frances: A Deep Learning NLP and Text Mining Web Tool to Unlock Historical Digital Collections

A Case Study on the Encyclopaedia Britannica

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

***Abstract*—This work presents *frances*, an integrated text mining tool that combines information extraction, knowledge graphs, NLP, deep learning, parallel processing and Semantic Web techniques to unlock the full value of historical digital textual collections, offering new capabilities for researchers to use powerful analysis methods without being distracted by the technology and middleware details. To demonstrate these capabilities, we use the first eight editions of the Encyclopaedia Britannica offered by the National Library of Scotland (NLS) as an example digital collection to mine and analyse. We have developed novel parallel heuristics to extract terms from the original collection (alongside metadata), which provides a mix of unstructured and semi-structured input data, and populated a new knowledge graph with this information. Our Natural Language Processing models enable *frances* to perform advanced analyses that go significantly beyond simple search using the information stored in the knowledge graph. Furthermore, *frances* also allows for creating and running complex text mining analyses at scale. Our results show that the novel computational techniques developed within *frances* provide a vehicle for researchers to formalize and connect findings and insights derived from the analysis of large-scale digital corpora such as the Encyclopaedia Britannica.**

***Index Terms*—information extraction, knowledge graphs, deep transfer learning, natural language processing, text mining, web tools, semantic web, parallel computing, digital tools, historical digital textual collections**

I. Introduction

The increasing availability of digital collections of historical texts presents a wealth of opportunities for advancing historical, cultural, and linguistics research. However, the scale and heterogeneity of these collections raises significant challenges when researchers attempt to find, link, and extract relevant concepts and their semantic relationships or perform text mining analyses that go beyond simple search and retrieval [1].

The National Library of Scotland (NLS) Data Foundry [1] offers a wide range of historical digital collections of textual resources that have the potential to provide an invaluable resource for historians, humanities, and computational linguistics researchers. One of those digital collections is the Encyclopaedia Britannica (EB), issued from 1768-1860. As is the case with most digital historical texts, its contents are provided in XML files derived from scanned manuscripts using Optical Character Recognition (OCR). The EB was the most authoritative general reference work of (part of) the eighteen, nineteenth, and much of the twentieth century and is the only encyclopaedia in any language to survive this 250-year period. It has long been used by researchers to document changes in individual concepts over time, since it provides evidence for when a concept could be called 'widely accepted'. But this data has much more to tell us than what happened to individual concepts – its continuity provides us with a unique opportunity to explore the broader question of how the structure of knowledge changed, and it allows us to compare different editions and identify patterns in its transformation.

While modern text mining and machine learning methods are available that could enable a much wider range of analyses to reveal useful information for digital humanities research, no tools are presently available that would enable researchers to apply these to the EB with ease. To address this shortcoming, we have developed *frances*, a novel web tool that enables researchers to accelerate the process of discovering insights from the EB without being distracted by the technology and middleware details.

This work also involved the automated extraction of EB terms (along with their metadata) across editions. To this end we employed *defoe* [2], [3], a Spark-based parallel processing library for analysing and mining textual datasets. Then, we created *EB-ontology* to represent the relations and properties between different editions, volumes, pages and terms, and used this ontology along with the extracted information to create the *EB Knowledge Graph* (*EB-KG*) in order to make the encyclopaedia searchable and analyzable. Later, we augmented the *EB-KG* by using transformer-based deep neural network language models.

frances interacts with the *EB-KG*, runs advanced NLP analyses (e.g. searches, term similarity, spell checking, etc) and submits *defoe* text mining queries, providing the results back to users. Although we have used for this work the EB, this could be extended to analyse other large digital collections with minor adaptations to the underlying codebase.

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

The remainder of the paper is structured as follows. Section II presents background on *defoe*. Section III introduces our parallel extraction heuristics. Section IV details the features of the *EB-ontology* and *EB-Knowledge Graph*. Section V explains how we have employed deep learning NLP-transformer models to augment the knowlege extracted from the encyclopaedia. Section VI presents the improvements performed to *defoe* along with a new set of *defoe* queries to mine the encyclopaedia. Section VII introduces the main features of the *frances* web tool. Finally, section VIII describes related work, and section IX concludes with a summary of achievements and future work.

II. Background

This section provides an overview of our previous work on *defoe* [2], [3] to introduce the necessary background for the functionalities presented in Sections II, VI and VII-E.

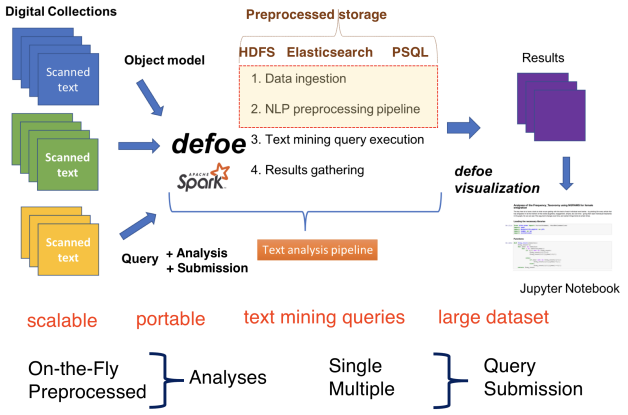


Fig. 1: Overview of *defoe*

defoe is a parallel Python library for analysing textual data. It allows for extracting knowledge from historical texts by running text mining queries in parallel via the Apache Spark framework [4] and storing the pre-processed data (for further queries) using several storage solutions, such as an HDFS file system, an ElasticSearch distributed engine or a PostgreSQL database (see Figure 1). *defoe* is able to extract, transform and load digital collections that comprise several XML schemata and physical representations. More specifically, *defoe* has four object models (`PAPERS`, `NZPP`, `ALTO`, `NLS`) to map the physical representations and XML schemas. `NLS` is the object model most relevant to this work, since it supports the ingestion of NLS digital collections including the Encyclopaedia Britannica.

defoe includes pre-processing techniques to mitigate against OCR errors and other issues such as long-S and line-break hyphenation, and to standardise the representation of the text. It has geoparsing capabilities [5], and is able to run single or multiple queries at once across digital collections.

All *defoe* text mining queries are based on a number of operations (`filter`, `flatMap`, `map`, `reduce`, etc) that are combined to perform text mining analyses. Figure 2 shows an implementation example, the `total_pages` query, in which a `flatMap operation` is applied to an archive to return the list of documents it contains (e.g. volumes, books). For each document, the `map operation` extracts the number of pages, gathering the total number of documents (volumes, books) and the total number of pages within those. Figure 3 shows the results running this query using the ten volumes of the Second Edition of the EB.

```
def do_query(archives, config_file=None, logger=None, \
                                                context=None):
    # [archive, archive, ...]
    documents = archives.flatMap(lambda archive:\
                                 list(archive))
    # [num_pages, num_pages, ...]
    num_pages = documents.map(lambda document: \
                                   document.num_pages)
    result = [documents.count(), num_pages.reduce(add)]
    return {"num_volumes": result[0],\
            "num_pages": result[1]}
```

Fig. 2: *defoe* `total_pages query`: Iterates through archives and counts the total number of documents (e.g. volume, book etc) and total number of pages.

```
Result:
num_volumes: 10
num_pages: 9448
```

Fig. 3: *defoe* `total_pages query` results using the ten volumes of the Second Edition of the encyclopaedia. This archive compromises ten documents, one per volume.

