Curatr: A Platform for Exploring and Curating Historical Text Corpora

Derek Greene, Karen Wade, Susan Leavy, and Gerardine Meaney

University College Dublin, Ireland
{derek.greene,karen.wade,susan.leavy,gerardine.meaney}@ucd.ie

Abstract. The increasing availability of digital collections of historical texts presents a wealth of possibilities for new research in the humanities. However, the scale and heterogeneity of such collections raises significant challenges when researchers attempt to find and extract relevant content. This work describes *Curatr*, an online platform that incorporates domain expertise and methods from machine learning to support the exploration and curation of large historical corpora. We discuss the use of this platform in making the British Library Digital Corpus of 18th and 19th century books more accessible to humanities researchers.

1 Introduction

The selection, curation, and interpretation of text is fundamental to knowledge generation in the humanities [4, 12]. This work presents *Curatr*[1], a web-based platform to facilitate the exploration and curation of historical text corpora. The platform was initially designed to uncover and track biopolitical and medical humanities topics in 19th-century texts, particularly those related to disease and migration. However, the platform has broader potential as a tool for exploring large historical corpora in general. While the current project is based on a collaboration between literary studies and computer science, the interface and methodologies are of significant potential interest to digital social sciences. Specifically, the platform aims to support a workflow that addresses the requirements of humanities scholars who are increasingly working with large collections of unstructured text. and will need to curate specific sub-corpora from large digital collections. This work builds on a range of literature that demonstrates the requirement for flexible and open digital humanities platforms. In particular, close reading functionality to provide context is an essential aspect of humanities research, as evidenced in the provision of this kind of functionality along with quantitative analysis in previous systems [3, 9].

Through a collaboration with the British Library Labs[2], we have access to a large digital corpus from the British Library, covering 35,918 English language fiction and non-fiction books dating from 1700 to 1899. This is equivalent to approximately 12.3 million individual pages of printed text. A resource such as

[1] http://curatr.ucd.ie
[2] https://www.bl.uk/projects/british-library-labs

this offers a wealth of possibilities for conducting new research on a variety of topics in the humanities. However, the scale and diversity of such collections presents challenges when attempting to identify and extract relevant content, particularly for humanities scholars who are interested in studying highly-specific themes. Researchers working with the British Library corpus have previously attempted to curate smaller sub-corpora related to niche topics. This has often been a painstaking task, requiring considerable manual effort to inspect and filter the corpus. The *Curatr* platform is designed to address and mitigate these challenges, making the British Library corpus more accessible and useful to a wider audience.

2 Curatr Platform

2.1 Corpus Search and Filtering

The *Curatr* platform indexes all English-language books and associated metadata from the British Library Digital Corpus, allowing the collection to be browsed, searched, and filtered based on fields such as year, author, and publication location. The information retrieval component of the system makes use of the open source Apache Solr engine[3]. When indexing, we consider two types of document: 1) complete volumes; 2) short text segments from volumes of maximum length 2,000 characters. This allows researchers to search volumes in context or focus on shorter text snippets containing specific words or phrases. These documents can be searched and filtered based on a range of criteria, including both their textual content and the metadata of the associated volumes, using standard search operators. Fig. 1 depicts the central *Curatr* search interface, while Fig. 2 shows a corresponding set of search results. Clicking a search result brings up the close reading interface, displaying the metadata and content text for the associated volume or segment.

Curatr also makes use of a digitised version of the topical classification index of volumes used by the British Library from 1823 to 1985. Previously no useful digital version of the index existed. Using a combination of OCR and manual annotation, we extracted the two top levels of the hierarchy of this index. We then classified 98% of the English language texts from the British Library Digital Corpus, by linking the index pressmarks to the shelfmarks in the corpus metadata. By incorporating this information into the search interface, books in the corpus can be further filtered by categories such as "Fiction", "Drama", and "Geography". Many of these categories also contain more granular subcategories (*e.g.* "Geography" → "Ireland and Scotland"). This classification corresponds to the physical arrangement of books at the library. Combining search queries with the classification index allows researchers to uncover rare material for research and teaching. Once a smaller curated sub-corpus of texts has been identified, the associated texts and metadata can be easily exported to other platforms for

