Intertextual Bridges: Search and Navigation across Heterogeneous Collections

White Paper and Final Report

1. Project Summary

The Intertextual Hub (https://intertextual-hub.org/) is an experimental digital humanities reading environment that aims to situate specific documents in their broader context of intertextual relations, whether in the form of direct or indirect borrowings, shared topics with other texts or parts of texts, or other kinds of lexical similarity. Intuitively, we believe that the conceptual relationships discovered by text mining algorithms among texts in large, heterogeneous collections can fruitfully inform and guide traditional close-reading approaches. More fundamentally, our contention is that scholarly reading in the digital age—and the true usefulness of computational analysis of texts—should be ~~*fore*~~grounded ~~on~~ in the discovery and navigation of intertextual relationships. The model we have developed here allows users to navigate between individual and larger groups of texts that are related through shared vocabulary, themes, ideas, and passages. What the Intertextual Hub offers, then, along with the scalable reading tools, is an approach to federating collections that can bypass the various competing problems of quality (OCR vs. curated) and access (pay vs. public) inherent in digital collections today, and still yield meaningful results.

More concretely, the Intertextual Hub is based on opening the traditional text analysis capabilities of PhiloLogic4 with a broad array of distant reading algorithms by leveraging the structured data generated by PhiloLogic4. Each corpus or collection is loaded into a PhiloLogic4 instance which is configured to suit best the characteristics of that collection. For example, only one of the collections, the debates of the *Archives Parlementaires*, has speakers identified as a feature, which is, of course, made available in the PhiloLogic4 instance. But, when put in the context of the Intertextual Hub, this example of the *Archives Parlementaires* becomes part of a larger set of collections combining a variety of different types of documents, ranging from all of the laws and decrees promulgated during the Revolution, to a large collection of Revolutionary pamphlets, as well as to the great works of major figures of the Enlightenment, including many English authors in translation. Data is extracted from each PhiloLogic4 instance using the ARTFL Preprocessing Library which supports numerous options including lemmatization, modernization, and part of speech identification as well as the structural metadata for each collection. This process works across all of the data of the various PhiloLogic4 instances and can be reiterated across as many collections as required. For this instance of the Intertextual Hub, we used the

Preprocessing Library to extract data for Topic Modeling, sequence alignment, ranked relevancy retrieval, object level similarity and word modeling across all of the collections. This allows the user to have full access to traditional text analysis tools, while making use of distant reading tools to select specific small parts of a text or document to read in this larger context. By opening the full data and services of PhiloLogic4, we have attempted to build a smarter silo, preserving the intellectual coherence of specific collections while merging many collections search, retrieval and examination in that broader context.

2. Project Origins and Goals

The Intertextual Hub seeks to address two long standing issues in digital humanities. The first is the integration of distance and close reading in a functional environment. The second is the architectural issue of how to handle the rapid accumulation of large disparate collections without reducing them all to a lowest common denominator.

The exponential expansion of available textual data across many disciplines, from Google Books/HathiTrust to the ever-growing number of language-,region- and period-specific collections, has sparked a revolution in techniques to examine these vast datasets based on text data mining and machine learning. The accumulation of data in the past decade or two has been nothing less than breathtaking, and it may be argued that in some subspecialities, such as French Revolution studies, the available holdings represent a significant portion of published materials identified in major bibliographies. The various analytical techniques employed to examine the growing mass of materials, grouped together as “distant reading”, have proven effective in opening new and exciting results and interpretations. Lost in this development, however, has been the individual text. This effort then attempts to situate individual texts in a broader textual and cultural environment by reconfiguring distant reading techniques around individual documents. This is accomplished by making explicit the many connections between texts that can be identified using distance reading techniques. These range from the identification of such “objects” as borrowed passages or specific parts of documents that share similar topic distributions,or- similar vocabulary. While reading a document or part of a text in the Intertextual Hub, the user is alerted to passages borrowed from other texts. The Hub displays in real-time the most similar portions of text from across the collections, and the user is thus encouraged to select relevant passages from the document to find other documents with similar passages.

As the number and scale of collections increase, the immediate response is to simply merge them all in a single instance. This is, of course, a perfectly serviceable approach

for some applications, but it comes at a cost: lost are the specific characteristics of different collections or datasets. Working on different kinds of documents –from dictionaries to collections of poems – requires systems that can handle these differences. PhiloLogic4 can be configured to treat a wide variety of types of document collections. Simply adding different collections to a single instance, is impractical since collections are so different. Maintaining and displaying the actionable data across a wide array of instances is a more practical, effective, and powerful way to merge collections. Thus, rather than a single instance of a database, we have adopted a federated search, retrieval and navigation model. This can be extended to any number of collections, while respecting the integrity of specific document types.