In previous work, *defoe* used the command line as its interface, meaning that users had to submit their queries via a computer terminal. As described in Section VII-E, in this work we have created a new web user interface to interact with *defoe* to increase its usability.

III. Extracting EB Terms and Metadata

The EB collection[2] comprises of eight editions and a total of 195 volumes with a total size of 44GB. It uses two XMLs schemata: METS [3] for descriptive, structural, technical and administrative metadata (Title, Author, Publisher, etc); and ALTO [4] for encoding the OCR text of a page. Therefore, each volume has a METS file describing different metadata information, and has one image file and ALTO file per page attached to it (see Figure 4). These make up a total of 195 METS files, 155,388 ALTO files, and 155,388 image files.

Given that ALTO files do not indicate the start and end of each EB term, the first part of our work involved the automated extraction of all terms (along with their metadata) across editions, so they can be analysed independently without the surrounding text. To this end, we developed a new set of information extraction heuristics encoded as *defoe* queries [5] for the `NLS` object model (see Section II). These extract

[2]https://data.nls.uk/data/digitised-collections/encyclopaedia-britannica/
[3]http://www.loc.gov/standards/mets/
[4]https://www.loc.gov/standards/alto/
[5]https://github.com/francesNLP/defoe/blob/master/defoe/nlsArticles/queries/

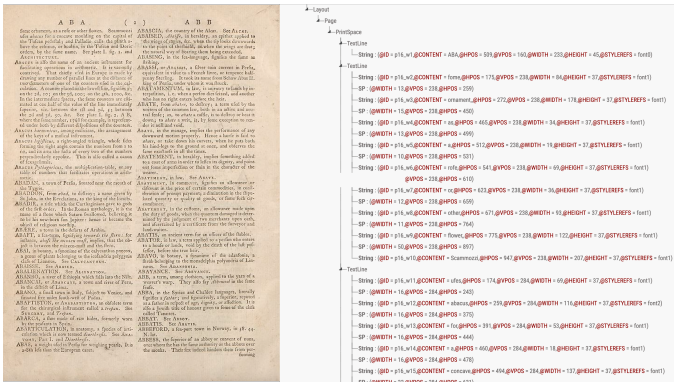
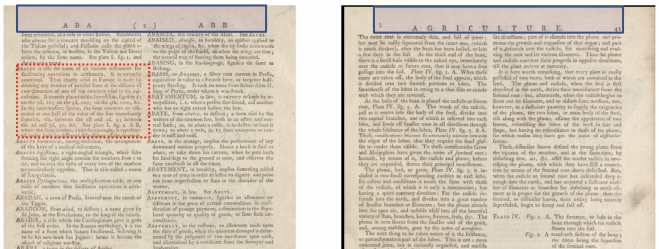


Fig. 4: Page 2 of the First Volume of the First Edition and its corresponding ALTO file. ALTO files contain a `TextLine` section to describe features per lines, which contain a `String` section to describe features per word within a line (content, position, etc).

structured information from the unstructured text available on ALTO files in parallel, making use of page headers and text patterns to classify terms between:

- *Articles*: Usually presented by a term in the main text in upper case followed by a "," (e.g. ABACUS,) and then a description of the term in one- or two-paragraph long text (similar to an entry in a dictionary). The headers of pages containing *Articles* have the first three letters of all *Articles* within each page (see left image in Figure 5).
- *Topics*: In this case, the encyclopaedia introduces a term (e.g. AGRICULTURE) in the header of a page (see right image in Figure 5). A *Topic* is typically described across several pages, often combining text, pictures, and tables.



Articles Topic

Fig. 5: *Article* (Page 2) and *Topic* (Page 41) page examples. Page 2 contains *Articles* that starts with the letters *ABA*, such as *Abacus* (highlighted in red), whereas Page 41 contains the start of the description for the *Agriculture Topic*; we have highlighted the headers in blue.

Our heuristics adapt to variations of page layouts, text patterns and headers across different editions. They also capture several definitions for the same term within each edition, in case those are available. Figure 6 shows the information extracted for the term *Science* from the First Edition issued in 1771.

```
term                                                          SCIENCE
definition          in philolophy, denotes any dpdfrine, deduced f...
relatedTerms                                                       []
header                                                         SCISCO
startsAt                                                          658
endsAt                                                            658
numberOfTerms                                                      24
numberOfWords                                                      15
numberOfPages                                                     872
positionPage                                                        7
typeTerm                                                      Article
editionTitle          First edition, 1771, Volume 3, M-Z
editionNum                                                          1
supplementTitle
supplementsTo                                                      []
year                                                             1771
place                                                      Edinburgh
volumeTitle         Encyclopaedia Britannica; or, A dictionary of ...
volumeNum                                                           3
letters                                                          M-Z
part                                                               0
altoXML                                  144133903/alto/144812443.34.xml
Name: 7454, dtype: object
```

Fig. 6: *Science* term information extracted from the First Edition

This work also involved the parallel extraction of edition and volume metadata based on the semi-structured information available on METS files. Figure 7 shows (part of) the information extracted for the volumes of the First Edition. Note that MMSID column refers to the Metadata Management System ID [6]. For each volume, we also added the permanent URL [7] for which the images of their pages can be visualized.

	MMSID	editionTitle	editor	editor_date	genre	language	termsOfAddress	numberOfPages	physicalDescription	place	...
0	992277653804341	First edition, 1771, Volume 1, A-B	Smellie, William	1740-1795	encyclopedia	eng	None	832	3 v., 160 plates : ill. ; 26 cm. (4to)	Edinburgh	... h
1	992277653804341	First edition, 1771, Volume 2, C-L	Smellie, William	1740-1795	encyclopedia	eng	None	1018	3 v., 160 plates : ill. ; 26 cm. (4to)	Edinburgh	... h
2	992277653804341	First edition, 1771, Volume 3, M-Z	Smellie, William	1740-1795	encyclopedia	eng	None	872	3 v., 160 plates : ill. ; 26 cm. (4to)	Edinburgh	... h
3	9929192893804340	First edition, 1773, Volume 1, A-B	Smellie, William	1740-1795	encyclopedia	eng	None	844	3 v. (viii, 697, [1] p., LVIII leaves of plate...	London	... h
4	9929192893804340	First edition, 1773, Volume 2, C-L	Smellie, William	1740-1795	encyclopedia	eng	None	1032	3 v. (viii, 697, [1] p., LVIII leaves of plate...	London	... h
5	9929192893804340	First edition, 1773, Volume 3, M-Z	Smellie, William	1740-1795	encyclopedia	eng	None	864	3 v. (viii, 697, [1] p., LVIII leaves of plate...	London	... h

Fig. 7: Subset of metadata extracted for the volumes of the First Edition. This edition is a 3-volume reference work, issued twice, in 1771 and 1773.

