frances: Cloud-based Historical Text Mining with Deep Learning and Parallel Processing

Lilin Yu
School of Computer Science
University of St. Andrews
St Andrews, UK
ly40@st-andrews.ac.uk

Ash Charlton
School of History, Classics and Archaeology
University of Edinburgh
Edinburgh, UK
A.R.Charlton@sms.ed.ac.uk

Wilfrid Askins
School of Computer Science
University of St. Andrews
St Andrews, UK
wilfridaskins@gmail.com

Melissa Terras
Design Informatics
University of Edinburgh
Edinburgh, UK
M.Terras@ed.ac.uk

Rosa Filgueira
School of Computer Science
University of St. Andrews
St Andrews, UK
rf208@st-andrews.ac.uk

***Abstract*—frances is an advanced cloud-based text mining digital platform that leverages information extraction, knowledge graphs, natural language processing (NLP), deep learning, and parallel processing techniques. It has been specifically designed to unlock the full potential of historical digital textual collections, such as those from the National Library of Scotland, offering cloud-based capabilities and extended support for complex NLP analyses and data visualizations. frances enables realtime recurrent operational text mining and provides robust capabilities for temporal analysis, accompanied by automatic visualizations for easy result inspection. In this paper, we present the motivation behind the development of frances, emphasizing its innovative design and novel implementation aspects. We also outline future development directions. Additionally, we evaluate the platform through two comprehensive case studies in history and publishing history. Feedback from participants in these studies demonstrates that frances accelerates their work and facilitates rapid testing and dissemination of ideas.**

***Index Terms*—digitised historical collections, cloud-based platform, information extraction, knowledge graphs, natural language processing, text mining, parallel processing, Apache Spark**

I. Introduction

The digitization of historical texts has opened up vast possibilities for advancing research in the fields of history, culture, and linguistics, and in recent years, there has been an increased interest in Digital Humanities (DH) textual datasets within the Natural Language Processing (NLP) community [1], [2]. However, the sheer scale and heterogeneity of these digital collections pose significant challenges for researchers aiming to extract meaningful information, establish semantic relationships, and perform comprehensive text mining analyses that transcend simple search and retrieval method [3]. This is particularly the case where a *temporal* analysis of textual material is involved, e.g., across multiple editions of newspapers, journals, gazetteers, chapbooks or encyclopaedias. *Temporal* analysis focuses on studying the evolution of language over different time periods and has gained increasing importance in humanities research, such as tracking the shifts in perceptions related to sensitive cultural topics like gender or race. Overcoming the impediments associated with historical textual analysis requires addressing various obstacles, including integrating data from disparate sources in different formats, correcting and cleaning text data extracted from digitised assets, segmenting text entries, identifying semantic connections, and processing large volumes of data spanning multiple editions of encyclopedic and journalistic works.

To address these challenges and facilitate research for non-technical researchers, we have developed frances [4], an advanced digital platform that combines deep learning, natural language processing (NLP), and text mining techniques. Initially, frances was a prototype limited to analyzing the first eight editions of the *Encyclopaedia Britannica* (EB) provided by the National Library of Scotland (NLS) Data Foundry [1]. However, due to architectural limitations and restricted performance on a single local machine, frances was not made available to the public, despite its access to advanced facilities.

In this work, we present significant advancements in frances. We introduce novel methods for automatically generating knowledge graphs that extract valuable information from historical document collections, offering a functionality previously unavailable to digital humanities researchers. Additionally, we have extensively improved frances by enhancing its architecture, expanding analysis functionalities, incorporating new data visualization capabilities, and enabling support for parallel text-mining queries in the cloud. These advancements provide researchers with a powerful platform to formalize and connect findings and insights derived from the *temporal* analysis of large-scale digital corpora. Furthermore, we evaluate our contributions through two comprehensive case studies. Participants in the case studies praised frances for its ability to accelerate workflows, enable new research practices, and support rapid testing and dissemination of ideas.

The rest of the report is structured as follows. Section II explores the background of defoe, *EB-ontology* and *EB-knowledge graph*. Section III gives an overview of the main features of frances. Section IV introduces our parallel extraction heuristics. Section V details the features of the *NLS ontology*, and the new knowledge graphs supported by frances. Section VI presents the new extensions of defoe,

[1] https://data.nls.uk/data/digitised-collections/encyclopaedia-britannica/

Section VII focuses on the new architecture of frances, while Section VIII introduces the different components of the frances User Interface. Section IX summarises the evaluations. Section X presents two cases studies: Commodity and enslaved labor in the *Encyclopaedia Britannica* and Geoparsing the *Chapbooks of Scotland*. These studies performed to validate the use of frances. Finally, section XI describes related work, and section XII concludes with a summary of achievements and future work.

II. Background

This section introduces the main technologies used in the original version of frances and that are relevant for this work: defoe, *EB Ontology* and *EB Knowledge Graph*.

A. defoe Parallel Library

defoe [5], [6] is a versatile and portable Python library specifically designed for the storage, processing, querying, and analysis of digital historical textual data, primarily in English. By leveraging the Apache Spark big data framework [7], defoe empowers researchers to extract valuable insights from text by running parallel text mining queries. It offers seamless ingestion, extraction, and transformation capabilities for digital collections with various XML schemas and physical representations, utilizing five distinct object models (PAPERS, NZPP, ALTO, NLS). Notably, the NLS model is employed for mapping the collections in this study (see Section IV).

Upon data ingestion, defoe introduces a preprocessing pipeline equipped with advanced NLP techniques to address optical character recognition (OCR) errors and other common issues such as long-S and line-break hyphenation. This preprocessing step significantly enhances the overall text quality. defoe encompasses a comprehensive set of text mining queries that enable efficient searching across large-scale datasets, providing results that can be further analyzed and interpreted. These queries are built on a combination of various operations, including filter, flatMap, map, and reduce. Furthermore, defoe includes a robust SPARQL connector that facilitates querying of SPARQL knowledge graphs. In our earlier work, this connector was primarily designed to query the *EB Knowledge Graph*, as outlined in Section II-B.

B. EB Ontology and EB Knowledge Graph

In our previous work, we introduced the *EB Ontology* [2] as a formal specification of knowledge within the domain of the *Encyclopaedia Britannica*. This ontology defined a comprehensive set of concepts and their relationships, establishing a shareable and reusable knowledge representation for the EB.

Leveraging the *EB Ontology* and the extracted information from the EB collection, frances utilized this knowledge to construct the *EB Knowledge Graph* (*EB-KG*). The *EB-KG*, stored as RDF triples in an *Apache Fuseki server* [3], facilitated efficient querying and analysis. To enhance the extracted knowledge, advanced NLP and deep learning analyses were incorporated. These analyses encompassed several tasks, including sentiment analysis to classify terms into positive or negative categories, topic modeling to cluster terms using Latent Dirichlet Allocation (LDA), term similarity identification, OCR error correction to address errors in automated text recognition, and text summarization for generating concise representations of historical text.

[2]https://francesnlp.github.io/EB-ontology/doc/index-en.html

[3]https://jena.apache.org/documentation/fuseki2/

III. frances - Overview

frances [4] represents a significant advancement in the field of deep learning NLP and text mining for historical textual analysis. It empowers researchers to analyze and extract valuable knowledge from diverse digital textual collections.

