary source provides users the capability to verify data accuracy. <https://enslaved.org/>

## 4.2 Traces Through Time Bell and Ranade (2015)

**Summary:**

This paper discusses the algorithmic approaches to perform matching records based on a person's name and age or date of birth and summarizes following concepts:

1. Record Linkage finds spelling differences between pairs of textual attributes. The basic approach is to use the ratio of probability that a pair of records refer to the same person and that of two different people. For this comparison, there are string comparison algorithms. For a fine grained approach, the paper uses fuzzy string matching.
2. For the matching algorithm, the paper discusses the Fuzzy matching algorithm. I worked on implementing and understanding algorithm for fuzzy string matching over sequence of words: <https://www.datacamp.com/tutorial/fuzzy-string-python>.

The records' attributes and approaches discussed in this paper:

1. The date of birth, age can be available, but for historical data which is handwritten, this might be problematic. Due to this reason, the authors used new techniques to get confidence scores for age, date of birth based on estimated distributions to account for error from scanned handwritten recognition tasks.
2. For the year attribute, the authors first created a probability distribution of likely values, then used a fuzzy match between year values, adjusted data quality and then derived the probability that underlying values are the same. In the absence of birth period or year values, authors derived a frequency distribution for the whole dataset, probability that person is in dataset A can be in dataset B.
3. An example for points 2 and 1 is: among soldiers in World War 1, age groups of 16-20 were more likely to have 2 birth years that would refer to the same person when compared to the 30-35 years of age group. Thus it seems like pairwise comparisons between each pair of ages which had two years mentioned that belonged to age group of 16-20, would have higher confidence level compared to two different years from 30-35 age groups.

Common problems in data where the approaches to fix them were discussed from the paper:

1. Factors such as: handwriting recognition errors, typographical errors and phonetic errors made when names are recorded as well as
2. Inaccuracies in recording such as mis-representation of age or rounding of declared ages - approach suggested is to the age distribution observed for each dataset is fed back into the algorithm to support a statistical approach to calculating the likelihood that two occurrences of a person with different recorded dates of birth, in fact, relate to the same individual. As each incremental enhancement of the algorithm improves the results of the matching process, these in turn, reveal further discrepancies in the data, from which the algorithm can learn.

Fuzzy Name comparisons: Name transcriptions for records representing the same person can be thought of as a function of several factors:

1. Regional spelling variations.
2. How the recorder hears the name, particularly with unfamiliar names and regional accents.
3. The recording medium such as – handwritten vs. typed.
4. Involuntary errors during data capture, spelling mistakes while writing or typing the original document. Involuntary errors during transcription, including those caused by difficult handwriting. Solution: Simply informing the model of the probability that Ts and Js have been interchanged has delivered good results, so the next step authors used was to use the data to identify a wider range of commonly occurring transcription errors. Jaro-Winkler measure to find similar strings and weightings to assign confidence depending on this measure.
5. Various types of text analytics techniques were broadly used in this paper and discussed.

Name frequency statistics:

” ...It can be difficult to accurately calculate the frequency of occurrence of a particular name in the population that appears in the records under consideration. Consider a dataset with 1,000 records including surnames such as Messrs Taylor and Zephania. From this alone, we could assess that

these surnames have an equal probability of 0.001 while, in reality, the former is more common and the latter is rare. Accumulating match results over multiple datasets allows us to create a larger population of individuals from which to derive probabilities. By clustering forenames and surnames together we have identified groups of names that typically occur together and which appear to align with national or ethnic groups - e.g. Irish, Italian, Hispanic/Portuguese...” (Bell and Ranade, 2015)

This paper thus provides an exhaustive number of inconsistencies in numerical as well as textual data such as names, naming conventions, root words which would vastly depend on culture, language evolution, location and age groups. Additionally the paper also developed exhaustive number of techniques for pattern recognition, pattern correction, error estimation, error adjustment to account for various handwritten recognized documents.

### 4.3 The Spatial Historian Schindling (2020)

#### Summary:

1. This is the thesis defense that integrates the tasks of information extraction, data management, and analysis while simultaneously giving primary emphasis to supporting the spatial and humanistic analysis and interpretation of the data contents.
2. The paper describes about Content Extraction with Synonym matching.
3. Paper gave several examples and approaches for contextualizing and normalizing the names. Names such as 1) “Fernández”, “Fernandez” and “Fernádes”, and 2) Alonso was written “Alo” “Franco” and Francisco was written “.
4. The lists are made up of basewords, which are the standard versions of words that are the names (proper nouns), and their associated spelling variants and abbreviations.
5. Abbreviations analysis: Franco is abbreviation of Francisco and the baseword entry is Francisco and variant spelled as Franco.



6. Normalized text is maintained instead of original text. For example, with baseword format, any variant is always replaced with its corresponding baseword. So normalized text will always contain the baseword instead of variant, if the original text consisted of the variant.
7. It is not clear if the paper used a Deep learning model to extract information such as names, events etc even though text preprocessing considers modern approaches such as basewords.
8. Most importantly, the paper seems to have extracted all names using information extraction tools, classify as baseword and variants and then used baseword to normalize the name values.
9. For our purpose of main goal, having basewords and spelling variants is certainly helpful to recognize and develop search based on basewords, variants likewise for same document with variation in ranking result. For example, we can rank 10 documents where a text such as Franco to display both baseword matches as well as spelling variants.