IV. EB Ontology and Knowledge Graph

Since one of our aims is to capture a shareable and reusable knowledge representation of the Encylocplaedia Britannica, we created the *EB Ontology* [8] (see Figure 8). The *EB Ontology* is a formal description of knowledge as a set of concepts

[6]The MMSID can be 8 to 19 digits long (with the first two digits referring to the record type and the last four digits referring to a unique identifier for the institution)

[7]Permanent URL for the First Volume of the First Edition (year 1771): https://digital.nls.uk/144133901

[8]https://w3id.org/eb/

(editions, volume, person, organizations, terms, etc) within the Encylocplaedia Britannica domain and the relationships that hold between them (publisher, startsAt, related terms, etc).

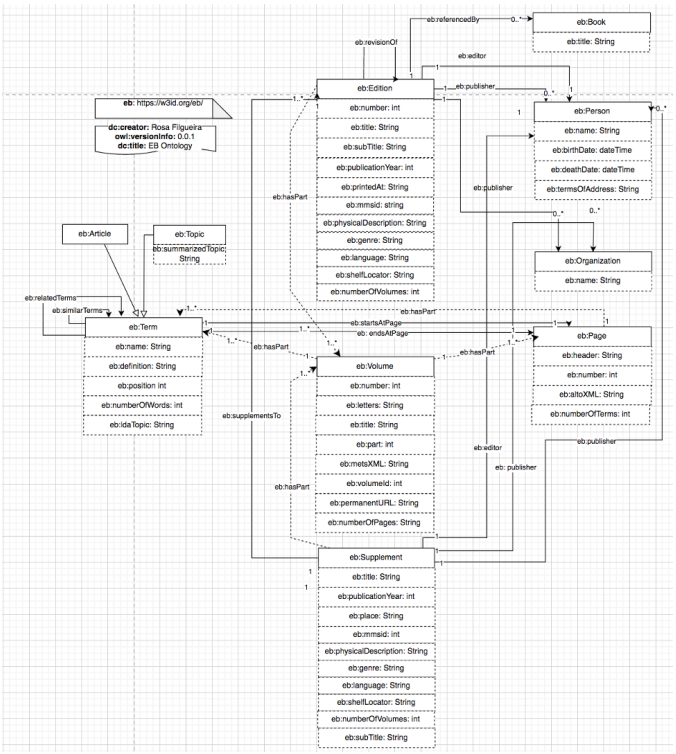


Fig. 8: Data Model of the *EB Ontology*.

The *EB Ontology* creation involved several phases. Taking into account all the information extracted during the first part of this work (see Section III), we first created an UML diagram to represent the conceptualization of the encyclopaedia. As shown in Figure 8, an *Edition* can have several *Volumes*, references to *Books*, *Supplements*; it also has an *Editor* and a *Publisher*, which can be a *Person* or an *Organization*. A *Volume* has several *Pages*, which can contain several *Terms*. And a *Term* can be either a *Topic* or an *Article*. In a second step, we converted the UML into an OWL ontology [9] using Chowlk [6]. After refining that ontology, we employed Widoco [7] to publish and create an enriched and customized documentation of the *EB ontology*, and configured the permanent identifier for this ontology using w3id.org service.

Finally, using the *EB Ontology* and the extracted information from the encyclopaedia, we created the first version of the *EB Knowledge Graph* (*EB-KG*), which contains 1,638,239 RDF [8] triples, and stored it into an *Apache Fuseki server* [10]. Each term, edition, page, volume, etc is a resource in our Knowledge Graph, and therefore has its own URI. Figure 9 shows a visualization of *EB-KG* terms. Storing our *EB-KG* in *Apache Fuseki* enables us to query the graph using SPARQL [11] (See Figure 10), a semantic query language to retrieve and manipulate RDF triples.

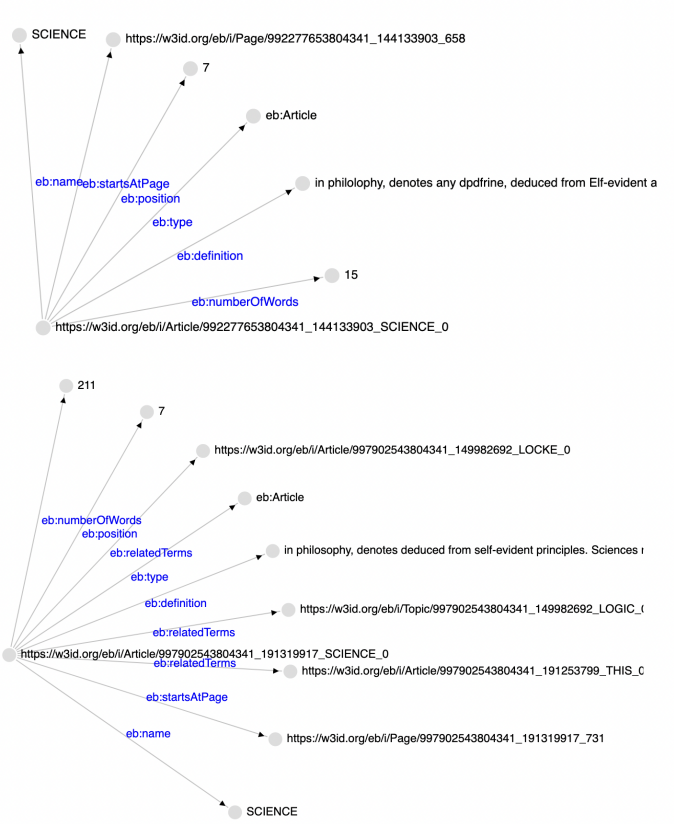


Fig. 9: Two representations of the *Science* term. The image at the top shows the information extracted for this term from the First Edition issued in 1771, while the other shows the information extracted from the Third Edition.

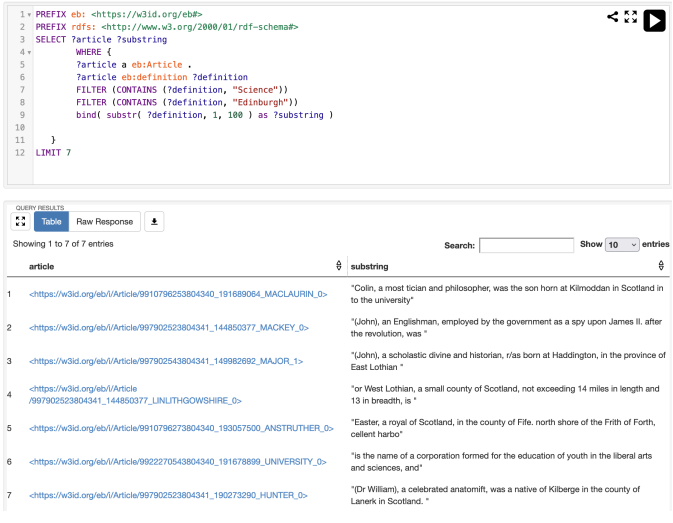


Fig. 10: A SPARQL query to retrieve the first seven *Articles* that contain the terms 'Science' and 'Edinburgh' in their definitions. For each definition, we show the first 100 characters.