In our earlier work, frances existed solely as a prototype, limited to exploring the *Encyclopaedia Britannica* and running on a local machine. In this work, we have expanded the capabilities of frances to facilitate seamless analysis of additional collections from the National Library of Scotland (see Sections IV and V). To accomplish this, we have developed new defoe text mining queries (described in Section VI) and integrated them into *frances'* user interface. Users now have the ability to create accounts and execute parallel defoe queries. To support this enhancement, we have completely restructured the architecture of frances, transforming it into a cloud-based digital platform (elaborated in Section VII). Another notable improvement in frances is its capacity to execute massively parallel defoe queries utilizing a Cloud-based Apache Spark Cluster. The details of this implementation can be found in Section VII-C. Furthermore, this version of frances boasts several enhancements to the User Interface, which are comprehensively outlined in Section VIII. These improvements optimize the user experience and streamline the process of accessing and interpreting analysis results.

IV. Extracting NLS Collections

The NLS data foundry[5] serves as a valuable resource by providing machine-readable formats of collections, facilitating computational research [8]. These collections encompass various types of data, including Digitised Textual material, Metadata, Map and Spatial data, and Organisational data. For the purpose of this study, our focus lies on the Digitised Textual material offered by the NLS, which currently consists of 19 collections [6]. Importantly, all digital textual collections provided by the NLS adhere to the principles of open data.

Each collection is accompanied by two XML-file outputs that serve distinct purposes. Firstly, the Analysed Layout and Text Object Extensible Markup Language (ALTO-XML) [7] files describe the layout information per page. Secondly, the Metadata Encoding and Transmission Standard (METS–XML) [8] files provide metadata information such as the title, author, and publisher for each collection. In this work, we have specifically chosen three NLS digital collections as examples, although our methodologies are applicable to any other collections due to their standardized nature. The selected collections are:

- **Chapbooks printed in Scotland** [9]: This dataset includes over 3,000 chapbooks printed in Scotland, which were popular reading materials from the late 17th to the 19th century. The collection consists of 47,329 ALTO XML

[4]https://github.com/frances-ai

[5]https://data.nls.uk/

[6]https://data.nls.uk/data/digitised-collections/

[7]https://www.loc.gov/standards/alto/

[8]http://www.loc.gov/standards/mets/

[9]https://data.nls.uk/data/digitised-collections/gazetteers-of-scotland/

files at the page level, containing a total of 10 million OCRed words.

- **Ladies' Edinburgh Debating Society** [10]: This collection comprises the complete runs of two Edinburgh journals, The Attempt and The Ladies Edinburgh Magazine, produced by the Ladies' Edinburgh Debating Society. It spans 10 volumes from 1865 to 1874 and 6 volumes from 1875 to 1880. The collection includes 6,354 ALTO XML files at the page level, with around 2.5 million words.
- **Gazetteers of Scotland collection** [11]: This collection consists of twenty volumes of descriptive historical gazetteers of Scotland from the 19th century, providing information on towns, cities, castles, and antiquities. It contains over 13,000 OCRed text files in ALTO-XML format, amounting to nearly 14.5 million words. These gazetteers serve as a comprehensive geographical encyclopedia of Scotland in the 19th century.

To facilitate the automatic extraction of semi-structured information from the selected collections or any other NLS digital collection, we have devised a novel heuristic approach. This approach entails parallel extraction of text from each page using ALTO-XML files, extracting metadata from both the collection and volume using METS-XML files, and subsequently structuring the extracted information. Additionally, for each volume, we have included the permanent URL that allows visualization of the page images. This heuristic is implemented in a new `defoe` query, *write_metadata_pages_yml* [12].

Currently, `frances` supports analyses for these three collections, in addition to the *Encyclopaedia Britannica*. It is worth noting that the extracted information already enables us to conduct rapid data explorations, as demonstrated in Figure 1 for the *Chapbooks printed in Scotland*. These explorations were not feasible before utilizing the collections' XML-files directly.

V. NLS Ontology and Knowledge Graphs

We have expanded the capabilities of `frances` to include support for new collections. As part of this extension, we developed a new ontology called the *NLS Ontology* [13], which represents all the collections from the Data Foundry Data collections of the National Library of Scotland [14].

Figure 2 illustrates the data model of the *NLS Ontology*. A *Serie* (e.g. *Gazetteer of Scotland* can consist of one or more *Volumes* (e.g. *Gazetteers of Scotland 1882*), which may reference *Books* and *Supplements*; A *Serie* also has an *Editor* (e.g., *Wilson, John Marius*) and a *Publisher* (e.g., *W. & A.K. Johnston*), which can be a *Person* or an *Organization*. Each *Volume* contains multiple *Pages*, which have various attributes, including the extracted text. Similar to the approach taken with the *EB Ontology*, the *NLS Ontology* was created in OWL ontology format [15] using Chowlk [9]. Subsequently, frances utilized Widoco [10] to publish enriched and customized documentation of the the *NLS Ontology* with a permanent identifier provided by the w3id.org service.

[10]https://data.nls.uk/data/digitised-collections/edinburgh-ladies-debating-society/

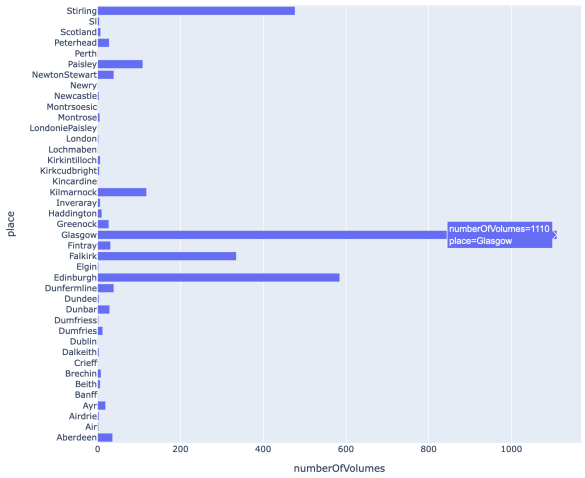
[11]https://data.nls.uk/data/digitised-collections/gazetteers-of-scotland/

[12]https://github.com/frances-ai/defoe_lib/blob/main/defoe/nls/queries/write_metadata_pages_yml.py

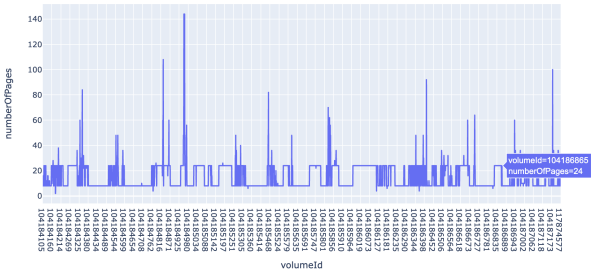
[13]https://francesnlp.github.io/NLS-ontology/doc/index-en.html

[14]https://data.nls.uk/data/digitised-collections/

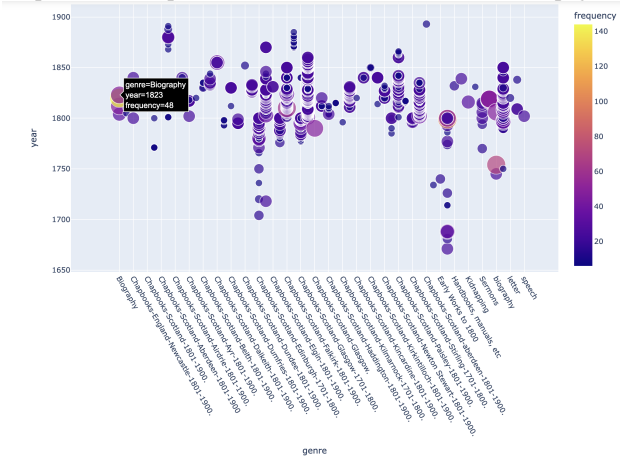
[15]https://www.w3.org/OWL/



(a) Number of volumes per publication place. Most of the chapbooks were published in Glasgow, follow by Edinburgh.



