White Paper

Report ID: 111509

Application Number: HJ-50185-14

Project Director: Robert Morrissey

Institution: University of Chicago

Reporting Period: 1/1/2014-12/31/2015

Report Due: 3/31/2016

Date Submitted: 3/31/2016

White Paper:
Commonplace Cultures: Mining Shared Passages in the 18th Century using Sequence Alignment and Visual Analytics

The ARTFL Project, University of Chicago

Introduction

In many ways, the 18th century can be seen as one of the last in a long line of "commonplace cultures" extending from Antiquity through the Renaissance and Early Modern periods. Recent scholarship has demonstrated that the various rhetorical, mnemonic, and authorial practices associated with commonplacing—the thematic organization of quotations and other passages for later recall and reuse—were highly effective strategies for dealing with the perceived "information overload" of the period, as well as for functioning successfully in polite society. But, the 18th century was also a crucial moment in the modern construction of a new sense of self-identity, defined through the dialectic of memory (tradition) and autonomy (originality), the resonances of which persist long into the 19th century and beyond. Our goal in this project was to explore this paradigm shift in 18th-century print culture from the perspective of commonplaces and through their textual and historical deployment in the various contexts of collecting, reading, writing, classifying, and learning. These practices allowed individuals to master a collective literary culture through the art of commonplacing, a nexus of intertextual activities that we uncover through the use of sequence alignment algorithms to compile a database of potential commonplaces drawn from the massive ECCO (Eighteenth Century Collections Online) collection provided to us by Gale Cengage. The ECCO corpus, which comprises some 200,000 volumes of texts published in England from 1700 to 1800, represents the most complete and comprehensive archive of 18th-century print culture available.

Our main objective was thus to move us beyond simply describing the underlying theories and methodologies of commonplacing, and allow us, for the first time, to consider, contextualize, and visualize the full spectrum of commonplace practices on a scale previously unimaginable. From this macroscopic level—viewed from the perspective of a collective cultural system—we can pose a broad range of new research questions: how and what commonplaces circulated, through which networks; their national and linguistic distribution; and their workings in collective cultural and discursive entities (e.g., Republic of Letters, Enlightenment vs. Counter-Enlightenment, Ancients vs. Moderns, etc.), as authors and individuals moved progressively from a culture of commonplaces to one of originality and autonomy.

During the initial stages of the project we worked closely with our partners at the University of Oxford to develop a visual analytics system that leverages image processing techniques for identifying and refining text alignment parameters. This system, called ViTA: Visualisation for Text Alignment, is now available as part of the Oxford Visual Infomatics Lab suite of tools (see http://ovii.oerc.ox.ac.uk/vita/) and is the subject of a recent joint article in *Computer Graphics Forum*, "Constructive Visual Analytics for Text Similarity Detection" (doi:10.1111/cgf.12798). Building upon this collaborative work and the various lessons learned from the ViTA development process, we at ARTFL developed several new approaches to handling the problem of text reuse on a large scale in order to identify reuses of the same passage over time. Since aligned passages that are shared by two texts can vary in size, this requires a high performance algorithm running on multiple cores based on frequencies of n-grams. This allows for the

identification of a single passage with many variants over time. ARTFL has extended the scale of the project beyond the ECCO corpus to include a large collection of English texts predating the 18th century and a complete collection of classical Latin texts. In this way, ARTFL has been able to track individual text borrowings from the source author chronologically through all subsequent reuses.

We have also worked to resolve two unanticipated phenomena arising from processing very large selections of documents—hundreds of thousands—rather than relatively small samples: passage reverberation and the fact that a small number of documents comprise the significant majority of text reuses. In the work done so far, citations from the Bible account for more than 70% of all of the similar passages in 16th and 17th century English literature and well over half for the 18th century. ARTFL has developed passage identification algorithms that filter out reuses from very high frequency texts or authors thus facilitating the identification of less "visible" texts that are more illustrative of significant patterns of reuse. The results of this work have been deployed in an easy to use web-based interface to a database of some 60 million aligned passages which is now open to the public.