#### **4.4 Equiano’s World Lovejoy and Chadha (2021)**

##### **Summary:**

1. This paper is about introducing Equiano’s World. The project on Gustavus Vassa (Olaudah Equiano) focuses on the movement to abolish the trans-Atlantic slave trade and ultimately to emancipate the Africans and their descendants who had been enslaved. The paper conveys the very perceptive story of the boy who later influenced the development of a website which makes his digital identity globally accessible, from the slave narrative, his historical significance and his continuing moral influence.
2. As quoted in the paper, ”...Equiano’s World is organized in a manner that allows users to access specific information, including a list of the various editions of The Interesting Narrative, a bibliography of publications about Vassa and his work, materials that can be useful for teaching purposes, background information on his associates, and a list of those who subscribed to the different editions of his autobiography. The different sections of the website are labeled “Context,” “Travels

of Vassa,” “Associates of Vassa,” “Questioning Equiano,” “Studying Equiano,” and “Resources”” Lovejoy and Chadha, 2021.

## 4.5 The enslaved ontology Shimizu et al. (2020)

### Summary:

The paper asks some compelling questions, which serve as good motivation:

1. How can we more effectively answer important moral questions?
2. How can we make those questions part of a broader public discourse? What sources are available? How can we give broad access to them?
3. And how in the decades to come will scholars answer questions about black bondage and its legacies when much valuable source material is deteriorating due to inattention, siloed scholarly activities, and under-funded archives?

This paper discusses attempts to develop a generalized model and ontology for the Slave Trade volumes:

1. They follow a specific Ontology modeling approach which is based on Use case based on Competency questions, answers while deciding which content can be more important for developing a model, use case etc.
2. Example competency questions are ”who did Thomas Jefferson enslave at Monticello?”, ”In what records does enslaved person named XXXX appear?”.
3. The paper used Web Ontology Language OWL as the language and formalized axioms required to express relationships broadly by emphasizing that history is related to geographical location, cultural heritage, and multimedia.
4. The paper explicitly restricted to knowledge representation using Semantic Web Ontology construction approach instead of deductive or inductive reasoning based on Formal Logic definitions, which is more commonly used in organizing web of information such as Google search.

5. As noted the paper did not seem to be well generalized or adaptable to various datasets.

Thus paper stands as a good example of developing Knowledge base for representing Semantic Web and developing an Ontology towards Slavery Trade documents.

## 5 Data and Method

In this section, a more elaborate information of definitions of the data structure, the content, language specific analysis is provided. We recognize the data for this project is historical data and thus could consist of language patterns, grammar which are a couple of centuries ago.

### 5.1 Data and Format

We first define the data structure objects in a Top down approach starting from Volumes to all the way to List of documents or contents of documents.

**Volume** : is the number of exported slaves, categorized by year, by ship, event and source, part of a court proceeding. **Collection**: is the collection of documents that belong to Classes such as Class A, B which hold list of documents which are in tabular format with subject, year with page number.

**Document** : consists of a detailed record of the type of event, transaction that occurred such as court proceeding, that include a case name or person name as recorded by a clerk which were hand written.

**List of Documents**: the list of documents that have part of court proceedings with the type of document, title, year and the location the document was recorded.

**About the publisher** : The Irish University Press Series of British Parliamentary Papers published several volumes of Parliamentary papers released by the court proceedings.

#### 5.1.1 Data format

Each of the volume consists of one more collections along with a table of contents page. Each collection is a PDF that consists of handwritten table of list of papers which contain document information of court proceedings of the emancipation processes, case information, boat information placed as a

table. Additionally the documents themselves are part of the collection with the actual document content.

The Table of contents pages examples from Slave Trade Volume 10:

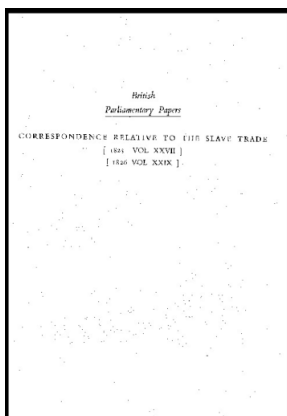


Figure 2: Page  
1

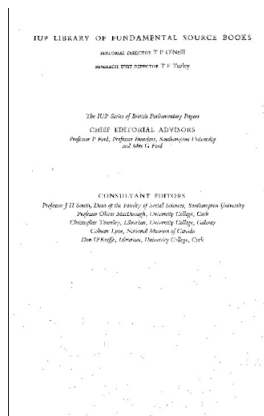


Figure 3: Page  
2

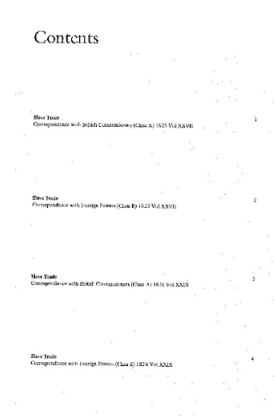


Figure 4: Page  
3

Figure 5: Table of Contents from Slave Trade Volume 10

The Class A Collection pages examples from Slave Trade Volume 10:

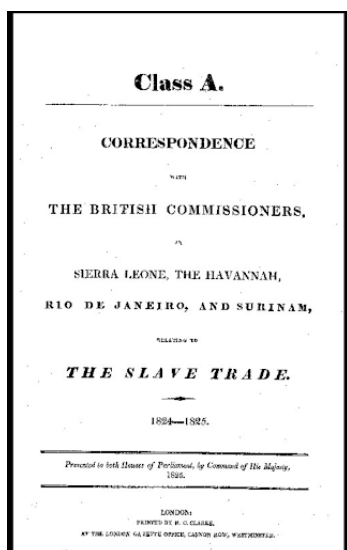


Figure 6: Page 1

LIST OF PAPERS		
SIERRA LEONE. (General.)		
No.	Date & Recd.	Page
1. Mr. Sec <sup>y</sup> Canning to H. M's. Comm <sup>r</sup>	— 18. Feb. 1825	1
One Enclosure		
2. Mr. Sec <sup>y</sup> Canning to H. M's. Comm <sup>r</sup>	— 18. Feb. 1825	2
3. H. M's. Comm <sup>r</sup> to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	3
4. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	13
5. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	13
6. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	14
7. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	14
8. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	14
9. Mr. Sec <sup>y</sup> Canning to H. M's. Comm <sup>r</sup>	— 18. Feb. 1825	15
Four Enclosures		
10. H. M's. Comm <sup>r</sup> to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	17
11. D. M. Hamilton, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	17
12. Mr. Sec <sup>y</sup> Canning to D. M. Hamilton, Esq.	— 18. Feb. 1825	18
13. Mr. Sec <sup>y</sup> Canning to D. M. Hamilton, Esq.	— 18. Feb. 1825	18
SIERRA LEONE. (Separate.)		
14. D. M. Hamilton, Esq. to J. Planta, Jun. Esq.	— 18. Feb. 1825	19
15. H. M's. Comm <sup>r</sup> to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	19
16. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	19
17. E. Gregory, Esq. to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	20
18. Lord Howard de Walden to E. Gregory, Esq.	— 18. Feb. 1825	20
Two Enclosures		
19. G. Rantall, Esq. to J. Bantick, Esq.	— 18. Feb. 1825	21
20. H. M's. Comm <sup>r</sup> to Mr. Sec <sup>y</sup> Canning	— 18. Feb. 1825	22
21. E. Gregory, Esq. to Lord Howard de Walden	— 18. Feb. 1825	22

Figure 7: Page 2

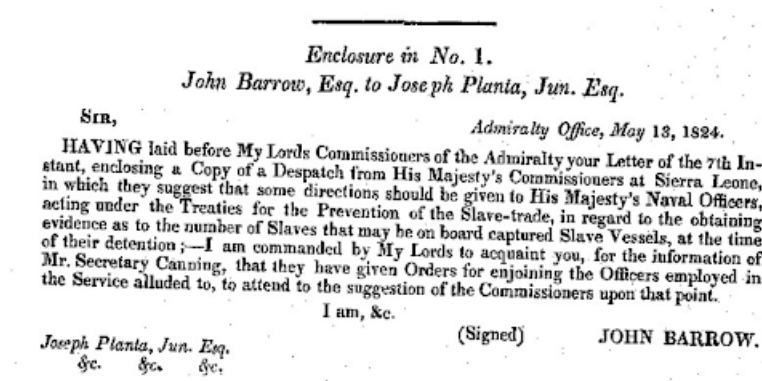
Figure 8: Class A Collection from Slave Trade Volume 10

## 5.2 About the Data & MacBERTh Manjavacas and Fonteyn (2021)

Historical English is vastly different from Contemporary English. More specifically, documents from 1800s are more influenced by Middle English. The first Phoenetic alphabet itself (from Proto-Canaanite script) was created by Semitic-speaking workers and slaves in the Sinai Peninsula. This Phoenetic alphabet was not available until late 1800 (updated version by Charles Morton): [https://en.wikipedia.org/wiki/Alphabet#/media/File:Orbis\\_eruditi\\_literatura\\_%C3%A0\\_caractere\\_Samaritico\\_deducta\\_1689.jpg](https://en.wikipedia.org/wiki/Alphabet#/media/File:Orbis_eruditi_literatura_%C3%A0_caractere_Samaritico_deducta_1689.jpg). The spellings in 1800 were more different, there was the usage of short and long S. There were pronouns such as “Thou”, “shall not”, then there was the semantic narrowing effects such as lexical set of words grouped semantically by meaning: example “To suffer” was used commonly in place of “to allow”. [https://en.wiktionary.org/wiki/suffer#:~:text=\(transitive\)%20To%20endure%2C%20undergo](https://en.wiktionary.org/wiki/suffer#:~:text=(transitive)%20To%20endure%2C%20undergo).

### 5.2.1 Analysis of language of Slave Trade volumes

As part of my understanding and learning curve, I analyzed the text documents. The English language used in Slavery Trade volume collections and reading the content seems to have a differences in language usage, grammar compared to contemporary English language. Aside from various other spelling corrections, auto-completion algorithms that were used and discussed for Fuzzy String matching algorithms, it seems the language itself is very different around the late 1800s. Some of the textual documents such as Class A collection from Slavery Trade volume 10, for example:



*Enclosure in No. 1.*  
*John Barrow, Esq. to Joseph Planta, Jun. Esq.*

SIR,

*Admiralty Office, May 13, 1824.*

HAVING laid before My Lords Commissioners of the Admiralty your Letter of the 7th Instant, enclosing a Copy of a Despatch from His Majesty's Commissioners at Sierra Leone, in which they suggest that some directions should be given to His Majesty's Naval Officers, acting under the Treaties for the Prevention of the Slave-trade, in regard to the obtaining evidence as to the number of Slaves that may be on board captured Slave Vessels, at the time of their detention;—I am commanded by My Lords to acquaint you, for the information of Mr. Secretary Canning, that they have given Orders for enjoining the Officers employed in the Service alluded to, to attend to the suggestion of the Commissioners upon that point.

I am, &c.

(Signed) JOHN BARROW.