(b) Number of pages per volume. This confirms that most of the chapbooks were printed into books of 8, 12, 16 and 24 pages.



(c) Frequency of genres per year. Each bubble indicates the frequency of a genre for a particular year

Fig. 1: Data-explorations using the *Chapbooks* extracted info.

In this work, we have developed a new methodology [16] to automatically generate NLS knowledge graphs. This methodology leverages the *NLS Ontology* and utilizes the information extracted by the `defoe` query presented in Section IV. The resulting knowledge graphs have Uniform Resource Identifiers (URIs) to identify individual resources such as series, volumes, and pages. They can be queried using SPARQL [17], a semantic

[16]https://github.com/francesNLP/frances/blob/main/NLS_Generic/Dataframe2RDF.ipynb

[17]https://www.w3.org/TR/rdf-sparql-query/

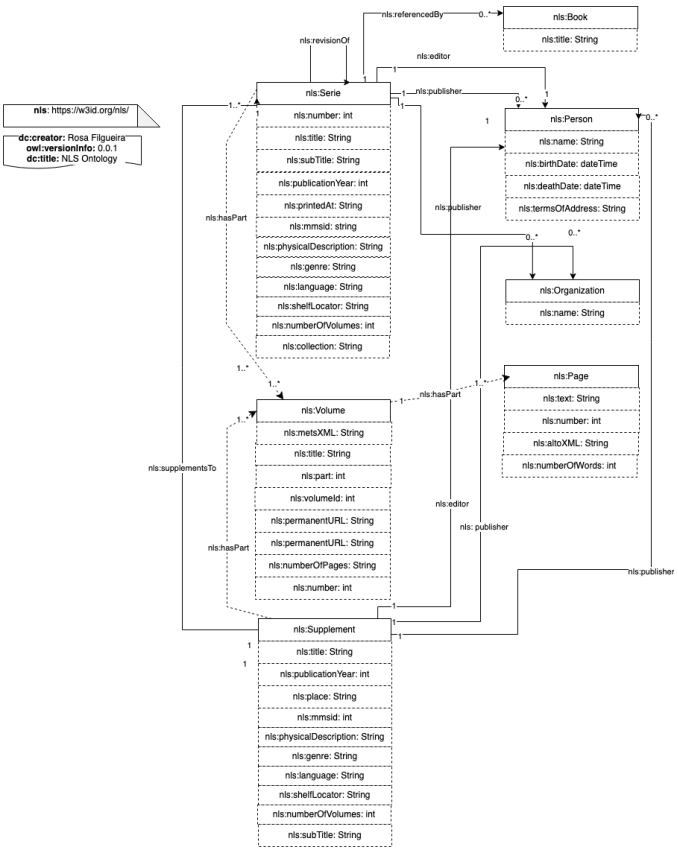


Fig. 2: Data Model of the *NLS Ontology*

query language for retrieving and manipulating RDF data.

To demonstrate the effectiveness of this approach, we have automatically created three knowledge graphs, one for each collection introduced in Section IV: *ChapbooksScotland-KG* [11], *LadiesDebating-KG* [12], and *GazetteersScotland-KG* [13]. These knowledge graphs are also stored in our *Apache Fuseki* alongside the *EB-KG*, allowing us to query them using the SPARQL language. Figure 3 provides an example [18] of quickly calculating the frequency of the term 'Mary' across the publications of the *Ladies' Edinburgh Debating Society* publications using the *LadiesDebating-KG*.

VI. Extensions to `defoe`

We modified the `defoe` SPARQL connector to query the new knowledge graphs introduced in Section V. Additionally, we developed new `defoe` queries and enhanced existing ones to interact with the *EB-KG* and any knowledge graph based on the *NLS Ontology*. These are:

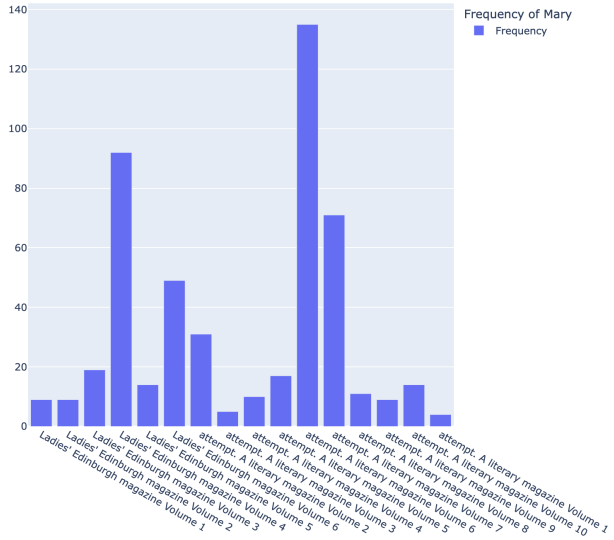
- `frequency-distribution`: Calculates the frequency of the most 'N' common tokens in terms definitions/pages. Users can specify the number of tokens (N) and provide a list of tokens to exclude.
- `lexicon-diversity`: Computes the lexical diversity metric for a given collection, which is the ratio of the vocabulary size to the total number of words in the text. The vocabulary consists of unique words in the text.

[18]Notebook at https://github.com/francesNLP/frances/blob/main/NLS_Generic/QueringRDF_LadiesDebating.ipynb

```
sparql = SPARQLWrapper("http://35.228.63.82:3030/ladies_debating/sparql")
sparql.setQuery("""
SELECT ?text ?title
        WHERE {
        ?page a nls:Page .
        ?page nls:text ?text .
        ?v nls:hasPart ?page .
        ?v nls:title ?title
        FILTER regex(?text, "Mary")
        }

""")
sparql.setReturnFormat(JSON)
results = sparql.query().convert()
```

(a) Query to obtain the frequency of 'Mary' across volumes.



(b) Visualization of the frequency results calculated above.

Fig. 3: SPARQL query to explore *LadiesDebating-KG*.

- `frequency_keysearch_by_year`: Determines the frequencies of one or multiple keywords or key sentences in terms definitions/pages, applying various pre-processing techniques. The results are grouped by year.
- `publication_normalized`: Counts the total number of volumes, pages, and words for a collection, providing results per year. For the *EB Knowledge* graph, it also calculates the number of terms per year.
- `fulltext_keysearch_by_year`: Searches and extracts full text definitions/pages based on different filtering settings. This query supports various pre-processing techniques, and the results are grouped by year.
- `snippet_keysearch_by_year`: Similar to the previous query, but instead of full text, it returns snippets of text. Users can configure the snippet size.
- `uris_keysearch`: Extracts URIs of terms/pages that contain selected keywords or key sentences in their definitions. This query utilizes different pre-processing techniques, and the results are grouped by URI.
- `geoparser_by_year`: Geo-locates locations mentioned in text and resolves them using the Edinburgh Geoparser [14]. Users can limit the geographical area and select the gazetteers for place name resolution.
- `person_entity_recognition`: Identifies people mentioned in text and estimates the gender distribution.

The previous queries interact with the *EB-KG* at term level, while for the new knowledge graphs, they work at page level.