[3] `https://lucene.apache.org/solr`

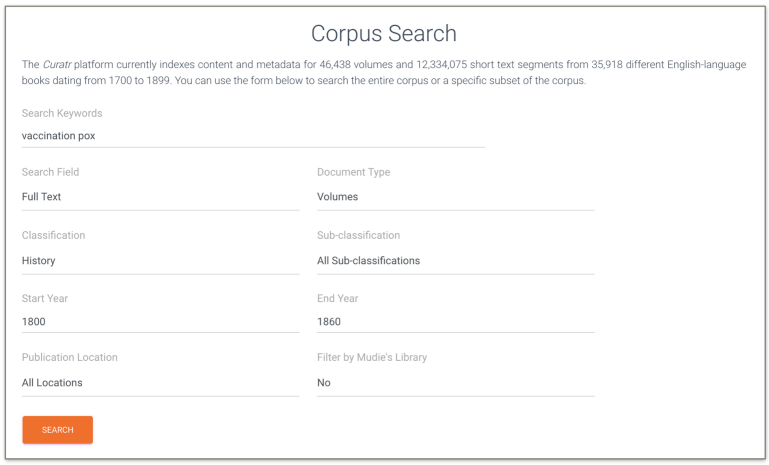

Fig. 1: *Curatr* corpus search interface.

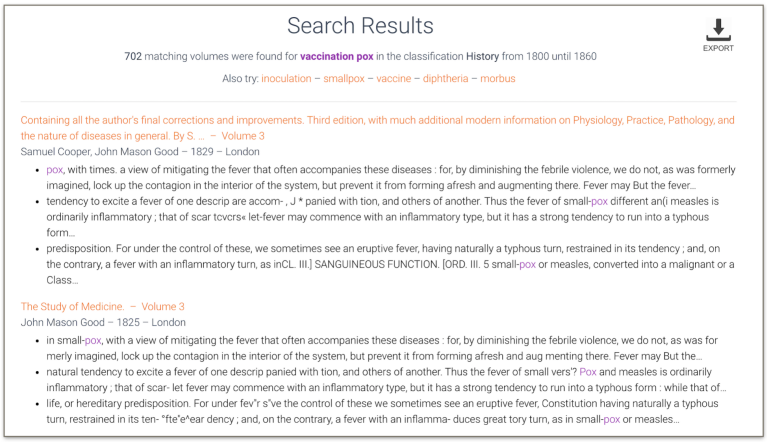

Fig. 2: *Curatr* search results for the query "vaccination pox". A list of matching volumes and other suggested search query words are shown.

further research and for close reading. Alternatively, the associated shelfmarks allow researchers to physically inspect the relevant books at the British Library.

An important step in the process of analysing a large dataset, such as the British Library corpus, involves understanding the representativeness and composition of its content. Therefore, we also integrate metadata extracted from catalogues originating from Mudie's Select Library [1], based on matches between catalogue entries and titles of books in the "Fiction" category of the British Library corpus. Comparisons can then be made between the contents of the corpus and those books which appear in the circulating library.

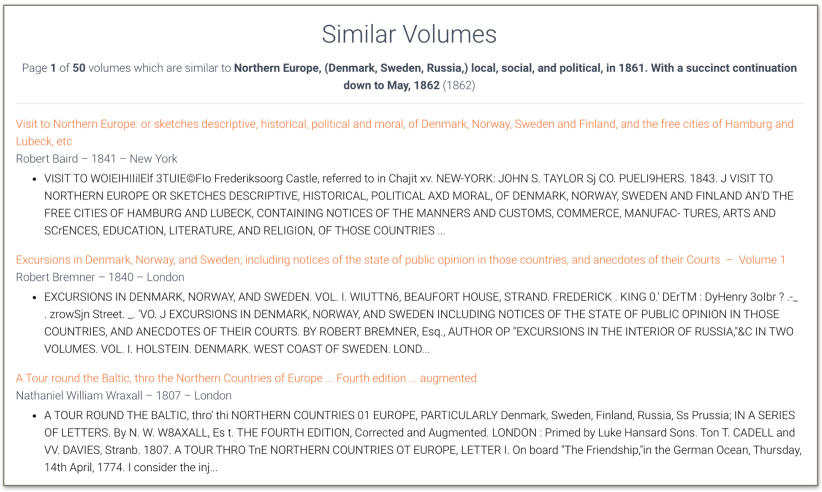

Fig. 3: Partial list of volume recommendation results generated by *Curatr* for an 1862 volume on the topography of Northern Europe.