Our work with many colleagues on French Enlightenment and Revolutionary collections served as the ideal testbed for this development and deployment. As shown in Appendix One, the ARTFL Project hosts a very comprehensive collection of materials covering this time period and region. Running from a significant collection of pamphlets published in the heat of the moment to the great works of the Enlightenment, we have developed a deep, cross-disciplinary understanding of the different resources and how they might be exploited and have attempted to reflect this model of how to work with these heterogeneous collections in the Intertextual Hub.

3. Project Activities, Team and Participants

Working closely with Robert Morrissey, Clovis Gladstone played the key role of lead developer on the project. Charles Cooney has contributed SQL and other components to the Hub. Mark Olsen has participated in data preparation and top level search and retrieval design. As a working team, we were somewhat impacted by the current pandemic situation, since we all work remotely on servers at the University of Chicago. In particular, this situation made hiring research assistants at the University of Chicago to work on specific coding issues problematic. We also encountered insurmountable difficulties in getting additional data resources from our colleagues in France and from Stanford University. We hope to be able to add these resources when they become available.

As to the data, we have assembled seven different collections that represent a wide array of documentary materials concerning the French Revolution. As shown in Appendix One, these collections include the Newberry FRC, the Archives Parlementaires (AP), the Baudouin Collection of Revolutionary Laws, the Journaux de Marat, as well as 18th-century holdings from the ARTFL Frantext Collection, the Goldsmith-Kress Collection, and French holdings of ECCO. We had planned to add new volumes of the Archives Parlementaires and significant runs of the Moniteur, but we

were unable to proceed on these data capture projects due to the pandemic. The data has been slightly reconfigured better to suit the global search and navigation tasks that are the aim of this project. This work required significant curatorial effort to identify duplicates or near duplicates across the various collections in order to eliminate uninteresting similarities, such as numerous editions or printings of a particular document. This effort evolved over the course of the project. At the outset, we used a near duplicate document detection system built into our sequence aligner, Text-PAIR, to identify similar documents. As the project progressed, we developed a similar document detector based on word vectors using Spotify Annoy (see below). We believe we struck a reasonable balance between removing duplicate documents and keeping documents that contain significant reuses, given the very fluid separation between these two kinds of duplications.

As a critical part of this project, we completed development of several important software components. The first is the ARTFL Text Preprocessing Library which is available open source on GitHub at

https://github.com/ARTFL-Project/text-preprocessing

This code extracts consistent, structured data from PhiloLogic4 database instances that is required for cross-collection processing, search, and navigation. We use this extracted data for global search and retrieval, generating topic models on individual collections, performing sequence alignment across collections, and generating topic based text segments. The library allows us to leverage PhiloLogic4 services to enhance search, reporting, and navigation.

The second element in this development is our open source release of TopoLogic, a Topic Model generator and navigation system which is also available on GitHub:

https://github.com/ARTFL-Project/TopoLogic

TopoLogic builds value-added services on top of the standard PhiloLogic index, leveraging topic-modeling techniques to offer an alternate way of exploring text collections. Topic-modeling, the algorithmic technique which we use for this new navigational tool, is an unsupervised machine learning approach designed to facilitate the exploration of large collections of texts where no topical information is provided. As such, this computational method can be a truly useful way of gaining a sense of the topical structure of a corpus -- i.e. to find out what's in there -- and how words are clustered together to form meaningful discourses. TopoLogic builds upon the topics and semantic fields generated by the algorithm to provide a web-based navigation system which lets users explore topics and discourses across time, as well as word usage within different contexts. The interaction of the three different schemes allows the user to navigate between alternative ways of considering topics across the collection.

We have made slight modifications to our existing TextPair package, which generates sequence alignments to identify and navigate similar passages in a single collection or between collections. TextPair is available on GitHub:

https://github.com/ARTFL-Project/text-pair

The modifications allow tighter integration with the preprocessing library and added services, generating visualizations of all similar passages in a single document, an important component of the Intertextual Bridges Document Reader.