VII. FRANCES ARCHITECTURE

The new architecture of frances is designed with a containerized approach, where the system is decoupled into four distinct containers, each serving a specific purpose, as it is shown in Figure 4. This containerization ensures high maintainability, efficient system deployment, and scalability. The first container is the *React Frontend*, responsible for handling the user interface. It enables a seamless and intuitive user experience, providing a user-friendly interface for interacting with the platform. The second container is the *Flask Backend*, which implements the core functionalities of frances. It handles data processing tasks and query handling, ensuring efficient execution of text mining operations and analysis on the digital textual collections.

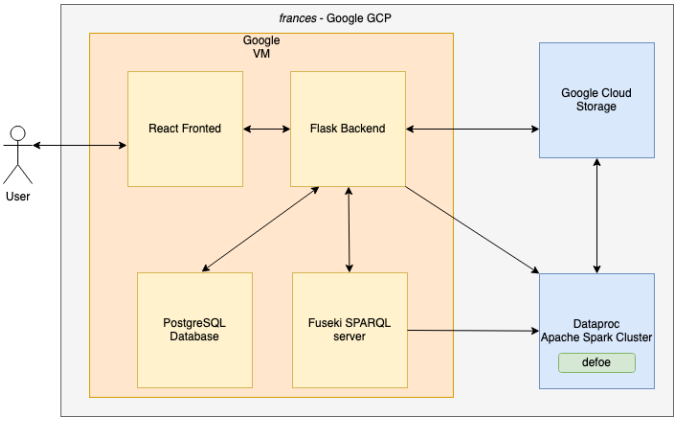


Fig. 4: frances architecture. Color-code: Yellow for containers, blue for Google GCP products, orange for the VM.

To improve file management and query submission, frances leverages two containerized services. The first service is the *Fuseki SPARQL Server*, which enables efficient storage and retrieval of semantic data through SPARQL queries. This server plays a vital role in managing and querying the interconnected knowledge graphs created from the textual collections. The second service is the *PostgreSQL Database*, responsible for managing user account data and defoe query tasks. It ensures accurate storage of configuration data and query results, maintaining data integrity and confidentiality.

By utilizing these containerized services, frances enhances the overall functionality and security of the platform, providing users with a reliable and robust environment for their data management and analysis needs. In the following subsections, we will delve deeper into the specific components and functionalities of this architecture.

A. System Security Measures

As frances is deployed in a cloud environment for public use, it is essential to mitigate Distributed Denial of Service (DDoS) [15] attacks and ensure the availability of resources. To achieve this, ingress filtering is applied to the network of the Google Cloud VM instance hosting frances, allowing only the *React Frontend* and *Flask Backend* to be accessed by the public. Rate limiting technology is employed in the *Flask Backend* to restrict the number of requests an IP address or user can make within a certain time period. Additionally, user account systems are implemented, and defoe-related functions are made accessible only to authenticated users, ensuring secure access to the costly resource, defoe.

B. Data Storage Solution

To address the challenge of mixed files among multiple users and ensure accurate passing of configuration data and query results, a *PostgreSQL database* [19] is implemented. The database manages user account and defoe query details (e.g. state, progress, submit time, error message, result file path, etc). Google Cloud Storage [20] is used to store users files (e.g. lexicons) and query results, leveraging its scalability and security features. Note that separate folders are automatically generated in the Google Cloud for each user to store query files, providing secure storage and restricting user access to their respective folders.

C. Apache Spark Cluster with Dataproc

The Apache Spark cluster with *Google Cloud Dataproc* serves as a robust distributed computing platform, combining the advanced data processing and analytics capabilities of Apache Spark with the scalable and managed infrastructure provided by Google Cloud. In the context of frances, we harnessed the power of *Google Cloud Dataproc* to establish an Apache Spark cluster comprising 18 nodes. We configure this cluster with 1 master (N2-series, 8 VCPU, 32GB memory), and 18 workers (E2-series, 8 VCPU, 32 GB memory). Having at the end a cluster with with 152 VCPus (144 for the workers).

This cluster was purposefully designed and configured to handle the execution of defoe queries, which are complex text mining operations. The integration of this Apache Spark cluster with frances yielded significant improvements in the performance and scalability of defoe queries. By distributing the workload across multiple nodes, the cluster enabled parallel processing, enhancing the speed and effectiveness of the defoe query execution. Note that this cluster is not continuously active in frances. It is specifically activated upon submission of a defoe query and automatically deactivated after two hours of inactivity. The process of starting the cluster currently takes around five minutes, and it is seamlessly automated, requiring no action from the user.

D. Cloud Deployment

frances leverages the power of the Google Cloud Platform (GCP) [21], utilizing a Google Virtual Machine (VM) to deploy the four containers mentioned earlier. In addition, it makes use of *Google Cloud Storage* to store users files and results and the *Dataproc* Apache Spark cluster for running defoe queries. These choices are driven by several reasons. GCP offers a highly scalable infrastructure capable of handling large-scale textual collections, while *Dataroc* enables distributed processing and efficient resource utilization through Apache Spark. The managed services provided by GCP handle infrastructure setup and maintenance, freeing developers to focus on enhancing the frances platform. The reliability, performance, and global network infrastructure of GCP contribute to the consistent and fast operation of frances, making it a powerful tool.

[19] https://www.postgresql.org/
[20] https://cloud.google.com/storage
[21] https://cloud.google.com/

Note that frances has been designed in a way that makes it adaptable to other cloud providers as well. The modular architecture of frances ensures that the core functionalities and components can be easily integrated into alternative cloud environments. By following cloud-agnostic design principles, such as using containerization and adhering to open standards, frances can be deployed on different cloud platforms without significant modifications.

VIII. FRANCES USER INTERFACE

frances offers a user-friendly interface (UI) that provides automatic abstractions for interacting with various knowledge graphs, such as the *EB-Knowledge Graph* (see Section II-B), *ChapbooksScotland-KG*, *LadiesDebating-KG*, *GazetteersScotland* (see Section V). It incorporates deep learning analysis and defoe, allowing users to extract complex knowledge quickly and transparently, without requiring expertise in data science. The UI consists of five main sections displayed at Figure 5.

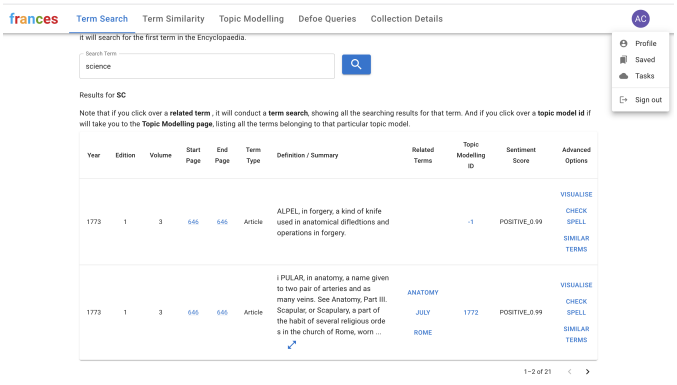


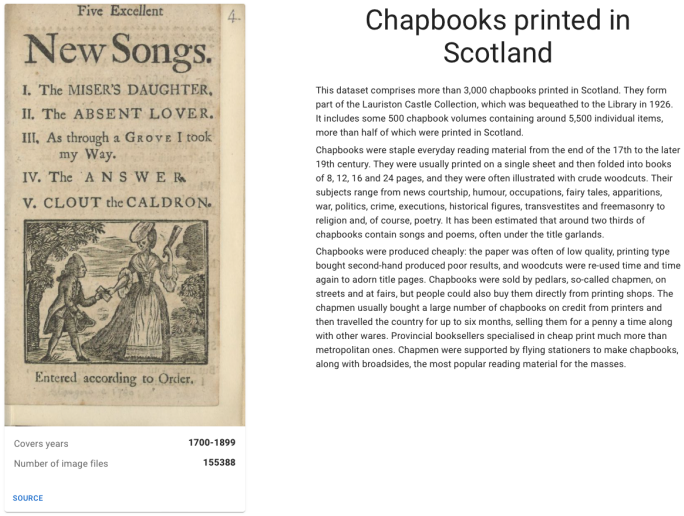
Fig. 5: frances new UI. Results for 'science' EB Term.