When a relevant volume has been identified, we also help the researcher to find other similar texts, through the use of a *content-based recommendation system* [7]. We implement a custom recommender, where other volumes are ranked based on the similarity of their textual content to the text of the current volume. Similarity scores are determined based on the pairwise cosine similarity of the vectors for volumes in a standard TF-IDF-normalised bag-of-words representation. Recommendations for each volume are pre-calculated in advance for performance reasons. Fig. 3 shows an example of a partial list of volume recommendations which were generated for an 1862 volume describing the topography of Northern Europe.

2.2 Advanced Functionality

In addition to using standard information retrieval technologies, the *Curatr* platform also makes use of neural word embeddings to support a range of more advanced forms of textual analysis. *Word embeddings* refer to a set of machine learning techniques, based on neural networks, which map the words in a corpus vocabulary to a numeric representation [6]. In this new representation, words which frequently appear together in the original corpus will be similar to one another, while words which do not frequently appear together will be dissimilar. So, for example, for the input word "influenza", an embedding might automatically recommend similar words such as "pneumonia" and "bronchitis". Word embedding methods have been used in digital humanities research to generate semantic lexicons for a range of purposes, including detecting language change over time [2], extracting social networks from literary texts [11], and semantic annotation of texts [5].

Rank	contagion	influenza	monarch	surgeon	ceylon	finland
1	infection	pneumonia	sovereign	assistant	java	bothnia
2	contagious	bronchitis	usurper	physician	sumatra	baltic
3	disease	pleurisy	throne	veterinary	singapore	riga
4	contamination	neuralgia	prince	chirurgeon	bombay	livonia
5	distemper	typhoid	king	medical	india	esthonia

Table 1: List of the top five words suggested for query expansion by *Curatr*, for a set of six example query words. The suggestions are generated based upon a *word2vec* embedding model.

A variety of different approaches have been proposed in the literature to construct embeddings. The word embedding algorithm used in this research is the popular *word2vec* approach [6]. The specific configuration is a 100-dimensional Continuous Bag-Of-Words (CBOW) *word2vec* model, trained on the full-text of the English language volumes in the British Library corpus, which yields a new representation of 719,735 unique words.

In *Curatr* we use word embeddings to support several layers of functionality. When searching large text corpora, a common task involves creating *word lexicons* [5]. That is, compiling lists of words of interest which all relate to a specific topic. Typically, this is a manual process, where a small initial set of "seed" words is expanded through a process of trial and error. We use word embeddings to reduce the effort required to create word lexicons, while keeping the scholar in the decision-making loop. Once a lexicon has been created with a small number of seed words provided by the user, the list can be augmented based on recommendations surfaced from the word embedding model, which are presented to the user for acceptance or rejection in an iterative manner. These recommendations are generated by identifying a ranked list of words in the embedding which are most similar to all of the existing words in the lexicon. The expanded lexicon can be subsequently used by the researcher to directly search the corpus. The documents identified by this process can then be exported as a sub-corpus for further inspection using other tools.

Word embeddings are also used to support other functionality in the *Curatr* platform. For instance, when searching, users are presented with suggested alternative search queries related to their current query. This process can be viewed as a form of *query expansion* [10], which uses the word embedding model to identify words which are highly similar to the current query word(s). Table 1 shows ranked lists of words that are suggested for a set of six example queries.