[9]https://www.w3.org/OWL/
[10]https://jena.apache.org/documentation/fuseki2/
[11]https://www.w3.org/TR/rdf-sparql-query/

V. Augmenting Knowledge with Deep Transfer learning for NLP

We augmented the knowledge previously extracted of the encyclopaedia with several advanced NLP and deep learning analyses: 1) classifying terms into categories expressing positive or negative attitudes (*sentiment analysis*); 2) clustering terms into topics (using *Latent Dirichlet Allocation (LDA) topic modelling*); 3) finding semantically similar terms (*term similarity*); 4) fixing OCR errors that frequently occur when applying automated text recognition to historical works (*spell checking*); and 5) producing shorter/more accessible representations of the historical text (*summarisation*).

For those analyses, we employed deep transfer learning, an approach where knowledge is transferred from one model to another [9], by making use of pre-trained transformer-based models such as *BERT* [10] or *GPT-3* [11].

In this work, we have used the state-of the art sentence transformer model, called *all-mpnet-base-v2* [12] and the *SentenceTransformers* framework [12] to train our terms embeddings (*eb_embeddings_model*). These use all terms definitions available in the *EB-KG* to capture their semantic information.

Furthermore, we have also employed additional transformer models for our advanced NLP analyses [13]:

1) *Sentiment analysis*: we used *RoBERTa-large* [13] model, which has been pretrained on a large corpus of English data in a self-supervised fashion.
2) *LDA Topic modelling*: we employed *BerTopic* [14], which is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing to produce easily interpretable topics whilst keeping important words in the topic descriptions.
3) *Semantic textual similarity*: we used embeddings of our terms (*eb_embeddings_model*) to compute cosine similarity between them, and compare all terms embeddings against all other terms, returning a list with the pairs sorted by their cosine similarity score.
4) *Spell checking*: We employed *neuspell* [15], a neural spelling correction library, and chose the *Elmosclstm-Checker* pre-trained neural model as the checker to use for our work. Figure 11 shows an example of how this analysis performs using the *Instrument* term definition from the First Edition issued in 1771.
5) *Summarization*: We used the *XLNet* [16] pre-trained model, which is an improved version of the *BERT* model that implements permutation language modeling in its architecture. For this task we also employed the *Bert Extractive Summarizer library* [14]. Note that we have only employed *summarization* for EB Topics definitions.

These transformers were selected after an extensive evaluation in which we compared the performance of different pre-trained models and various configurations. The selected models were trained and run using a Google Cloud Debian VM configured with 32 vCPUs, 120 GB memory and 4 NVIDIA Tesla P4 GPUs. The results of these analyses themselves were stored in the *EB-KG*, augmenting the knowledge stored in graph. Sections VII-A to VII-D detail the use of these results across several *frances* facilities.

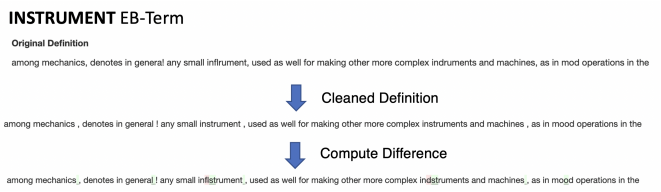


Fig. 11: Top: original definition of the *Instrument* term stored in the *EB-KG*; middle: cleaned version computed using *neuspell*; bottom:difference between the two.

VI. Enabling defoe to query the EB-KG

As described in Section II, initially *defoe* was only able to either ingest digital collections directly from XML files by indicating the appropriate object model (e.g. `NLS` object model for the encyclopaedia), or from pre-processed data provided through different storage solutions. However, it did not include support for querying SPARQL knowledge graphs.

In this work we have implemented a new SPARQL connector to make use of the full potential of *defoe* and perform further text mining analyses using our *EB-KG* as a data source. Furthermore, we have created a new set of *defoe* queries [15] for this connector:

- `frequency_keyseach_by_year`: calculates the frequencies of one or several keywords or key sentences in terms definitions, applying different pre-processing techniques (normalization, lemmatization, stemming). Results are grouped by year.
- `publication_normalized`: counts the total number of volumes, pages and words of the encyclopaedia and returns results per year.
- `terms_fulltext_keysearch_by_year`: searches and extracts full text definitions according different filtering settings. This query also allows us to apply different pre-processing techniques. Results are grouped by year.
- `terms_snippet_keysearch_by_year`: similar to the previous query, but instead returns snippets of text definitions. It allows for snippet size configuration.
- `uris_keysearch`: extracts URIs of terms that contain the selected keywords or key sentences in their definitions. It uses different pre-processing techniques and results are grouped by URI.
- `geoparser_terms`: geo-locates locations in terms definitions and geo-resolves them using the Edinburgh Geoparser [17].

These new *defoe* queries are fully configurable, which was not the case for previous *defoe* queries developed in earlier

[12]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[13]Our deep learning NLP models with their configurations are available from https://github.com/francesNLP/frances/tree/main/NLS_EB/frances_nlp_scripts

[14]https://github.com/dmmiller612/bert-extractive-summarizer

[15]https://github.com/francesNLP/defoe/tree/master/defoe/sparql/queries

work. They allow us to capture different configurations by choosing filtering options, target, lexicon, period of time, hit count, etc. These can be run across the different encyclopaedia editions in parallel, processing the information further stored in the *EB-KG*. Figure 12 shows an overview of how *defoe* interacts with the *EB-KG* to run a given *defoe* query.

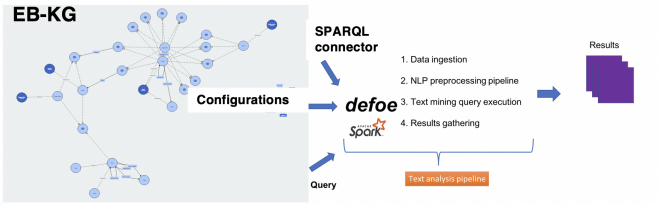


Fig. 12: Overview of *defoe* interacting with the *EB-KG*

Furthermore, in this work we have created a new web user interface to interact with *defoe* and the *EB-KG*. It enables users to select, configure and run *defoe* queries (from the ones listed above). This is described in detail in Section VII-E.

VII. Frances Web Tool

Finally, to unlock the full value of the Encyclopaedia Britannica, we created *frances* (see Figure 13), a novel Flask-based web application [16] that interacts with the *EB-KG* to a) extract information from the encyclopaedia (using SPARQL in the backend) and display the desired information and b) perform further text mining analyses (using *defoe* as a backend) on the *EB-KG* providing the results back to users, as well as visualizing them.