The 'Term Search' section enables users to search for specific terms within the EB. The 'Term Similarity' feature helps find similar terms based on an EB term or free text input applying. 'Topic Modelling' allows for clustering of EB terms (deep learning for topic modelling). The 'Defoe Queries' section allows users to select a collection and defoe query for analysis, with various query configuration options available. Lastly, the 'Collection Details' section provides comprehensive information about the collections.

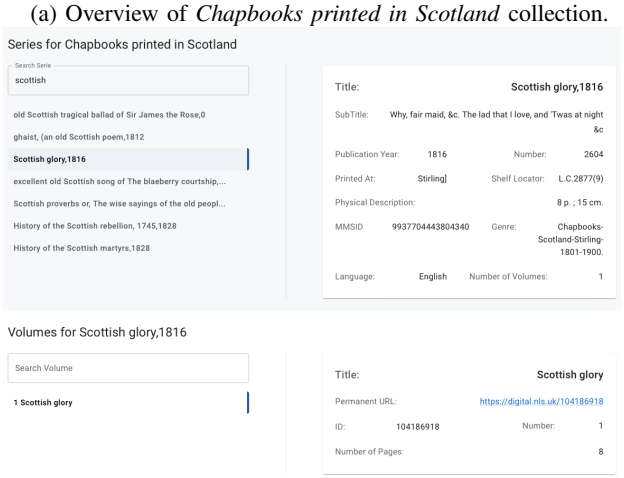
In order to enhance frances usability we have implemented substantial improvements across all sections of the UI. Additionally, we have introduced user login functionality, allowing users to create accounts and personalize their experience. Among the sections, the 'Defoe Queries' and 'Collection Details' sections have undergone significant transformations, which will be discussed in detail in the following subsections. It is important to note that detailed explanations of all deep learning analyses supported by frances can be found in our previous publication [4].

A. Collection Details

The 'Collection Details' section serves as an interactive interface between users and the various knowledge graphs supported by frances. It presents supported collections in the form of cards, featuring a cover picture, name, and year range for each collection. Users can click on a collection card to navigate to a dedicated page where they can access detailed information about the selected collection. This frances section enhances user interaction and facilitates exploration of the available knowledge graphs in the platform.



(a) Overview of *Chapbooks printed in Scotland* collection.



(b) Series & Volumes details of *Chapbooks printed in Scotland*.

Fig. 6: Details of *Chapbooks printed in Scotland*.

Figure 6a demonstrates the details page for *Chapbooks printed in Scotland* collection, which is internally referred to as the *ChapbooksScotland-KG*. This page highlights a comprehensive description of the collection, along with its metadata and a link to its source in the National Library of Scotland. Furthermore, the page presents the metadata of its series and volumes, which is retrieved from the *ChapbooksScotland-KG* through the API backend application. Figure 6b provides a closer look at the details of one the books of this collection. In particular, the *Scottish Glory* series from 1816, along with its sole volume. To enhance accessibility, the page also includes a search function for both series and volumes, allowing users to easily locate specific series or volumes by name, with the scrollable search results for improved efficiency.

B. Defoe Queries

The 'Defoe Queries' section empowers users to configure and submit queries for supported collections, as discussed in Section VI. This section is exclusively available to authenticated users for security purposes, while other sections are open

to all users. Figure 7 displays the ‘defoe query submit page’, where users can select different collections to query, a feature not available in the previous prototype. Users can also upload a lexicon (list of words or sentences), and indicate several filtering options such as target words (those must appear in terms/page to select them) and year range.

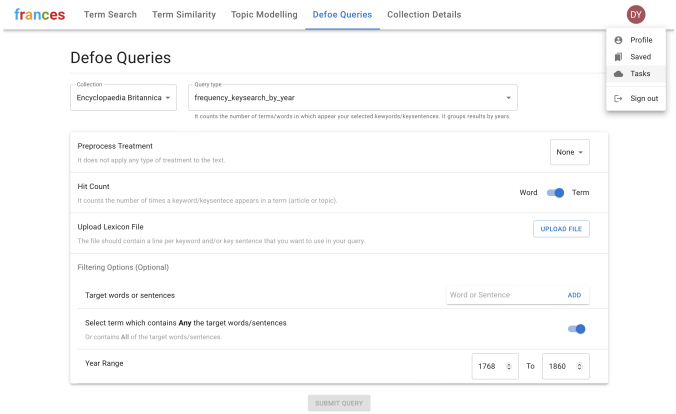


Fig. 7: ‘defoe query submit page’ when EB collection and *frequency_keysearch_by_year* are selected.

The configuration options dynamically change based on the selected collections and queries. For instance, when the EB collection is chosen, users can query the frequency in either texts or term definitions by selecting ‘word’ or ‘term’ for the hit count option. Conversely, if other NLS collections are selected, the hit count option offers ‘word’ and ‘page’ choices for querying frequency in texts or pages. When selecting the *geoparser_by_year* query, the gazetteer [22] and bounding box (to select a place coordenates) options are enabled. Figure 8 presents the dialog for bounding box editing, allowing users to either manually fill the box or search for a bounding place.

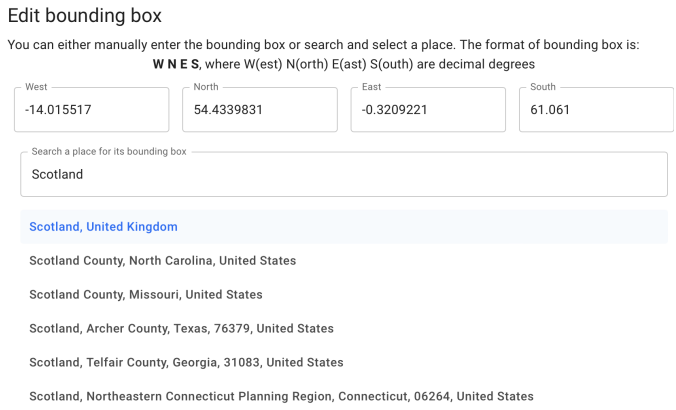


Fig. 8: Bounding box editing dialog. Searching for ‘Scotland’.

Figure 9 displays the defoe query tasks page, in which all submitted queries are listed along with their configuration data, execution state and submission time. Users can view the result of a previous submitted query by clicking the *view* button of that task result in this page. Furthermore, for enhanced user convenience, a link to this page is readily available within the user menu component located in the page header (refer to the top right corner of Figure 7) as well as on a query result page. These improvements in the UI of frances enhance the user experience by providing a more intuitive and interactive platform for exploring collections, conducting defoe queries, and visualizing data. The enhanced usability and expanded functionalities cater to the diverse needs of researchers, enabling them to gain valuable insights from the vast historical collections available.

[22]Users can select among several online gazetteers as the source authority against which to ground placenames.