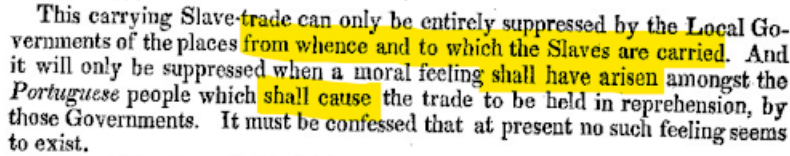
Joseph Planta, Jun. Esq.  
&c. &c. &c.

Figure 9: Database Schema

### Why BERT (Devlin et al., 2018) will not work for Slavery Trade volumes?

The address such as My Lords Commissioners is very less likely to be classified or learnt by modern Language model such as Bidirectional Encoder Representations from Transformers i.e BERT. BERT was originally trained on Book corpus and English Wikipedia, which uses more modern versions of books. The book corpus specifically used for the BERT language model consists of 11,038 unpublished books from 16 different genres and English Wikipedia consists of 2,500 million words from text passages. (Devlin et al., 2018) BERT uses modern Language constructs, i.e. from the 21st century. The language learnt by BERT is expected to work on search ranking, information extraction, event extraction, named entity extraction and Question Answer systems.

The highlighted words/phrases from above image such as the use of words “from whence”, “shall have arisen”, “shall cause”. For example, whence <https://en.wiktionary.org/wiki/whence> is noted as uncommon in contemporary English language.



This carrying Slave-trade can only be entirely suppressed by the Local Governments of the places from whence and to which the Slaves are carried. And it will only be suppressed when a moral feeling shall have arisen amongst the Portuguese people which shall cause the trade to be held in reprehension, by those Governments. It must be confessed that at present no such feeling seems to exist.

Figure 10: Example English language sentences from 1824 collection (Slave Trade Volume 10)

Having a Language model such as MacBERT<sub>h</sub> Manjavacas and Fonteyn (2021) from older English versions helps with word sense disambiguation, Named Entity Recognition tasks. (Manjavacas and Fonteyn, 2022) We anticipate the need for Named Entity Recognition. Thus for addressing Henry’s problem definition, we use MacBERT<sub>h</sub> model to train over the extracted text data and develop a further pipeline. In addition to this, historical documents which are digitized often contain unstable Orthography, OCR noise and changes in lexis and grammar (Manjavacas and Fonteyn, 2021). So by using MacBERT<sub>h</sub> we reduce problems from OCR noise and changes in lexis and grammar.

## 6 Problem Definition

### 6.1 Henry’s Problem Definition

The main goal is to provide a search space such that given a set of search terms (is a row of event details in text format), we query the terms in documents and then extract information from Named Entity Corpus and auto-fill an inventory table with explanations.

Search space is a set of volumes. Each volume consists of about 1 or 3 collections. Usual structure for each volume:

1. Table of contents - consists of title page, lists of collections
2. Collections - consists of documents along with an index of List of Papers which are referred as list of documents (LOD).

1 collection structure: The index pages consists of title page, List of papers.

The manually collected data from Henry, which should be target output of a search space is as follows:

Year	Month	Day	Event	Location	Period	Category	Count	Count	Count
1807	January	1	Arrival	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	February	1	Departure	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	March	1	Arrival	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	April	1	Departure	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	May	1	Arrival	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	June	1	Departure	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	July	1	Arrival	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	August	1	Departure	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	September	1	Arrival	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	October	1	Departure	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	November	1	Arrival	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70
1807	December	1	Departure	London	Period 1	Colonial and Foreign Offices, Great Britain	184	156	70

Figure 11: Dataset 1

Liberated Africans Digital Archive			
Metadata Documentation			
Descriptive metadata for the distribution/access copies of digital resources in project			
Place Scheme			
Place: A definitive location that remains constant through time and space			
#	Required/Optional	Term Name/DC Map to DC	Definition/Description & Comments
01	Required	Place Name	Label or name given to the place (if relevant)
02	Required	Place Type	General category or class of location - distinct from the place name aspects, used to group together similar places (such as ports or coasts) (if relevant)

Figure 12: Dataset 2

Figure 13: Henry's data manually transcribed

### 6.1.1 Solutions

1. Build index for the volumes for title page, table of contents, collections and list of documents.
2. Using the index, match search year or year ranges from indices and narrow down document search space.
3. Search for boat name or case name from the search space from step 2.

This is essentially a document structure representation, knowledge representation task with Named Entity Recognition.

### 6.1.2 Approaches to extract information

1. Extract Events (year mentions that match boat or case names).
2. Extract all pages with boat names and bookmark.

### 6.1.3 Search Space

For the purpose of this goal, having basewords and spelling variants is certainly helpful to recognize and develop search based on basewords, variants likewise for same document with variation in ranking result. For example, we can rank 10 documents where a text such as Franco to display both base-word matches as well as spelling variants. We could use all proper names,



such as case names, boat names, person names and create variants for each of the words. Then we can take all locations from Henry’s data (refer figure 13) and create basewords as locations without variants. We assume a certain level of manual correction.

#### 6.1.4 Solution design

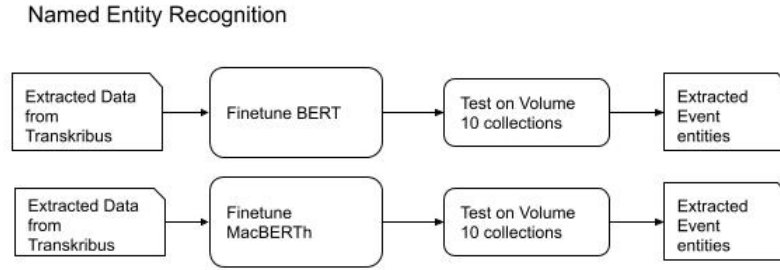


Figure 14:

Figure 15: Henry’s data manually transcribed

## 6.2 Kartikay’s Problem Definition

The goal of this problem is to build index from table of contents from Slave Trade Volumes. We first build indices of all important content that provide information about search space. A few

1. We take Volumes with collections as the input which are PDF files.
2. Extract table of contents, list of documents.
3. Finally build an excel sheet of index for each volume.