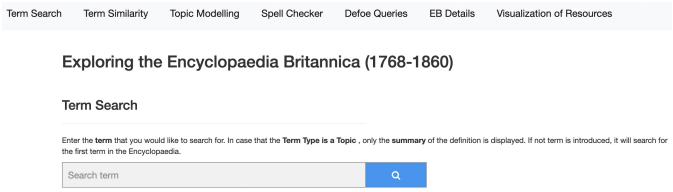


Fig. 13: *frances* enables us to explore the encyclopaedia with a variety of deep learning NLP and text-mining facilities

frances provides automatic abstractions for interacting with the Knowledge Graph (see Section IV), deep learning NLP analyses (see Section V), and *defoe* (see Section VI), so that users can extract complex knowledge from the encyclopaedia in a fast and transparent manner without having to be an expert data scientist. These abstractions are offered as *frances* facilities, which are described in the following subsections.

A. Term Search

frances allows users to search for terms across different editions of the encyclopaedia and obtain their definitions, metadata information, and advanced NLP analyses results (previously calculated). Figure 14 shows the results of searching for *Flower*. The results are displayed in a table with a row per term-resource found in the *EB-KG*. In this example, nine *Flower* term-resources were found displaying two at the time.

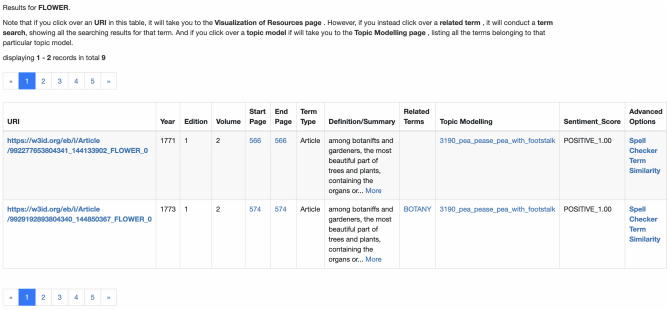


Fig. 14: Results obtained when we search for *Flower*. This term has been found nine times across the eight editions (note that the First Edition was issued twice, in 1771 and 1773).

The metadata information includes the year, edition, volume, and start and end page of each match. The start and end pages contain links to the URLs where their images are permanently stored in the NLS collections, enabling a visual validation of the information extracted by the heuristics introduced in Section III.

These results also show each term-resource classification (either *Article* or *Topic*) across editions. For *Topics*, the text shown corresponds to the *summarization* analysis of their definitions. Also, whenever our extraction heuristics detect related terms within each definition, those are also shown in this table (e.g. *BOTANY* in Figure 14). Clicking on any of the related terms triggers a new *term search* using that particular related term as the new search term.

Furthermore, results include links to other *frances* facilities, such as *LDA topic modelling*, *spell checking*, and *term similarity*. They also display their *sentiment analysis* (and score). Note that all NLP analyses have been previously calculated (and stored) by applying different deep learning transformer models (see Section V).

Finally, results include URIs for each term-resource match. Clicking them allows us to visualize their information stored in our *EB-KG*. This is explained further in Section VII.G.

B. Term Similarity

frances allows us to search for semantic textual similarity of terms as previously calculated (see Section V) by applying cosine similarity to term definitions.

This facility allow us to provide the URI of a term-resource for which we want to find the most similar terms, or we can click on any of the *term similarity* links that we get when we search for terms as shown in Figure 14. Figure 15 shows the most similar terms for *Flower* as defined in the First Edition (issued in 1773). The results are sorted by similitude rank, displaying the 20 most similar term-resources in a table. In this table, we get the URI of each similar term-resource and some metadata (edition, year, volume), along with the term name, definition, *LDA topic modelling* results and similitude rank.

[16] Our Source code available from https://github.com/francesNLP/frances/tree/main/web-app

URI	Edtion	Year	Volume	Term	Definition	Topic Modelling	Similitud Rank
https://w3id.org/eb/i/Article /992277653804341_144133902_FLOWER_0	1	1771	2	FLOWER	among botanifts and gardeners, the most beautiful part of trees and plants, containing the organs or... More	3190_pea_pease_pea_with_footstalk	0.9776949
https://w3id.org/eb/i/Article /997902543804341_149981189_FLOWER_0	3	1797	7	FLOWER	Flos, among botanifts and gardeners, the most beautiful part of trees and plants, containing the org... More	-1_he_his_was_in	0.81424284
https://w3id.org/eb/i/Article /9910796253804340_192015836_FLOWER_0	6	1823	8	FLOWER	Flos, among botanists and gardeners, the most beautiful part of trees and plants, containing the org... More	-1_he_his_was_in	0.8098287
https://w3id.org/eb/i/Article /9922270543804340_191678898_FLOWER_0	5	1815	8	FLOWER	Flos, among botanifts and gardeners, the molt beautiful part of trees and plants, containing the org... More	-1_he_his_was_in	0.8073469
https://w3id.org/eb/i/Article /9929192893804340_144850367_LILIADEOUS_0	1	1773	2	LILIADEOUS	an appellation given to such flowers as resemble that of the lily.	3798_mint_horehound_horehound_and_in_rings_the_lily	0.704466
https://w3id.org/eb/i/Article /992277653804341_144133902_LILIADEOUS_0	1	1771	2	LILIADEOUS	an appellation given to such flowers as resemble that of the lily.	3798_mint_horehound_horehound_and_in_rings_the_lily	0.704466
https://w3id.org/eb/i/Article /9910796253804340_192692756_CALYX_0	6	1823	5	CALYX	among botanists, a general term, expressing the cup of a flower*, or that part of a plant which surr... More	226_botanifts_among_botanifts_of_flower_petals	0.68875444
https://w3id.org/eb/i/Article /997902523804341_144850377_LILIACEOUS_0	2	1778	6	LILIACEOUS	in botany, an appellation given to such flowers as referable those of the lily.	3798_mint_horehound_horehound_and_in_rings_the_lily	0.68756175

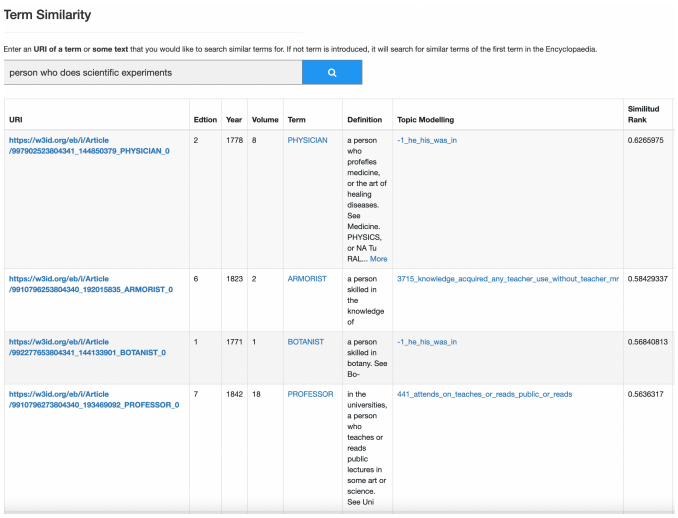
Fig. 15: Similarity results for the term-resource *Flower* from the First Edition; in this example we have opted to show just the top eight most similar term-resources.