For the federated word and metadata search component of the Hub, we opted to use the open source database engine, SQLite (https://www.sqlite.org/index.html), and its full text search module (FTS5 https://www.sqlite.org/fts5.html). The purpose of this component is to allow users to conduct term queries or simple bibliographic queries across all of the individual collections. Term queries return lists of documents ranked by relevance and can be delimited by bibliographic criteria. After experimenting with other database engines with full text search capacities, such as PostgreSQL, we chose SQLite primarily because it offers built-in BM25 ranking functionality for full text query results (https://en.wikipedia.org/wiki/Okapi_BM25). The SQLite database engine is not without limitations. We wanted users to have the choice of executing ranked relevancy searches with stemmed forms of query terms, to give as full and reasonable a sense of document relevancy as possible. Unfortunately, SQLite does not offer reliable tokenizing/stemming for the French language, only for English (https://www.sqlite.org/fts5.html#porter_tokenizer). We therefore turned to an open source extension to SQLite that allows FTS5 to use a Snowball stemming library (https://github.com/abiliojr/fts5-snowball). This extension proved effective, but in the end, the SQLite full text index we built with the extension did not satisfy all of our requirements for combined metadata and text search.For this reason we decided to usea python stemming package and constructed two separate full text indexes, one with terms pre-stemmed, the other with surface forms. When the option for stemmed word search is checked, the Hub's search code stems terms upon query execution. We feel that this approach offers the best and most flexible means of generating ranked relevancy results that can also be delimited by many different bibliographic criteria: author names; work titles; dates and date ranges; source collection; and time periods -- pre-Revolutionary, post-Revolutionary, or both. Searches can be executed using stemmed words, non-stemmed words, and topic vectors joined by Boolean ANDs or ORs. And finally, users can conduct bibliographic searches just to retrieve documents.

Considerable effort went into the design and implementation of a user interface that would allow users to navigate different types of intertextual relationships. The main objective of this web application was to provide an innovative, seamless integration of both close and distant reading perspectives. In essence, we aimed to offer both a

bottom-up perspective where readers of particular texts contained in our collections would have immediate access to any other related texts, and a top-down lens where intertextual relationships can be queried and explored within the close reading interface. While the actual Intertextual Hub web application provides distant reading points of entry (federated search, text-reuse search, topic modeling browser, word usage and evolution), the end points of all these top-down approaches is a portion of the text itself, which then provides links to other texts within the collection.

Critical to these objectives is the close reading interface, implemented in order to enhance the exploration of texts in the Hub environment. It works to guide the reader to similar texts by displaying if any reuses of earlier texts as well as any later reuses. This guided reading interface highlights within the text all of these text reuses, and additionally allows readers to hone in on any particular passage in order to find similar passages across the collections in the Intertextual Hub.

Finally, we subsequently decided to provide analytics for word usage across the texts included in the Intertextual Hub, showing word associations based on topic modeling, as well as word evolution across time using time slices of word2vec models across the entire 18th century. While the design of the Intertextual Hub app was guided by our focus on 18th century France, we kept the code as generic as possible, with an eye to making a generalizable interface for any collection of texts in any other language. We have released the code as open-source on the GitHub platform: https://github.com/ARTFL-Project/Intertextual-Hub-App

The Intertextual Hub has a significant number of interactional functionalities that are automatically integrated into the build process. The steps of the build process are sequential, building out from the PhiloLogic4 instances as follows.

- PhiloLogic: PhiloLogic is a corpus query engine which provides concordances for all the texts in the different collections which are part of the Hub. Through its indexing system, it provides deep linkages between all texts from document-level metadata to words for all data-mining tools to use. It is deeply integrated into the Intertextual Hub with all links to text views provided by the Intertextual Hub document viewer.
- TextPAIR: TextPAIR is a sequence aligner for humanities text analysis designed to identify text-reuses in large collections of texts. TextPAIR was used to generate all text-reuses across all texts in the Intertextual Hub collections.
- Similar documents: We built on our experience with text similarity detection to build two separate document similarity detection schemes:

    - A pre-calculated measure where all texts from the Hub collections were compared to one another, and the top 20 most similar texts for each document were saved in the Hub database.
    - A live runtime similarity measure which relies on the Annoy package (https://github.com/spotify/annoy), which provides fast and accurate similarity measures. The document viewer feature where users can select a passage and find the most similar document in the Hub collections relies on this feature.
- Topic modeling: We built a standalone tool, TopoLogic, during the course of the grant in order to explore ways to leverage results from topic models for corpus analysis and exploration. We subsequently integrated many of the components of TopoLogic within the Intertextual Hub app. The model we built relies on all of the texts within the Hub collections, and was refined over several weeks of testing text preprocessing and model hyperparameters.
- Word embeddings: We explored a number of different word embedding models (word2vec, Glove, Single Value Decomposition, Fastext) to provide a view of word evolution across the whole 18th century. After careful analysis, we decided to use word2vec as it provided the most consistent results across all the different algorithms. We split all of the texts in the Hub collections into four periods (1700-1725, 1725-1750, 1775-1800), and ran the word2vec algorithm on each group of texts. We then generated lists of the most similar words for each word in each period, and subsequently integrated these results in a word cloud view displayed in the Intertextual Hub application. We also built a global word2vec model based on all the text in the Hub collections for the most associated words feature in the Intertextual Hub application, where users have the ability to automatically append the most similar words to a term from the search form.

4. Project Outcomes

There are several significant project outcomes. The first is the public release of the Intertextual Hub Web site and supporting resources:

https://intertextual-hub.org/

As part of the supporting resources, we have provided a variety of use cases and given an overview of the kinds of results that the Intertextual Hub may support in a series of blog posts and slide presentations listed on

https://intertextual-hub.org/presentations-2/

The use cases stem from real work using the Hub based on previous research. We have also developed a screencast presentation outlining the overall functionality of the Hub:

https://intertextual-hub.uchicago.edu/intertextual_hub_screencast.mp4

The prototype Hub itself is the most important of the project outcomes:
https://intertextual-hub.uchicago.edu/
This instance of the Intertextual Hub is open to the public. By agreements with Gale/Cengage, we can only permit access to the full text of their collections to authorized users. The Hub otherwise provides full functionality in terms of topics, similar passages, merged searching and so on. The is running under a virtual machine on one of ARTFL's servers at the University of Chicago.

As noted above, we have provided open source releases of all of the primary components of the Intertextual Hub on GitHub. These are:

- PhiloLogic4: https://github.com/ARTFL-Project/PhiloLogic4
- Text-PAIR: https://github.com/ARTFL-Project/text-pair
- TopoLogic: https://github.com/ARTFL-Project/TopoLogic
- Text Preprocessing Library: https://github.com/ARTFL-Project/text-preprocessing
- Hub App: https://github.com/ARTFL-Project/Intertextual-Hub-App

As we pursued our work, we determined that it would be best to defer to a future development round one element of the proposal.. We had anticipated a user-submit mode, where users could upload TEI texts, which would be added to a user-contributed collection, and subsequently integrated into the Intertextual Hub We ended up dropping the idea of having a user-submit mode for the Intertextual Hub for a number of reasons. First, as we learned throughout the process of coordinating a number of different collections with distinct TEI encoding schemes and internal structure, the inclusion of additional texts requires both quality control and an analysis of the nature of the document (literary work, newspaper, dictionary...etc). This makes any automatic processing of new texts unfeasible. Additionally, the process of building and connecting the different analytics and data-mining results included in the Hub, while partially automated, is not conducive to the easy integration of new texts on the fly. As a result, we decided to defer the implementation of such a feature, and reconsider its inclusion once the build system and the structure analysis of texts is more automated.

We have presented preliminary results from this project to a couple of venues. Clovis Gladstone presented the ARTFL Project's last work around intertextuality during the Colloquium on Digital Humanities and Computer Science on November 9th 2019. In particular, he discussed the development of TopoloLogic as a first step to a generalized web application for intertextual exploration. He also discussed the progress being made on the Intertextual Bridges Web application front at a talk on March 2nd 2020 at the University of Chicago. In particular, he presented the first elements of the Hub

document viewer, and the possibilities offered by latent guided reading features displayed in the close reading environment of the Intertextual Hub.

5. Project Evaluation and Impact

The scope of the current project was the development and deployment of a working prototype over a 12 month period. Given the complications of the pandemic, we requested an additional 4 months to complete the development and deployment. We released the prototype to our Advisory Board in the Fall of 2020 and to the wider scholarly community in early 2021 via various lists aimed at French historians and digital humanists more generally.

An important element of the next phase of the project, which will continue without direct NEH support, is encouraging user feedback and evaluation. We have created a feedback reporting system and are monitoring use and feedback. We have received expected bug reports of various kinds, such as access control questions, as well as more general suggestions about future development and directions. We anticipate additional feedback from users in the next several months.