### 6.2.1 Example

We consider an example for Slave Trade 10 which is the volume that consists of:

- 1) Table of Contents

- 2) Class A collections
- 3) Class B collections

Each collection has documents listed in a table. Documents are the "List of Papers" table summaries with page numbers, boat names and/or case-names. Please refer Data and Format section for further description and reference.

The expected output from List of Papers is as follows:

1. Location of the Court : SIERRA LEONE (Spain.) index number from List of papers/documents: 32.
2. Title/Destructor : Mr. Secy. Canning to H. M's. Commrs.:
3. D. 29 May 1825 -i date (each document has a D and R and if lines provided means they follow previous row's year). R. (-i Received)
4. Case name or Destructor: Fabiana
5. Page Number from the List of documents/papers (NOT the PDF) : 27
6. PDF Page number: 24

Expected output from Table of Contents page for the volume, following information comes from Table of Contents (generally, assuming same structure):

1. Name of the Editors: **\*\*in\*\***(remove this) T. P. O'Neill, T. F. Turley, et al., eds.,
2. Name of the Collection: "Correspondence with British Commissioners (Class A) 1825 Vol XXVII,"
3. Name of the Publisher: Irish University Press Series of British Parliamentary Papers
4. Volume number: Slave Trade, vol. 10
5. Original Publisher: (Shannon: Irish University Press, 1968-1969).

## 6.2.2 Known Challenges

From volumes and collections analysis, there are known challenges as follows:

1. Between each of the Volumes, collections vary in number and structure. For example, we can see three different Volume structure each consisting of collection where table of contents or multiple collections is not necessarily the same between each of the volume. Slave Trade Volume 10 has 4 collections and a Table of contents while Slave Trade Volume 30 has only single document.

Volume	Collection Name	Publication Date	Pages
Slave Trade 10	Table of Contents	1815	6
	Correspondence with the British commissioners at Sierra Leone, the Havannah, Rio de Janeiro, and Barbados, relating to the slave trade, 1804-1805	1805	100
	Foreign powers, relating to the slave trade, 1804-1805	1805	100
	Correspondence with the British commissioners at Sierra Leone, the Havannah, Rio de Janeiro, and Barbados, relating to the slave trade, 1805-1806	1805	98
Slave Trade 30	Table of Contents	1815	6
	Correspondence with the British commissioners at Sierra Leone, the Havannah, Rio de Janeiro, and Barbados, relating to the slave trade, 1804-1805	1805	100
Slave Trade 32	Table of Contents	1815	6
	Correspondence with the British commissioners at Sierra Leone, the Havannah, Rio de Janeiro, and Barbados, relating to the slave trade, 1804-1805	1805	100
	Foreign powers, relating to the slave trade, 1804-1805	1805	100
	Correspondence with the British commissioners at Sierra Leone, the Havannah, Rio de Janeiro, and Barbados, relating to the slave trade, 1805-1806	1805	98

Figure 16: Comparison of collection structure between volumes.

2. Another challenge, that cannot be generalized is that page numbers are different between each of the collections. The true page number is different from handwritten page number on the pages. Example: Suppose we want to find TRUE page number for Fabiana from List of papers in page 3. Fabiana is referred in Page number 27. So we navigate to page number 27 but it shows page number as 21, but we know true page number is 33, which does match page 27 listed as per List of Papers table.

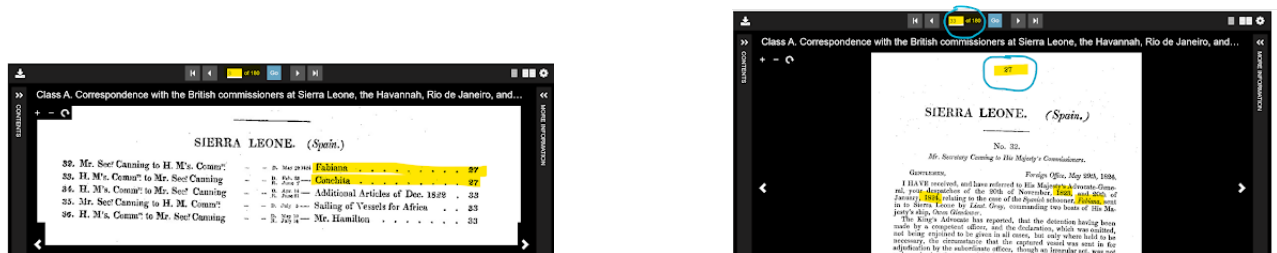


Figure 17: Comparison of collection structure between volumes.

3. Table of Contents from List of Papers/documents all are in single document. Example: Slave Trade 30, there is only one document.