This table includes links to another *frances* facilities. For example, if we click on a URI link, the information of that resource will be visualized. If we click on a term name link, it will perform a new *term similarity* search using that particular term-resource. And if we click on a *topic modelling* link, it will show which other term-resources belong to the same *LDA topic*. This is further described in Section VII.C.

frances also enables semantic similarity searches using free text. Figure 16 shows the results obtained when we search for *'person who does scientific experiments'*. This search calculates the first 20 most similar term-resources (in the example shown in Figure 16 we opted to display just the first four). In this case, the semantic similarity is calculated at the time of the search, applying the same methodology as that described in Section V. *frances* first calculates the sentence embedding for the free text query and then calculates the cosine similarity between its embedding and those of all other terms, returning a list with the pairs sorted by their cosine similarity score.

C. Topic Modelling

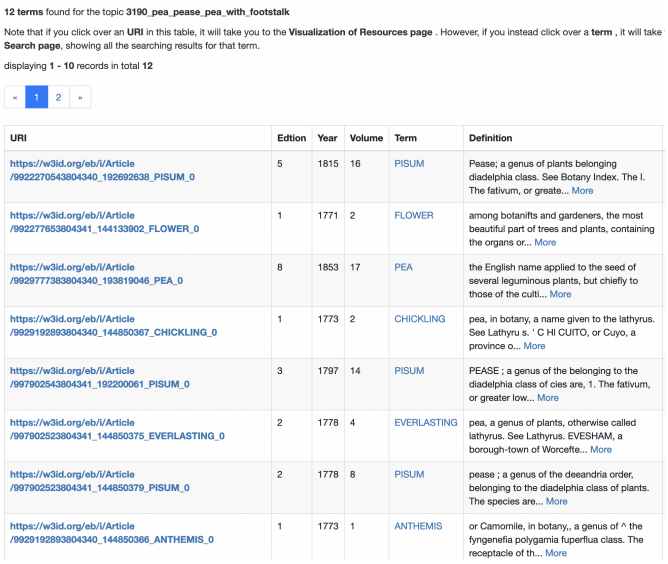
Another functionality of *frances* is that it enables us to visualize all term-resources that have been previously clustered together applying *LDA Topic Modelling* (see Section V), displaying the results in a table. In this table, we obtain the URI of each term-resource and some metadata (edition, year, volume), along with the term name and definition. Clicking on any of the term names will result in a new *term search* as shown in Section VII.A, while clicking on any URI will



URI	Edtion	Year	Volume	Term	Definition	Topic Modelling	Similitud Rank
https://w3id.org/eb/i/Article /997902523804341_144850379_PHYSICIAN_0	2	1778	8	PHYSICIAN	a person who profefies medicine, or the art of healing diseases. See Medicine. PHYSICS, or NA Tu RAL... More	-1_he_his_was_in	0.6265975
https://w3id.org/eb/i/Article /9910796253804340_192015835_ARMORIST_0	6	1823	2	ARMORIST	a person skilled in the knowledge of	3715_knowledge_acquired_any_teacher_use_without_teacher_mr	0.58429337
https://w3id.org/eb/i/Article /992277653804341_144133901_BOTANIST_0	1	1771	1	BOTANIST	a person skilled in botany. See Bo-	-1_he_his_was_in	0.56840813
https://w3id.org/eb/i/Article /9910796273804340_193469092_PROFESSOR_0	7	1842	18	PROFESSOR	in the universities, a person who teaches or reads public lectures in some art or science. See Uni	441_attends_on_teaches_or_reads_public_or_reads	0.5636317

Fig. 16: Similarity results for the free-text: *person who does scientific experiments*.

visualize the information of those resources stored in our *EB-KG*.



URI	Edtion	Year	Volume	Term	Definition
https://w3id.org/eb/i/Article /9922270543804340_192692638_PISUM_0	5	1815	16	PISUM	Pease; a genus of plants belonging diadelphia class. See Botany Index. The I. The fativum, or greate... More
https://w3id.org/eb/i/Article /992277653804341_144133902_FLOWER_0	1	1771	2	FLOWER	among botanifts and gardeners, the most beautiful part of trees and plants, containing the organs or... More
https://w3id.org/eb/i/Article /9929777383804340_193819046_PEA_0	8	1853	17	PEA	the English name applied to the seed of several leguminous plants, but chiefly to those of the culti... More
https://w3id.org/eb/i/Article /9929192893804340_144850367_CHICKLING_0	1	1773	2	CHICKLING	pea, in botany, a name given to the lathyrus. See Lathyru s. ' C HI CUITO, or Cuyo, a province o... More
https://w3id.org/eb/i/Article /997902543804341_192200061_PISUM_0	3	1797	14	PISUM	PEASE ; a genus of the belonging to the diadelphia class of cies are, 1. The fativum, or greater low... More
https://w3id.org/eb/i/Article /997902523804341_144850375_EVERLASTING_0	2	1778	4	EVERLASTING	pea, a genus of plants, otherwise called lathyrus. See Lathyrus. EVESHAM, a borough-town of Worcefte... More
https://w3id.org/eb/i/Article /997902523804341_144850379_PISUM_0	2	1778	8	PISUM	pease ; a genus of the deeandria order, belonging to the diadelphia class of plants. The species are... More
https://w3id.org/eb/i/Article /9929192893804340_144850366_ANTHEMIS_0	1	1773	1	ANTHEMIS	or Camomile, in botany,, a genus of ^ the fyngenefia polygamia fuperflua class. The receptacle of th... More

Fig. 17: Visualization of the first ten term-resources clustered in the *3190_pea_pease_pea_with_footstalk* topic modelling

We can either indicate the name of the *LDA Topic* to visualize, or click on any *topic modelling* links provided in either *term search* (introduced in Section VII.A) or *term similarity* (introduced in Section VII.B) results. Figure 17 shows the term-resources that belong to *3190_pea_pease_pea_with_footstalk LDA Topic*. Note that this *LDA Topic* corresponds to the results shown in Figure 14, after searching for the term *Flower*.

D. Spell Checking

When we perform a *term search* in *frances*, the results include a link to the *spell checker* facility. This enables us to check the pre-computed clean version of a term-resource

definition as shown in Figure 11. Furthermore, this facility also enables us to indicate the URI of a term-resource whose spelling we want to check.

E. Defoe Queries

As described in Section II, *defoe* is a Python library that allows for running text mining queries across large digital collections in parallel. *frances* provides a new web interface for *defoe* to mine the Encyclopaedia Britannica. Users can select, configure and run any of the new *defoe* text-mining queries introduced in Section VI, which use the new SPARQL connector to mine the *EB-KG*.