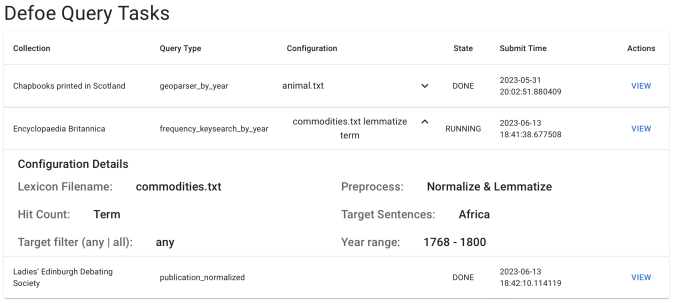


Fig. 9: defoe query tasks page

IX. Evaluation

In this section, we evaluated the performance of defoe with different Spark properties. The assessment was conducted on the *Dataproc* Apache Spark cluster introduced in section VII-C, consisting of 18 worker nodes with a total of 144 virtual cores allocated to the workers. We focused on two key properties: *spark.cores.max*, which determines the maximum number of CPU cores, and *spark.executor.instances*, which defines the number of executors [7]. We measured the total time to run a *frequency_keysearch_by_year* query with 18 different combinations of these two properties. This query counts the number of terms in EB collection where appear eight commodities (‘tobbaco’, ‘rum’, ‘indigo’, ‘sugar’, ‘coffee’, ‘rice’, ‘cotton’, ‘molasses’) without applying any filtering options but applying normalization and lemmatization prepossessing treatments. The query mentioned here is the initial one employed in the Case Study X-A to generate Figure 10, and the corresponding results can be found in Table I.

cores	4	8	16	32	64	128	144	288	300
executors	1	1	2	4	8	16	18	18	18
times (s)	962	996	940	904	620	404	433	422	449
cores	4	8	16	32	64	128	**144**	288	300
executors	2	2	4	8	16	32	**36**	36	36
times (s)	985	970	946	623	426	331	**297**	308	300

TABLE I: Evaluation with different spark properties.

This evaluation demonstrates a reduction in processing time as the number of executors increases substantially, particularly when it reaches or surpasses 8, aligning closely with the number of worker nodes. However, the maximum number of CPU cores has minimal impact on the observed results. In this case, this query runs about 70% faster when the number of executor increases from 1 to 36. This finding emphasizes the positive effect of scaling up the number of executors on overall processing efficiency. Therefore, frances has adopted the 144 cores and 36 executors for the mentioned spark properties.

X. Case studies

A. Commodity and Enslaved Labor in the EB

This case study explores the depiction of race, slavery, and commodities produced with enslaved labor in the *Encyclopaedia Britannica* (EB) during a period of intellectual and industrial development. By utilizing the `frances` platform, the study conducts a diachronic analysis of references to these topics, with a specific focus on cotton. Previous analyses have primarily examined individual entries, but this study takes a broader approach by investigating the occurrences of commodities across different editions of the EB [16]. Digital research methods are employed due to the extensive textual data involved.

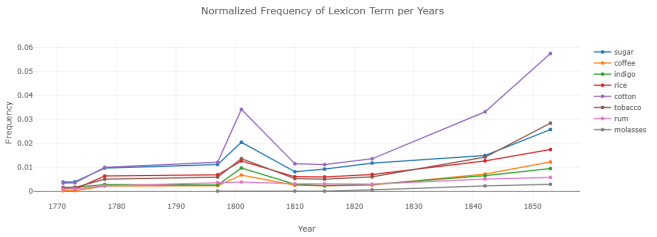
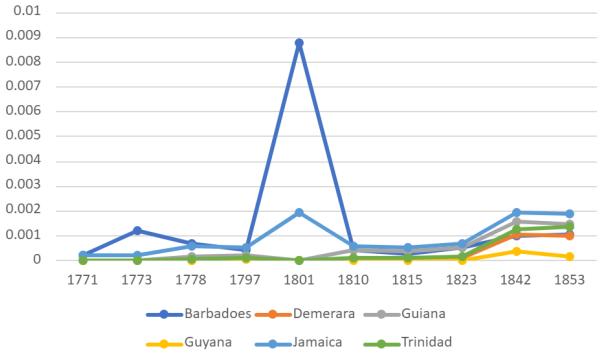


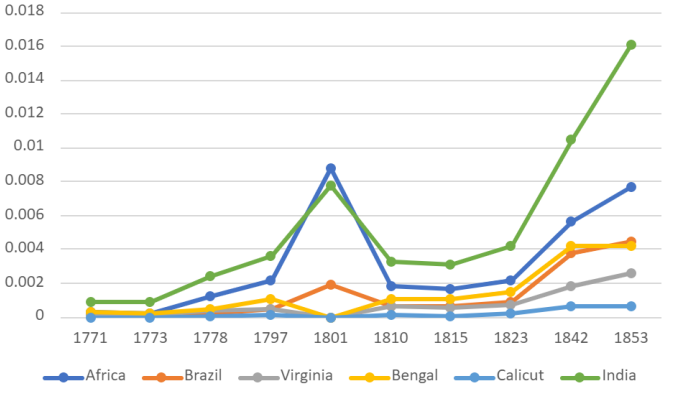
Fig. 10: N-gram to visualise the normalized frequencies of eight commodities: 'tobacco', 'rum', 'indigo', 'sugar', 'coffee', 'rice', 'cotton' and 'molasses', produced in `frances`.

Using `frances`, we executed the `frequency_keysearch_by_year defoe` query (introduced in Section VI) to analyze the frequencies of eight commodity types associated with slavery in the *Encyclopaedia Britannica* (EB). This query identified occurrences of each commodity in each term of the EB, counting them as hits. Figure 10 presents the normalized frequency results, obtained by dividing the total quantity of terms containing a commodity per year by the total quantity of terms per year. The analysis uncovered shifts in the frequency of commodity mentions over time, with 'cotton' surpassing 'sugar' as the most referenced commodity after 1801. The notable increase in terms mentioning cotton can be attributed to Britain's advancements in cotton technologies during the late eighteenth and nineteenth centuries, facilitating mass production and manufacturing [17].

Furthermore, `frances` facilitated in-depth analysis by examining the occurrence of commodities in conjunction with geographic locations. Twelve `frequency_keysearch_by_year` queries were executed using different target words representing place names, filtering out irrelevant terms and returning commodity frequency counts. Those queries were submitted in parallel taking an average of five minutes to get the results of each of them. While cotton remained a frequently mentioned commodity in EB terms related to non-Caribbean areas as displayed in Figure 11b, its frequencies stagnated or decreased in articles focusing on Caribbean areas (see Figure 11a), contradicting the overall trend. However, a significant increase in references to cotton was observed in articles referencing India, and these numbers sharply spiked after 1823, likely due to the international trading activities of Britain's East India Company who were trading commodities



(a) Occurrences of 'cotton' in EB terms mentioning Caribbean locations: 'Barbadoes', 'Demerara', 'Guiana', 'Guyana', 'Jamaica' and 'Trinidad'.



(b) Occurrences of 'cotton' in EB terms mentioning non-Caribbean places: 'Africa', 'Brazil', 'Virginia', 'Bengal', 'Calicut', 'India'.