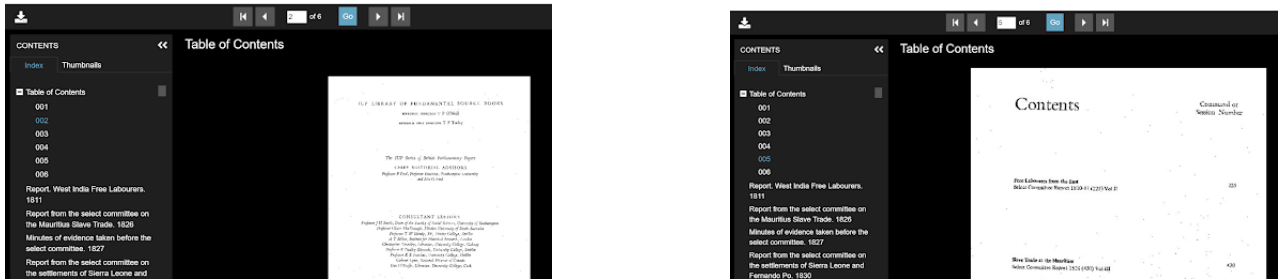


Figure 18: Comparison of collection structure between volumes.

- Year mentioned in List of documents table is not continuous - blank for few years is assumed to mean it is same as last written year from previous rows in the table. Example: we assume year 1824 (from previously filled year entry) for all the missing year entries in Date & Receipt column until year appears in another row. This is however a common technique used in handwritten numbers or repeated entries during large-scale manual record collection tasks which are handwritten.

iv

LIST OF PAPERS.

No.		Date & Receipt.		STANDARD.	Page
37.	Mr. Sec <sup>r</sup> Canning to H. M.'s Comm <sup>r</sup>	—	Do. Sept. 3 1825	Fabiana	34
	One Enclosure				
38.	E. Gregory, Esq. to Mr. Sec <sup>r</sup> Canning	—	Do. Aug. 5 1825	Fabiana. Conchita	35
39.	E. Gregory, Esq. to Mr. Sec <sup>r</sup> Canning	—	Do. Oct. 2 1825	Ditto	35
40.	E. Gregory, Esq. to Mr. Sec <sup>r</sup> Canning	—	Do. Oct. 2 1825	Fabiana	35
41.	E. Gregory, Esq. to Mr. Sec <sup>r</sup> Canning	—	Do. Oct. 2 1825	Conquistador, Nicanor, and Nueva Havannah	36
42.	H. M.'s Comm <sup>r</sup> to Mr. Sec <sup>r</sup> Canning	—	Do. Nov. 3 1825	Fabiana	36
43.	H. M.'s Comm <sup>r</sup> to Mr. Sec <sup>r</sup> Canning	—	Do. Nov. 11 1825	Ditto	37
	Two Enclosures				

### 6.2.3 Research Questions

We would like to find how to extract the the text and tabular information without loosing document structure.

- What different methods we can use to extract metadata from OCR text documents?
- How accurate is the meta data to the ground truth (which is going to be built based on original pdf collections)?

#### 6.2.4 Approach to solve by Brute Force Method:

We assess the results by using a Brute force method, assuming generic document structure.

1. Manually look for table of contents, title list of papers document and extract content.
2. We expect this method will give errors due to differences in entries, other known problems (missing entries, no entry for year, change in columns etc).
3. We provide analysis and results for this approach.

## 7 Brute Force Analysis

In this approach, we first manually look for table of contents, title and list of papers from collections and extract content. And analyse the results and problems. To conduct Brute force method, we explore following methods, tools that are commonly used:

1. The tools that exist to extract page-wise raw text from PDFs.
  - (a) The full list of locations to extract.
  - (b) The full list of Case names or Descriptors.
  - (c) To extract paragraphs, tabular rows that match “keywords”.
2. For table of contents and List of Documents/Papers, extract following simple, brute force rules:

Columns	Extraction Mechanism (Brute force)
Name of the Editors	Extract next lines between IUP LIBRARY OF FUNDAMENTAL SOURCE BOOKS and The IUP Series of British Parliamentary Papers
Name of the Collection	Content that starts with "Correspondence with"
Name of the Publisher	Content that starts with "Irish University Press Series of British Parliamentary Papers"
Volume number	Content that starts with "Slave Trade"
Original Publisher	Line containing "Irish University Press"

Columns	Extraction Mechanism (Brute force)
Location of the Court	Extract location between "Class A - CORRESPONDENCE THE BRITISH COMMISSIONERS, AT" and "Relating to Slave trade"
Index number from List of papers/documents	The first <u>column's</u> row value in List of papers
Title/Descriptor	The second <u>column's</u> row value in List of papers
D. and Received dates	The Date & receipt <u>column's</u> row value in List of papers
Subject	The subject <u>column's</u> row value
Page number	Give TRUE page number

Figure 19: Brute force approach to extract information to fill in the target CSV file.

We explored following solutions, we present results from each tool:

1. Extract text from PDF using PyPDF2 software <https://pypi.org/project/PyPDF2/>.

- (a) The content is split into text, but words are not more together.
- (b) The letters, spaces and special characters are inserted during text extraction from pdf.

```
doc = 'E:\\cu\\summer2022\\independent-study\\documents\\Slave Trade Volume 10\\toc.pdf'
pdf_file_obj = open(doc, 'rb')
pdf_reader = PyPDF2.PdfFileReader(pdf_file_obj)
print(pdf_reader.numPages)
filename = os.path.splitext(doc)[0] + ".txt"
with open(filename, 'a') as ofile:
    for pg_num in range(pdf_reader.numPages):
        this_page = pdf_reader.getPage(pg_num).extractText()
        print(this_page)
#pdf_file_obj.close()

Fo
Identifier
:
Range:
S
c
Download
d
e
r
for
Resear
c
with
the
qu
a
```

Figure 20: PyPDF2 output

2. Extract full text and copy and paste to text document manually.
  - (a) This method is tedious and time taking
  - (b) Also consists of some special characters when copying text manually.
3. Extract text using Apache pdfbox: <https://pdfbox.apache.org/andusingpython> and <https://github.com/lebedov/python-pdfbox>
  - (a) Less time taking to extract text content
  - (b) Less computing power
  - (c) The content is split into text, but words are not together.
  - (d) Sentences are scattered, tabular structure is not preserved.