The following will first outline the processing steps that we used to identify similar passages as well as those required to be able to track the "same" passage over time. The will be followed by a discussion of the resulting database and interface that was developed in order to most effectively exploit this data.

Processing Steps for ECCO

With more than 200,000 volumes, the ECCO database is a prime example of a "big data" resource for humanities research. Given the scale of its contents, it is fortunate that Gale Digital Collections has put together a classification scheme by topic in order to give an overview and facilitate navigation within the dataset. There are seven main rubrics: General Reference, History and Geography, Law, Literature and Language, Medicine and Sciences, Religion and Philosophy, and Social Sciences and Fine Arts. Building on our extensive experience in the development of digital tools for detecting similar passages between two different texts—or sequence alignment—at scale, we took a three step approach in our attempt to gather and organize text reuses in the ECCO collection: focus on the earliest editions of any given text by eliminating subsequent editions, generate a list of all shared passages in the dataset, group together similar shared passages in order to track reuses across time.

As we examined the content of the ECCO collection, we discovered a great number of re-editions of texts. Since the goal of our project was to uncover text reuses across the whole 18th century, and not to analyze reuses within a single author, we decided early on only to keep from our text corpus all the earliest editions of any given text. In order to do this, we had to devise a method to detect and remove any later editions of any single work, a far from trivial task given the large variations in titles in 18th century works. Though the most obvious method to flag duplicates may seem to be to compare the contents of each document with all others—and define a similarity threshold beyond which we consider two works to be the same—we dismissed this strategy because such a comparison is computationally expensive, and in our case mostly unreliable given the quality of the OCR in the ECCO dataset. As a result, we decided to focus our effort on comparing document metadata—which is of excellent quality in the ECCO collection—as a means to detect reeditions.

Our methodology consisted in computing the cosine distance of vector space representations of the titles of each work. By using this technique, one obtains a similarity score between 0 and 1 based on the lexical similarity between two titles. For our purposes, we determined a minimal similarity score to automatically determine whether two texts were similar enough for us to flag the later work as a reedition of the earlier one. We then built a basic decision tree to evaluate the probability that any given text was a duplicate, and were able to reduce the size of the collection by 43%, that is by 88,850 documents, to retain finally 116,700 documents.

The effect of this deduplication step was to reduce drastically our computational problem since detecting identical or similar passage requires a one to one document comparison of every text in the dataset. But rather than compare the 116,700 texts of our corpus, a very lengthy process, we decided to leverage the division of the ECCO collection in modules, and to limit the comparison inside each individual module. This choice was also motivated by our analysis of sample inter-module comparisons, which showed that text reuse across different rubrics tended to be less frequent: in other words, a historian is less likely to borrow from a literary source than he is from another fellow historian.

For the purpose of text reuse detection, we used PhiloLine, our in-house sequence alignment tool, as it can scale up quite well. Without going into details, the basic functioning principle of PhiloLine is to compare sequences of words, or n-grams, and determine the presence of a shared passage according to the number of common contiguous n-grams between two textual sequences. Below is an example of two passages that reached the necessary threshold of shared n-grams, and were therefore identified as a shared passage:

an Hour of virtuous Liberty, Is worth a whole Eternity in Bondage
hour_virtuous_liberty, virtuous_liberty_eternity, liberty_eternity_bondage
- Joseph Addison

an hour, of virtuous liberty Is worth a whole eternity in bondage
hour_virtuous_liberty, virtuous_liberty_eternity, liberty_eternity_bondage
- James Thomson

PhiloLine ended up finding more than 43 million passages across the entire ECCO dataset, a number which certainly seemed intimidating at first. Identifying these common passages was only a first step of our project as our goal was to extract commonplaces out of these millions of passage alignments.