Fig. 11: N-grams visualising the normalized frequency of lexicon term per year with data exported from `frances`

internationally throughout this period. This shift suggests a transition from cotton production in the Caribbean to stronger associations with regions like the United States [18] and South Asia. The decline in West Indian cotton production and supply at the turn of the nineteenth century due to crop disease and detrimental weather, the comparative growth in sugar trading [17], as well as rising anti-slavery sentiment within Britain and towards its colonies in the early nineteenth century likely contributed here. This shift in the seventh and eighth (1842, 1853) editions took place after the Slavery Abolition Act 1833 was passed, enforcing gradual abolition within the British Empire, although notably this did not include territories owned by the East India Company and may be a contributing factor to the patterns seen here. The study demonstrates the comprehensive analysis capabilities of `frances`, shedding light on the dynamics of slavery and commodity production in the *Encyclopaedia Britannica*. Further use of `frances` will explore additional aspects of the EB, contributing to a deeper understanding of the information environment during the Enlightenment and Victorian periods, including the examination of racialized language and its implications.

B. Geoparsing the Chapbooks printed in Scotland

Geoparsing the extensive collection of over 3,000 *Chapbooks printed in Scotland* during the 18th and 19th centuries

holds immense value for historians, as these chapbooks serve as invaluable cultural artifacts offering insights into the social, economic, and historical contexts of that era. Through the process of geoparsing, historians can delve deeper into the geographic distribution and spatial references within the texts, shedding light on regional language variations, customs, folklore, and the dissemination of ideas and popular culture across different areas of Scotland.

To prioritize the significance of Scottish locations within this collection, we have instructed `frances` to assign them higher weighting during the geo-resolution step using the Bounding Box (see Figure 8). The configuration details and the obtained results of the submitted query are depicted in Figures 12a and 12b, and the advanced visualizations (Figures 12c, 12d, and 12e) are automatically generated by `frances`, readily available for download at any time, alongside the results. By geoparsing the chapbooks using `frances`, historians gain the ability to unravel the spatial dimensions of Scottish literature from the 18th and 19th centuries.

XI. Related Work

Digital textual analysis and the development of platforms for analyzing large-scale textual collections have been active areas of research in the fields of Digital Humanities and computational analysis. Several works and technologies have laid the foundation for the advancements in `frances`.

Text analysis frameworks such as NLTK [19] (Natural Language Toolkit), GATE (General Architecture for Text Engineering) [20], and Apache Tika [21] have facilitated the exploration and analysis of textual data. These frameworks have influenced the enhancement of `frances`.

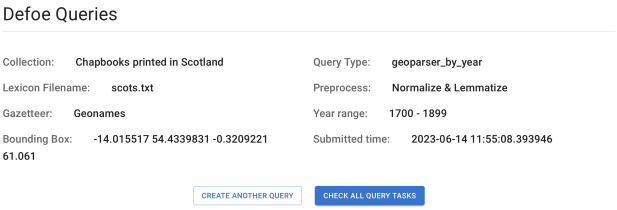
The construction of knowledge graphs from textual data has been a significant focus. Works like DBpedia [23], YAGO [22], and Freebase [23] have demonstrated the value of representing structured information extracted from unstructured text. The introduction of the *NLS Ontology* and the creation of knowledge graphs for NLS collections build upon these efforts.

Cloud-based text mining has revolutionized the field by providing scalable and cost-effective infrastructure for processing large-scale textual collections. Technologies like Apache Hadoop [24] and Apache Spark [7] enable distributed processing and analysis of big data. Leveraging Apache Spark and the cloud-based *Google Dataproc* [24] service, `frances` harnesses the power of distributed computing for efficient text mining.

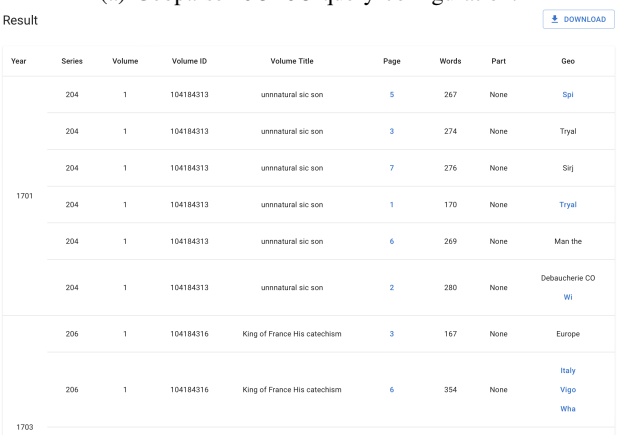
Semantic web and linked data principles have enhanced the accessibility and interoperability of structured data. Ontologies like FOAF [25] (Friend of a Friend) and DBpedia Ontology [25] enable the integration of heterogeneous data sources. Linking the NLS collections with external ontologies, such as FRBRoo [26], DBpedia, and EOL [27], enriches the semantic richness of knowledge graphs.

Curatr [28] is an online platform designed for exploring and curating the British Library corpus. It leverages natural language processing and text mining methods to facilitate tasks such as text search, volume recommendation, semantic search, and sub-corpus curation. *frances* offers several advantages over Curatr, supporting a wider range of collections and utilizing the *NLS Ontology* to create knowledge graphs, and it enables

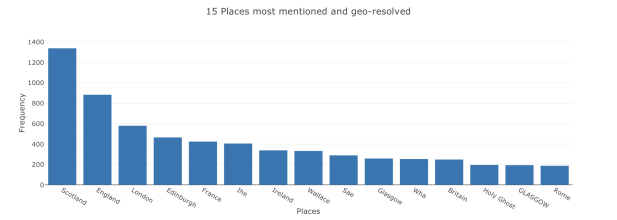
[23] https://dbpedia.org
[24] https://cloud.google.com/dataproc?hl=en
[25] https://dbpedia.org/ontology/



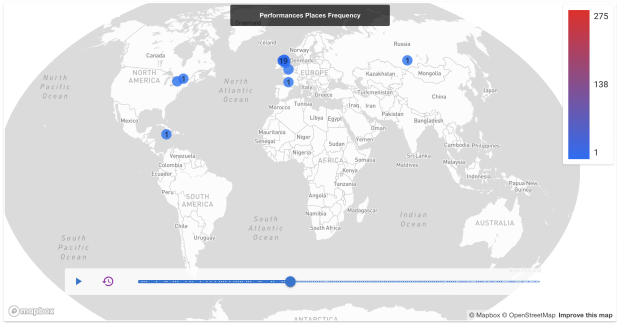
(a) Geoparser `defoe` query configuration.



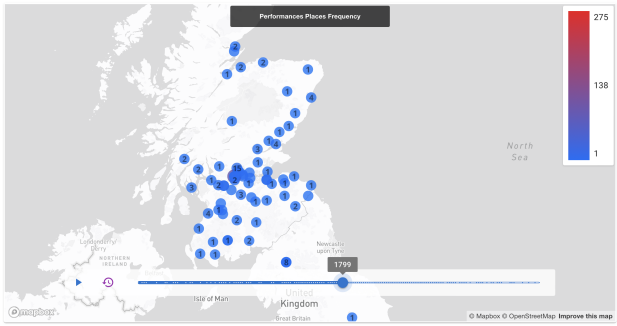
(b) Frequency results sorted by year of publication.



(c) 15 Places most mentioned in this collection.



(d) Map of georesolved places frequencies over time.



(e) Zoom in 'Scotland'. Frequency of geolocated places in 1799.

Fig. 12: Geoparsing the *Chapbooks printed in Scotland*. The visualizations are automatically generated by `frances`.

efficient and scalable text mining through an Apache Spark cluster. This research extends prior works and contributes to the growing knowledge in digital textual analysis.

XII. Conclusion and Future Work

This paper introduces significant advancements to the `frances` platform, which contribute to the ongoing progress in analyzing digital historical textual collections. The cloud-based architecture ensures maintainability, efficient deployment, and scalability. The *NLS Ontology* standardizes the representation of NLS Data Foundry collections, enabling the creation of knowledge graphs that capture intricate relationships within the collections.