[List-57]  
[1] DATE OF PAPERS.v0n0.vn1.vn2.vn9.n0.vn1.vn2.vn1.vn3.vn4.vn5.vn5.vn1.vn7.v8.vn9.vn20.vn21.vnSERRA LEONE, (Generat.)vN  
[2] , Date & Receipt.SUBJECT.vN.M. Hamilton, Esq. to Mr. Secy- Canning n.d. 27.1825 Mr. Hamilton and Mr. Refeill.vnApr16 vN  
[3] ~Judgeships at Sierra Leone.vM. I . . Secy: a D. March 14 .D. M. Hamilton, Esq. to Mr. Seer "Canning June 14: 77 Abstracto  
f Proceedings in 1824Mr. Secy. Canning to H. M.'s Comms., D. June 15 = Papers laid before Parliament.vN.M.'s Comms-to-  
[4] ~vN.H. M. Hamilton, Esq. to Mr. Secy- Canning 1825 vNJune 15 - Annual Report of the Sierra Leone Settlers.  
[5] 4 = Receipt of Despatches ~vM.R. Secy. Canning to J. T. Williams, Esq. - D. Sept. 10 - General Instructions ~vM. M's Commr:  
to Mr. Secy. Canning - rz gne = Arrival of Mr. Smith ~vN.Planta, Jun. Esq. to H. M.'s Comms., D.Sep't 27 - Mr. Williams -  
[6] ~vN.H. M's Comms. to Mr. Secy- Canning - Rsept. es, te ~vofParliamentary ~vSERRA LEONE, (Separate.)vN.M. Consul-General Clarke  
to Feb. 1825 vNSept. 1825 vNSept. 1825 - "Seail - Two Slave Vessels ~vM.Renda, Esq. to Mr. Secy- Canning 1825  
[7] ~vEmancipated Slaves ~vN.M. Hamilton, Esq. to Mr. Secy- Canning z frit 13 - Leave of Absence ~vN.Refeill, Esq. to Joseph P  
lanta, Jun. Esq. - R au ib - Leave of Absence ~vM.R. Secy- Canning to H. M.'s Comms., D.Oct. 7 - Vacancies in the Mixed Com  
Two Enclosures missions ~vSERRA LEONE, (Spain.)vN.April 13 1825, Espafola ~vN.M.'s Comms.- to Mr. Secy- Canning - R13  
[8] ~vN.M. Hamilton, Esq. to Mr. Secy- Canning 1825 vNSept. 1825 - "Seail - Two Slave Vessels ~vM.Renda, Esq. to Mr. Secy- Canning 1825  
EONIE, (Portugal.)vN.M. Secy- Canning to H. M.'s Comm, D. April 23 1825, Dez de Fevereiro ~vInthree EnclosuresvND. M. Hamilton  
[9] N, Esq. to Mr. Secy: Canning 2May {~ Diana, Dos Amigos Brazil ~vThree Enclosures ~v lieros, and Avizo ~vN.M. Comm to Mr.  
Secy: Canning - n May i - Bella Eliza ~vInfor EnclosurevN.M.'s Comms. to Mr. Secy- Canning \*June 14 - Overrating of Tomna  
[10] ~vN.M. Hamilton, Esq. to Mr. Secy- Canning 1825 vNJune 15 - Annual Report of the Sierra Leone Settlers.  
1819.vn19.vn20.vn20.vn20.vn20.vn25.vn28.vn30.vn31.vn32.vn33.vn34.vn35.vn36.vn37.vn38.vn39.vn40.vn41.vn42.vn43.vn44.vn45.vn46.vn47

Figure 21: PDFBOX output

## 4. Using Tesseract

- (a) We need to extract page as an Image of JPEG or PNG format.
- (b) By manually extracting required pages into PDF and then by using pdf2jpg package converters, we can extract required pages into separate documents and then extract images from the pages.
- (c) The content is split into text, but words are not together.
- (d) Sentences are scattered, tabular structure is preserved to some extent, but columnar structure seems to have been lost (Years column is lost).

```
key = list(volumes_img_dict.keys())[0]
file = volumes_img_dict[key][0]
print(pytestesseract.image_to_string(file, timeout=2000))
```

---

OF PAPERS

---

SIERRA LEONE.

No.

1. Mr. Sec' Canning to H. M's. Comm'' ~

2. Mr. Sec! Canning to H. M's. Comm'' ~

3, H

A

5. E. Gregory, Esq. to Mr. Sect Canning =~

6. E. Gregory, Esq. to Mr. See? Canning ~

7. E. Gregory, Esq. to Mr. SectCanning = -

8. E. Gregory, Esq. to Mr. Sec!Canning = -

9. Mr. Sec! Canning ta H. M's, Comm? ~

Figure 22: PDFBOX output

## Jupyter Notebook Solutions for Bruteforce:

1. Brute force method without comments in code (Jupyter Notebook as HTML): <https://drive.google.com/file/d/1pbs2aL-KpNt0Ij9urEort7nD1yVJH2q0/view?usp=sharing>
2. Brute force method with Tesseract and with detailed comments in code (Jupyter Notebook as HTML): <https://drive.google.com/file/d/11Uq0DgvJuhC1Q9j3XTzrnf7NN6nhvEnm/view?usp=sharing>

Broadly we found two general problems from Brute force method:

1. Table structure is not preserved by PDF extraction tools, especially scanned documents that were OCR converted.
2. Given any random OCR scanned Text document, we cannot retain PDF encoded Document structure (such as tables, headings, titles, page numbers).
3. Portable Document Format (PDF) Encodings are different from regular encodings as they preserve metadata, layouts. Reference: <https://opensource.adobe.com/dc-acrobat-sdk-docs/> On Windows: WinAnsiEncoding is used by Adobe.  
Default: StandardEncoding is used by Adobe.