To provide an alternative visual search interface, *Curatr* also allows researchers to identify further relevant search terms by constructing a *semantic network*, based on one or more user-specified query words. Semantic networks have previously been used in a variety of disciplines to explore the lexical environment around concepts extracted from unstructured text [8]. While traditionally semantic relations between words have been generated directly from raw word

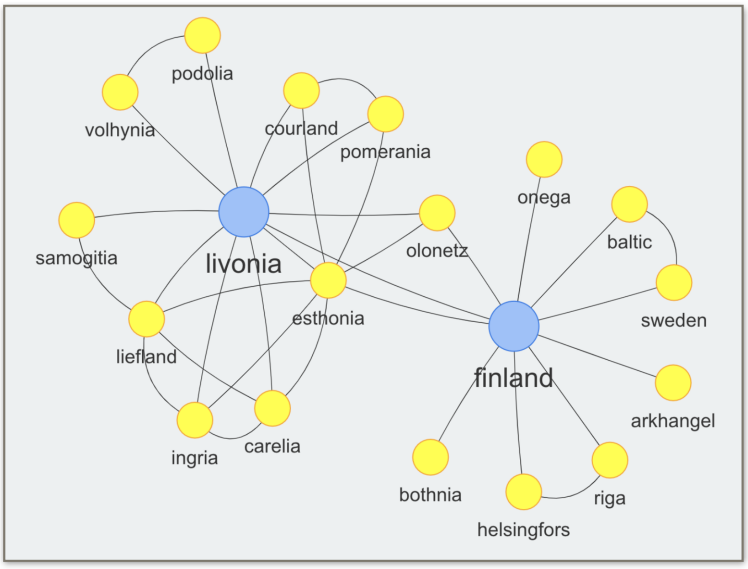

Fig. 4: *Curatr* visualisation of a semantic network created for two sample keywords "livonia" and "finland". Each node in the network represents a word, while an edge indicates a semantic association between two words.

co-occurrence counts, we make use of the *word2vec* embedding model described previously for this purpose. Information from the embedding is depicted visually as an interactive web-based network diagram, where words are represented by nodes, and an edge connecting two nodes indicates that the two corresponding words are highly similar to one another in the embedding. Clicking on a node performs a search for the corresponding word. An example of this type of network is shown in Fig. 4, for the two keywords "livonia" and "finland".

3 Conclusion

This project presents *Curatr*, a platform for exploring large digital corpora, which makes use of techniques from information retrieval and machine learning to support text search, volume recommendation, semantic search, and sub-corpus curation. While the primary focus of our work has been on providing improved access to the British Library Digital Corpus of 18th and 19th century texts, we plan to investigate extending the platform to include other corpora and different types of content, such as texts from historical newspaper archives.

Acknowledgement. This research project was supported by the Irish Research Council (IRC) and Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2.

References

1. Griest, G.L.: A victorian leviathan: Mudie's select library. Nineteenth-Century Fiction 20(2), 103–126 (1965)
2. Hamilton, W.L., Clark, K., Leskovec, J., Jurafsky, D.: Inducing domain-specific sentiment lexicons from unlabeled corpora. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). p. 595 (2016)
3. Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., Coates, C.M.: Trading consequences: A case study of combining text mining and visualization to facilitate document exploration. Digital Scholarship in the Humanities 30 (2015)
4. Jockers, M.: Detecting and characterizing national style in the 19th century novel. In: Digital Humanities 2011, Stanford, CA (2011)
5. Leavy, S., Keane, M.T., Pine, E.: Mining the cultural memory of irish industrial schools using word embedding and text classification. In: Digital Humanities 2017, Montreal, Canada (2017)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
7. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the 5th ACM conference on Digital Libraries. pp. 195–204. ACM (2000)
8. Nulty, P.: Network visualisations for exploring political concepts. In: 12th International Conference on Computational Semantics (IWCS) (2017)
9. Vane, O.: Text visualisation tool for exploring digitised historical documents. In: Proceedings of the 19th International SIGACCESS Conference on Computers and Accessibility. pp. 153–158. ACM (2018)
10. Voorhees, E.M.: Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR'94). pp. 61–69. Springer (1994)
11. Wohlgenannt, G., Chernyak, E., Ilvovsky, D.: Extracting social networks from literary text with word embedding tools. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities. pp. 18–25 (2016)
12. Wolfe, J.: Annotations and the collaborative digital library: Effects of an aligned annotation interface on student argumentation and reading strategies. International Journal of Computer-Supported Collaborative Learning 3(2), 141 (2008)