6. Project Continuation and Long-Term Impact

The Intertextual Hub is installed on the ARTFL Project's supported production servers. We plan to keep this instance running for several years so as to provide a coherent environment for future testing and evaluation. We are planning a second upgrade instance which will feature some additional material, Moniteur and 18 volumes of AP as they become available. We would also like to open negotiations with Gale to acquire page images and new data from ECCO and Goldsmith.

The prototype has already revealed a set of future development goals, some of which would correct limitations in the prototype and others to open the uses of the Hub to other languages and time periods.

We believe that the Hub architecture has proven itself to be robust and flexible. We are using PhiloLogic4 services from word indexing to document access and browsing as the basis for a variety of functions that merge the collections in various ways. In this first instance, we did all of the processing locally, on a single machine. For a relatively small set of collections, such as the materials surrounding the Revolution, this is a reasonable approach. Moving to sets of much larger collections, it will become necessary to adopt a more widely distributed architecture with instances of PhiloLogic4 running on many hosts. The preprocessing library can be readily fitted with a function to remotely gather

salient data from PhiloLogic instances, and perform higher level functions on a different machine. Federated functions, such as search, topic models, similar documents and so on can be performed on one machine (or set of machines) while document services and more traditional text analysis applications will be handled by separate PhiloLogic instances.

Appendix One: French Eighteenth Century Digital Resources

The ARTFL Project has assembled an unparalleled collection of digital resources related to the French Eighteenth Century and Revolution upon which to base this pilot project. In the course of this development work, which has spanned a number of years, we have developed a deep understanding of the different resources and how they might be exploited.

ARTFL's French Revolutionary holdings are anchored by four open access collections developed in collaboration with institutions in the United States and France. The largest of these is the French Revolution Collection of the Newberry Library in Chicago, two versions of which we are hosting at ARTFL.[1] We have done considerable preliminary work on this extraordinary collection, most notably to create a deduplicated and machine corrected subset of 26,445 volumes (121M words) for the period of the Revolution. In collaboration with Stanford University, we have released 83 volumes of the *Archive Parlementaires* (AP), which is the daily legislative record from the beginning of the Revolution to the fall of Robespierre, containing all of the sessions, speeches, debates (in which speakers have been identified), documents, and appendices.[2] The companion dataset to the AP is our complete collection Revolutionary Laws. Developed in collaboration with researchers at the Sorbonne, it contains 56,000 laws, forming the complete record of legislation during the Revolutionary period.[3] ARTFL and Stanford have also released *Les Journaux de Marat*, which contains 932 numbers of Marat's Revolutionary journals.[4] We have also accumulated smaller collections containing works by Robespierre, Condorcet, Danton, and other central Revolutionary figures.

ARTFL's holdings of 18th-century collections are equally as broad and varied. The ARTFL Frantext database contains over 700 texts dating from 1700 to 1789, which offer significant representation of Enlightenment and counter-Enlightenment traditions. This will be supplemented by texts drawn from our collection of some 250 plays from the 18th century; the complete works of Rousseau, Voltaire, and Diderot; the *Encyclopédie* of Diderot and d'Alembert; and significant holdings of the Encyclopédie Méthodique.[5] Through our collaboration with OBVIL at Sorbonne University, we have recently acquired Tres Grande Bibliotheque (TGB) collection, which contains well over 6,000 documents from the 18th century.[6] The ARTFL 18th century holdings also contain

---

[1] https://artfl-project.uchicago.edu/content/french_revolutionary_collection
[2] https://artfl-project.uchicago.edu/node/148
[3] https://artflsrv03.uchicago.edu/philologic4/revlawsall/
[4] https://artfl-project.uchicago.edu/content/lami-du-peuple
[5] See http://artfl-project.uchicago.edu/content/public-databases for links to many of these collections.
[6] http://api.bnf.fr/documents-de-gallica-produits-au-format-tei-par-obvil

over 2,700 texts related to economics found in the GoldSmith Kress collection[7] and more than 4,300 French imprints, primarily from London, found in the ECCO-French collection.[8]

All of these databases have been loaded into PhiloLogic4. Because of some previous work, the Revolutionary collections are nearly ready to be integrated. We expect that we will devote some effort as part of this project to deduplicating some of the holdings from the broader 18th century and possibly running automated correction algorithms on some of these datasets.

[7] https://www.gale.com/c/making-of-the-modern-world-part-i
[8] https://www.gale.com/intl/c/eighteenth-century-collections-online-part-i