### 7.0.1 Suggested Solutions

We follow a different approach to now address the problem by using Transkribus (Bell and Ranade, 2015):

1. To use Transkribus scanned documents - Extracted using Transkribus <https://transkribus.eu/lite/home>
2. Define data structure for tabular data, Table of Contents. Then define and develop a benchmark for the data based on Text data standards. Finally design a dataset format in JSON or CSV files.
3. OPTIONAL: Use PDF Structure Classification from Extracted text from step 1.



## 7.0.2 Extract information from Transkribus

We next studied Transkribus (Kahle et al., 2017). By understanding steps to detect layout, extracting information, we follow the steps from <https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-i> by using an example from <https://readcoop.eu/transkribus/howto/how-to-transcribe-docu>

By using List of papers pages from each collection from Slave Trade volume 10, we can see following example layout detected from pre-trained Transkribus model for English language.

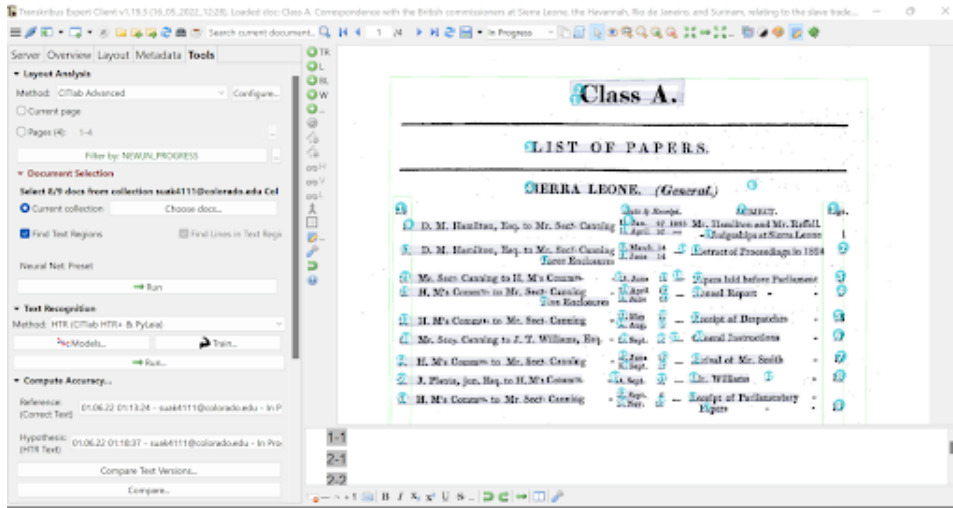


Figure 23: Trankribus example.

## 8 Conclusion and Discussion

To address Henry's problem definition, we first need to extract data and layout from the collections. To extract data and layout from PDF collections we follow approaches broadly discussed in Henry's Problem definition and Kartikay's Problem definition.

## 9 Future Work

We will first address solutions discussed in Kartikay's Problem definition section using Slave Trade Volume 10, 11, 12. Then once data was extracted, we

correct from manual transcribing. We will provide and define a pipeline, description for how-to conduct this and generalize this data extraction for future documents. We then develop a simple model over MacBERTh (Manjavacas and Fonteyn, 2021) to train the extracted text and conduct a Named Entity Recognition to bookmark pages for a simple case such as Case names, boat names. We then provide a detailed final report with solutions, challenges and analysis conducted for the above two broad classes of problems.

## 10 List of tools and methods used

This section lists all the tools and methods used.

### 10.1 Tools and software

- Python3
  - NLTK
  - re
- PyPDF2
- Apache PDFBOX
- Tesseract
- Transkribus

<https://www.colorado.edu/lab/dsrl/collaborators>

## References

- Bell, M. and Ranade, S. (2015). Traces through time: a case-study of applying statistical methods to refine algorithms for linking biographical data. In *BD*, pages 24–32.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

- Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (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.
- Lovejoy, H. B. (2020). Who did what when? acknowledging collaborative contributions in digital history projects. *Esclavages & Post-esclavages. Slaveries & Post-Slaveries*, (3).
- Lovejoy, P. E. and Chadha, K. (2021). Equiano’s world: Chronicling the life and times of gustavus vassa. *Esclavages & Post-esclavages. Slaveries & Post-Slaveries*, (4).
- Manjavacas, E. and Fonteyn, L. (2021). Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*.
- Manjavacas, E. and Fonteyn, L. (2022). Adapting vs pre-training language models for historical languages.
- Schindling, J. P. (2020). *The Spatial Historian: Creating a Spatially Aware Historical Research System*. West Virginia University.
- Shimizu, C., Hitzler, P., Hirt, Q., Rehberger, D., Estrecha, S. G., Foley, C., Sheill, A. M., Hawthorne, W., Mixter, J., Watrall, E., et al. (2020). The enslaved ontology: Peoples of the historic slave trade. *Journal of Web Semantics*, 63:100567.