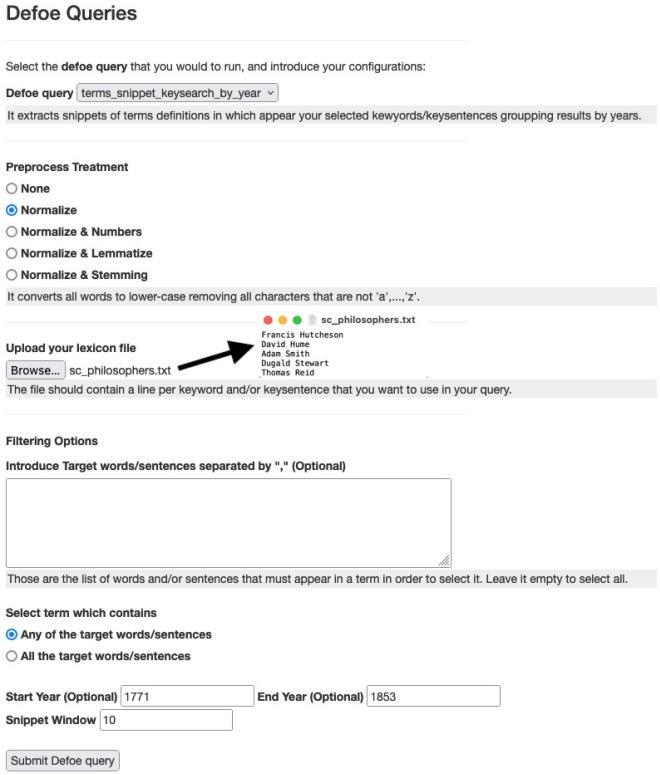


Fig. 18: Configuring the `terms_snippet_keysearch_by_year` query with the lexicon: *Francis Hutcheson, David Hume, Adam Smith, Dugald Stewart.*

Figure 18 shows an example in which we have selected the `terms_snippet_keysearch_by_year` query for retrieving snippets of texts (e.g. 10 words before and after each match) that contain any of the Scottish Philosophers stored in the 'sc_philosophers.txt' lexicon: *Francis Hutcheson, David Hume, Adam Smith, Dugald Stewart* and *Thomas Reid*. Lexicons should contain the list of keysearch words and/or sentences to use in queries. Furthermore, in this example, we selected `normalization` to pre-process the definition text, which converts all words to lower case removing all characters that are not 'a',...,'z'.

More filtering options are available to users, such as specifying an additional 'target' list of words/sentences that must appear in the text of a term's definition in order to "select it",

Fig. 19: The top image displays the results of the query configured in Figure 18, while the bottom image displays the frequency n-gram obtained after running the `frequency_keyseach_by_year` using the same configurations as in the previous query

which we did not use in this example, or selecting terms that are from a specific period of time (e.g. 1771 to 1853).

The image at the top in Figure 19 shows (part of) the results of running the previous *defoe* query. Results are sorted by year with one row per snippet along with further information, the keysearch word/sentence, edition number, volume number, page number, ALTO file, etc. These results are also available to users to download as a zip file so they can explore them further if they so wish. The image at the bottom of Figure 19 shows the frequency n-gram automatically obtained after running the `frequency_keyseach_by_year` using the Scottish Philosophers lexicon and the same filtering options as the ones in Figure 18. Additionally, this query returns the results in a table sorted by year, which can be downloaded, along with a second automatic frequency n-gram in which results are normalised by the number of terms per year.

F. EB Details

All the details regarding edition and volume metadata can be consulted in the *EB Details* facility. This allows us to select edition and volume, and *frances* performs a number of SPARQL queries in the backend to retrieve the pertinent information from the *EB-KG*. Figure 20 shows the details extracted for the First Volume of the Third Edition.

Among the details obtained, we have the edition-resource URI that we can click on to visualize the information available for this resource in our *EB-KG*, and the Volume Permanent URL (e.g. https://digital.nls.uk/190273291) which corresponds to the URL where NLS hold an online version of page images for each volume.

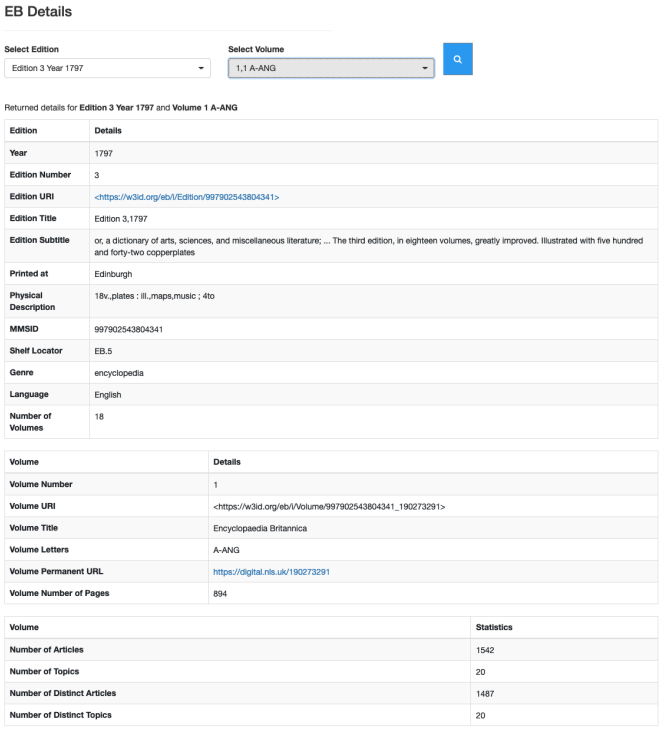


Fig. 20: Details of the Third Edition and its First Volume

Note that *EB Details* also provides statistics about the volume we are consulting. In our example, we can see that for this particular volume the *EB-KG* stores 1542 *Articles* (of which 1487 are distinct – without more than one definition per *Article*), and 20 *Topics*. This information would not be available without the information extraction heuristics developed for this work.

G. Visualizations of Resources

As described in previous subsections, *frances* allows us to visualize resources stored in our *EB-KG*. We can use this facility to visualize all the information about term-resources as shown in Figure 9, in which we visualize two *Science* term-resources from different editions. Alternatively, we can visualize the information about edition-resources as shown in Figure 21.

VIII. Related Work

A number of web tools have been developed for analysing historical digital collections in recent years. In this section we review those most relevant to our work. Curatr [18] is an online platform for the exploration and curation of historical digital books from the British Library. It provides facilities for creating n-grams and lexicon generation. Voyant-tools [17] is a web-based reading and analysis environment for digital texts. Google Books NGram Viewer [18] analyzes historical word occurrence, usage and changes over time and allows users to download data for more intensive research. The British

[17]https://voyant-tools.org/

[18]https://books.google.com/ngrams

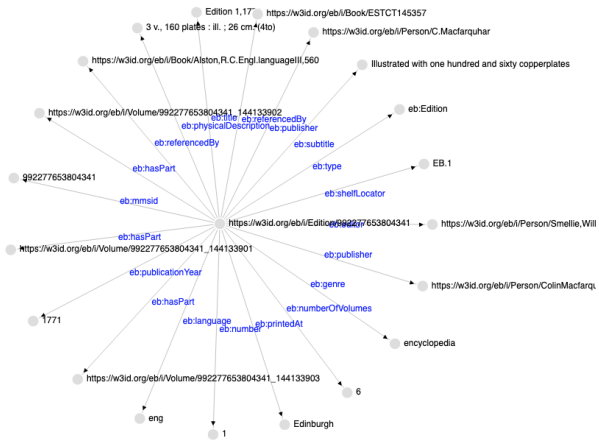


Fig. 21: Visualization of the First Edition (issued in 1771) resource stored in the *EB-KG*

Library also offers online tools [19] to search in their linked data instance of the British National Bibliography (BNB) [20].