The main difficulty in uncovering these commonplaces is in defining the commonplace computationally, that is to come up with a set of rules that allow the machine to identify commonplaces across many millions of shared passages. The first clue to finding a commonplace is the repeated use of the same passage - more or less similar - in a minimum number of different authors. A phenomena we discovered was that authors tend to borrow different parts of the same source passage, so the real challenge for us was to merge together all variants of the same passage. If we look at the following passage from the Scottish poet James Thomson:

Then infant reason grows apace, and calls For the kind hand of an assiduous care. Delightful talk! to rear the tender thought, To teach the young idea how to shoot, To pour the freft infiruAion o'er the mind, 1150 To breathe enlivening spirit, and to fix The generous purpose in the glowing breast.

We noticed that the reuse of this passage in other authors could vary significantly. For instance in the Gentleman of the Middle Temple (1775) :

> *How glorious would her matron employments be, to hear the tender thought, to teach the young idea how to Jhoot; to be at once the precept and example to her family of every thing that was good, every thing that was virtuous.*

Or in Mrs Lovechild (1790):

> *Happy the Mother "Distilling knowledge through the lips of " love !"- ' Delightful talk! to rear the tender thought, " To teach the young idea how to shoot, " To pour the fresh inltrution o'er the mind !'Lines which will never cease to be quoted...*

The variability in the reuse of any given passage as seen above, as well as the very uneven quality of the OCR, led us to develop a new algorithm that could match similar passages in a way that was precise, yet more flexible than PhiloLine. We chose to chunk our passages in overlapping two-token skip-grams as a way to account for OCR errors and modifications by later authors: a two-token skip-gram is basically a tri-gram with the middle word ignored. Below is an example of how of these skip-grams look:

> *Then infant reason grows apace, and calls For the kind hand of an assiduous care.*
>
> *(infant, grows), (reason, apace), (grows, calls), (apace, kind), (calls, hand), (kind, assiduous), (hand, care)*

For two passages to be considered as similar enough to be grouped together, we devised a set of rules for the match:

- Each skip-gram is counted once per passage
- Bag of skip-grams model: order of skip-grams doesn't matter
- Define a minimum of N shared skip-grams (determined by the length of each passage) between two passages in order to consider them variants of one another.

The flexibility we introduced in our algorithm allowed us to merge the different uses of a single source passage, and determine a fixed number of unique authors which had reused any given passage. In the case of the passage by James Thomson above, we found that 57 different authors had reused one part or another of the source passage, a number that from our perspective indicates that this source passage is a commonplace.

Database Extensions

Early use of the database of aligned passages derived from the Gale ECCO database led to two unanticipated yet significant developments. The first was the addition of two collections of texts which predate the 18th century:

1) EBBO-TCP covering the early modern period from the end of the 15th to the end of the 17th centuries. This collection contains 25,368 works and, unlike the data drawn from the ECCO database, this collection was keyboarded and well encoded, and;
2) Packard Humanities Institute of Classical Latin corpus of some 827 texts.

We added these two collections to the alignment database because we found that we could not provide proper source identification for a significant number of aligned passages that predated

the holdings of the ECCO database. Thus, passages from Shakespeare, Milton, Locke, and many others could not be properly identified and were often displayed as the earliest use which could be from an early 18th century edition, a commentary, or from a commonplace book. Similarly, given the extensive use of Latin passages and reprinting of Latin texts in various ways meant that limiting the database to the original 18th century-materials would make it almost impossible to reasonably identify reuses of, for example, Lucretius or Cicero.

One of the advantages of both databases is that they are well curated and encoded. We followed the same procedures for these two corpora as we had for ECCO. We aligned the contents of the ECCO modules, generated the output format and merged all of the databases together in a default chronological order. Special care was taken to make sure that we could reliably identify passages from significant authors in English and Latin back to their sources. While this is a significant modification to the original project specifications, we believe it greatly enhances the utility of the final product.