`frances` has evolved into a cloud-based digital platform, utilizing an Apache Spark cluster hosted in the cloud for text mining queries. This enhancement enhances scalability and performance, enabling researchers to efficiently analyze large-scale textual collections. The development of novel `defoe` queries and modifications to existing ones equip researchers with a diverse toolkit for extracting valuable insights and conducting profound analyses. The case studies showcases the power of `frances` in uncovering insights within vast datasets. Researchers used `frances` to analyze commodities related to slavery, revealing shifting patterns of commodity references over time, as well to Geoparsing collections. Future improvements include expanding the creation of knowledge graphs to include additional collections, collaborating with other institutions, and conducting extensive evaluation and validation. Optimization of scalability and performance, integration with external linked data sources, and gathering user feedback are also important areas of focus. Overall, the advancements presented in this paper open up new possibilities for analyzing digital textual collections, inspiring further research in the fields of Digital Humanities, History and eScience.

References

[1] B. McGillivray, T. Poibeau, and P. R. Fabo, "Digital humanities and natural language processing: "je t'aime... moi nonplus"," *DHQ: Digital Humanities Quarterly*, vol. 14, no. 2, 2020. [Online]. Available: https://example.com

[2] N. Pedrazzini and B. McGillivray, "Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers," in *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*. Taipei, Taiwan: Association for Computational Linguistics, Nov. 2022, pp. 85–95. [Online]. Available: https://aclanthology.org/2022.nlp4dh-1.12

[3] A. Hawkins, "Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web," *Archival Science*, 2021. [Online]. Available: https://doi.org/10.1007/s10502-021-09381-0

[4] R. Filgueira, "frances: a deep learning nlp and text mining web tool to unlock historical digital collections," in *2022 IEEE 18th International Conference on eScience*. IEEE, Jul. 2022, funding: This work was supported by the NLS Digital Fellowship and by the Google Cloud Platform research credit program.; 18th IEEE International eScience Conference (eScience 2022), eScience 2022 ; Conference date: 10-10-2022 Through 14-10-2022.

[5] R. Filgueira Vicente, M. Jackson, A. Roubickova, A. Krause, R. Ahnert, T. Hauswedell, J. Nyhan, D. Beavan, T. Hobson, M. Coll Ardanuy, G. Colavizza, J. Hetherington, and M. Terras, "defoe: A spark-based toolbox for analysing digital historical textual data," in *2019 IEEE 15th International Conference on e-Science (e-Science)*. United States: Institute of Electrical and Electronics Engineers (IEEE), Mar. 2020, pp. 235–242, 2019 IEEE 15th International Conference on e-Science (e-Science), e-Science 2019 ; Conference date: 24-09-2019 Through 27-09-2019. [Online]. Available: https://escience2019.sdsc.edu/

[6] R. Filgueira, C. Grover, V. Karaiskos, B. Alex, S. Van Eyndhoven, L. Gotthard, and M. Terras, "Extending defoe for the efficient analysis of historical texts at scale," in *2021 IEEE 17th International Conference on eScience (eScience)*. IEEE, Oct. 2021, pp. 21–29, iEEE eScience 2021 - 17th IEEE eScience 2021 International Conference, eScience 2021 ; Conference date: 20-09-2021 Through 23-09-2021. [Online]. Available: https://www.escience2021.org

[7] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, and I. Stoica, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, p. 56–65, oct 2016. [Online]. Available: https://doi.org/10.1145/2934664

[8] S. Ames, "Transparency, provenance and collections as data: The national library of scotland's data foundry," *LIBER Quarterly: The Journal of the Association of European Research Libraries*, vol. 31, no. 1, p. 1–13, Feb. 2021. [Online]. Available: https://liberquarterly.eu/article/view/10880

[9] S. C. Feria, "Web ontology language (owl)," https://chowlk.linkeddata.es.

[10] D. G. Verdejo, "Wizard for documenting ontologies (widoco)," https://github.com/dgarijo/Widoco.

[11] R. Filgueira, "ChapbooksScotland-KG: A Knowlege Graph for representing the "Chapbooks Printed In Scotland" (1671 - 1893)," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6673995

[12] ——, "LadiesDebating-KG: A Knowlege Graph for representing the "Edinburgh Ladies' Debating Society Digital Collection" (1865 - 1880)," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6686596

[13] ——, "GazetteersScotland-KG: A Knowlege Graph for representing the Gazetteers of Scotland (1803-1901)," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6686829

[14] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball, "Use of the Edinburgh Geoparser for georeferencing digitized historical collections," *Philosophical Transactions of the Royal Society A*, vol. 368, no. 1925, pp. 3875–3889, 2010. [Online]. Available: https://doi.org/10.1098/rsta.2010.0149

[15] R. V. Deshmukh and K. K. Devadkar, "Understanding ddos attack & its effect in cloud environment," *Procedia Computer Science*, vol. 49, pp. 202–210, 2015, proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915007541

[16] S. Sebastiani, "Conjectural history vs. the bible: Eighteenth-century scottish historians and the idea of history in the encyclopaedia britannica," *Lumen*, vol. 21, pp. 213–231, 2002. [Online]. Available: https://doi.org/10.7202/1012276ar

[17] B. Schoen, *The Fragile Fabric of Union: Cotton, Federal Politics, and the Global Origins of the Civil War*, ser. Studies in Early American Economy and Society from the Library Company of Philadelphia. New Haven: Yale University Press, 2011.

[18] G. Riello, *Cotton: The Fabric that Made the Modern World*. Cambridge University Press, 2020.

[19] E. Loper and S. Bird, "Nltk: The natural language toolkit," 2002. [Online]. Available: https://arxiv.org/abs/cs/0205028

[20] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, pp. 223–254, 2002. [Online]. Available: https://doi.org/10.1023/A:1014348124664

[21] The Apache Software Foundation, "Apache tika," https://tika.apache.org, 2021, accessed: June 13, 2023.

[22] YAGO, "Yago," https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/, 2021, accessed: June 13, 2023.

[23] Freebase, "Freebase," https://developers.google.com/freebase, 2021, accessed: June 13, 2023.

[24] Apache Software Foundation, "Apache hadoop," https://hadoop.apache.org/, 2021, accessed: June 13, 2023.

[25] D. Brickley and L. Miller, "FOAF Vocabulary Specification," FOAF Project, Namespace Document 2 Sept 2004, http://xmlns.com/foaf/0.1/.

[26] *FRBR - object-oriented definition and mapping to FRBRer*, 1st ed., 2009, may 2009. [Online]. Available: http://cidoc.ics.forth.gr/docs/frbr_oo/frbr_docs/FRBRoo_V1.0_draft__2009_may_.pdf

[27] E. Pafilis *et al.*, "Environments and eol: identification of environment ontology terms in text and the annotation of the encyclopedia of life," *Bioinformatics*, vol. 31, no. 11, pp. 1872–1874, 2015.

[28] D. Greene, K. Wade, S. Leavy, and G. Meaney, "Curatr: A platform for exploring and curating historical text corpora," in *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, Riga, Latvia, October 21-23, 2020*, ser. CEUR Workshop Proceedings, S. Reinsone, I. Skadina, A. Baklane, and J. Daugavietis, Eds., vol. 2612. CEUR-WS.org, 2020, pp. 247–253. [Online]. Available: http://ceur-ws.org/Vol-2612/short9.pdf