IX. Conclusions and Future Work

In this paper we have presented *frances*, a new deep learning NLP and text mining powered web tool that enables researchers to analyse and extract knowledge from the Encyclopaedia Britannica with ease and explore how the encyclopaedia has changed over the years. In this work we have developed new parallel information extraction heuristics to extract, classify and structure terms across eight editions of the encyclopaedia. We have proposed a new ontology, *EB-Ontology*, that represents the information extracted from the encyclopaedia and its editions, volumes, pages and terms. Our approach combines knowledge graphs, deep transfer learning, parallel processing and Semantic Web techniques for formalizing and connecting findings and insights derived from the analysis of encyclopaedic corpora. Furthermore, we have enriched *defoe* by enabling it to mine knowledge graphs with a new set of configurable text mining analyses (*defoe* queries) and a new web interface for submitting, running and visualising analyses.

This research shows how deep learning NLP models and knowledge graphs can unlock the potential for using artificial intelligence to support digital humanities research, and to transform the ways in which we study and analyse historical textual collections.

Our immediate next step will be to make *frances* available as part of the tools offered by the NLS Data Foundry [21]. Furthermore, to account for differences between historical and current language use, we plan to experiment with NLP language models trained on a large historical dataset of books

[19]https://www.bl.uk/collection-metadata/metadata-services

[20]Available at http://bnb.data.bl.uk and http://bnb.data.bl.uk/sparql

[21]https://data.nls.uk/tools/

in English published between 1760-1900 [19], as well as those pre-trained on present-day language and adapt them to the historical domain, a technique that has shown promise results [20]. Also, in support of key tenets of linked open data, such as Findability, Accessibility, Interoperability, Repeatability (FAIR), in the near future the knowledge graph developed in this work (*EB-KG*) will be published under an open licensing scheme and connected to general-purpose knowledge bases such as Wikidata [22] or DBpedia [23].

Although we have used the Encyclopaedia Britannica for this work, in the future we plan to extended it to handle, mine and analyse other digital collections effectively (e.g. from the NLS or from the British Library), with minimum changes to incorporate the necessary information into the knowledge graph, or re-using the information from another existing knowledge graph.

X. Acknowledgements

This work was supported by the NLS Digital Fellowship [24] and by the Google Cloud Platform research credit program. The authors wish to thank, Daniel Garijo at the University Politécnica of Madrid, Sarah Ames and Ines Byrne at the NLS for their help and support during this work.

References

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

[2] 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, M. Terras, defoe: A spark-based toolbox for analysing digital historical textual data, in: 2019 IEEE 15th International Conference on e-Science (e-Science), Institute of Electrical and Electronics Engineers (IEEE), United States, 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. doi:10.1109/eScience.2019.00033.
URL https://escience2019.sdsc.edu/

[3] R. Filgueira, C. Grover, V. Karaiskos, B. Alex, S. Van Eyndhoven, L. Gotthard, M. Terras, Extending defoe for the efficient analysis of historical texts at scale, in: 2021 IEEE 17th International Conference on eScience (eScience), 2021, pp. 21–29. doi:10.1109/eScience51609.2021.00012.

[4] 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, I. Stoica, Apache spark: A unified engine for big data processing, Commun. ACM 59 (11) (2016) 56–65. doi:10.1145/2934664.

[5] R. Filgueira Vicente, C. Grover, M. Terras, B. Alex, Geoparsing the historical gazetteers of scotland: Accurately computing location in mass digitised texts, in: Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora, European Language Resources Association (ELRA), 2020, p. 24–30, 8th Workshop on the Challenges in the Management of Large Corpora, CMLC-8 ; Conference date: 16-05-2020 Through 16-05-2020.
URL http://corpora.ids-mannheim.de/cmlc-2020.html

[6] S. C. Feria, R. García-Castro, M. Poveda-Villalón, Converting UML-based ontology conceptualizations to OWL with chowlk, in: ESWC2021 Poster and Demo Track, 2021.
URL https://openreview.net/forum?id=u1Vp2y_QE1

[7] D. Garijo, WIDOCO: A wizard for documenting ontologies, in: C. d'Amato, M. Fernández, V. A. M. Tamma, F. Lécué, P. Cudré-Mauroux, J. F. Sequeda, C. Lange, J. Heflin (Eds.), The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II, Vol. 10588 of Lecture Notes in Computer Science, Springer, 2017, pp. 94–102. doi:10.1007/978-3-319-68204-4_9.
URL https://doi.org/10.1007/978-3-319-68204-4_9

[8] G. Klyne, J. J. Carroll, Resource description framework (rdf): Concepts and abstract syntax, W3C Recommendation (2004).
URL http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.
URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
URL https://aclanthology.org/N19-1423

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners (2020). arXiv:2005.14165.

[12] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
URL https://aclanthology.org/D19-1410

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, cite arxiv:1907.11692 (2019).
URL http://arxiv.org/abs/1907.11692

[14] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[15] S. M. Jayanthi, D. Pruthi, G. Neubig, Neuspell: A neural spelling correction toolkit, in: Conference on Empirical Methods in Natural Language Processing (EMNLP) Demo Track, Online, 2020.
URL https://arxiv.org/abs/2010.11085

[16] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, cite arxiv:1906.08237Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet (2019).
URL http://arxiv.org/abs/1906.08237

[17] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, J. Ball, Use of the Edinburgh Geoparser for georeferencing digitized historical collections, Philosophical Transactions of the Royal Society A 368 (1925) (2010) 3875–3889.
URL https://doi.org/10.1098/rsta.2010.0149

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

[19] K. Hosseini, K. Beelen, G. Colavizza, M. C. Ardanuy, Neural language models for nineteenth-century english, CoRR abs/2105.11321 (2021). arXiv:2105.11321.
URL https://arxiv.org/abs/2105.11321

[20] E. Manjavacas, L. Fonteyn, Adapting vs Pre-training Language Models for Historical Languages, working paper or preprint (Apr. 2022).
URL https://hal.inria.fr/hal-03592137

[22] https://www.wikidata.org/wiki/Wikidata:Main_Page

[23] https://www.dbpedia.org/

[24] https://www.nls.uk/using-the-library/academic-research/fellowships/digital-scholarship/