The second major addition to the entire database consisting of alignments of Latin works, works drawn from EEBO-TCP and the ECCO corpus is the identification of probable Bible passages. This is due to the fact that a staggering number of aligned passages are probably Bible passages: more than 70% of all of the similar passages in 16th and 17th century texts and well over half for the 18th century. We determined that users would most likely want a mechanism to filter other types of searches—by author, title or time period—with the likelihood that passages were either probably derived from the Bible or not. Thus, we built a variant of the PhiloLine sequence aligner which used measure of n-grams in the aligned passages compared to a fairly generic but clean version of the King James Bible. We sampled a number of passages and determined that it worked reasonably well. In many cases, particularly in sermons, prayer books, and other religious writing, the difference between Biblical reference and writing about the Bible in a religious vein is sometimes difficult to distinguish. Thus, Biblical paraphrases and allusions may be identified as Bible passages.

Final Result Database Overview

With the addition of new sources (EEBO-TCP and Classical Latin corpus), we ended storing 60,442,024 text reuses in our database:

- 168,603 alignments from Classical Latin sources among which 54,874 are the probable sources of the reuses
- 16,306,951 alignments from EEBO-TCP among which 1,187,408 are the probable sources of the reuses
- 6,713,392 alignments from the ECCO Literature and Language module among which 219,706 are the probable sources of the reuses
- 24,266,904 alignments from the ECCO Religion and Philosophy module among which 420,273 are the probable sources of the reuses
- 1,475,636 alignments from the ECCO General Reference module among which 42,634 are the probable sources of the reuses
- 3,473,358 alignments from the ECCO History and Geography module among which 117,532 are the probable sources of the reuses
- 3,743,431 alignments from the ECCO Social Science and Fine Arts module among which 115,911 are the probable sources of the reuses

- 1,310,301 alignments from the ECCO Medicine, Science and Technology module among which 51,175 are the probable sources of the reuses
- 2,897,966 alignments from the ECCO Law module among which 80,060 are the probable sources of the reuses

If we consider the output of our Bible passage identifier, 35,261,455 reuses were flagged as originating from the Bible, which comes down to 58.5% of all reuses. When considering numbers coming out of our similar passage merger, of the 2,289,551 distinct passages we identified, 670,360 (29%) most likely come from the Bible. The discrepancy between the proportion of total number of reuses and the proportion of actual distinct passages seems to indicate that bible passages were far more likely to be reused many times as opposed to other types of reuses. Indeed, 6,202,602 passages out of the 11,853,736 (52%) were actual reuses from the Bible. In other words, a single Bible passage was reused on average 9 times (6,202,602 reuses out of 670,360 source passages), whereas any non-bible passage was reused about 3.5 times (5,651,134 reuses out of 1,619,191 source passages).

Very early on, we realized that our end product would need to scale at an unprecedented level for typical digital humanities projects. This is why we built a very powerful dedicated server, and wrote the entire server side query engine in a compiled language (Go). The other component of the end product that needed a lot of work was the user interface. The choice of the various components of our product was carefully evaluated after testing various options:

- Database engine: we tried MongoDB and PostgreSQL. We finally settled on a fork of MySQL called MariaDB. The decision was based on our familiarity with MySQL, and the improved performance MariaDB provided over our previous options.
- Web server: Since we knew we wanted to get optimal performance on the server side because of the scale of our dataset, we quickly dismissed Python as on option, and instead tried running a Node.js server/application (Express.js). Performance was better than Python, but we felt we could get better results by switching to a compiled language. This is why we finally settled with a Golang solution (the Gin web framework), which offered very high performance and also better stability. This allowed us to offload a lot of the work we were delegating to the client (web browser), and as a result speed up the retrieval and display of query results.
- Client side application: We had previous experience with AngularJS and were quite happy with its ease of use and performance in the browser. We also used the Bootstrap front-end framework for the UI for its ease of use and flexibility.

A crucial aspect of the application development phase was that we needed to make sure we were presenting the results of our data processing in a way that was both user friendly and conducive to discovery. Our decision to rely on a search form as the main tool for navigation was based on the scale of our data. We found that combining traditional search and faceted browsing offered the best solution for navigating inside the millions of text reuses we had stored. In order to provide users with as many filters as possible for their queries, we stored important metadata (author, title, publication date) for each passage, as well as information on the passage (origin of digital text, length of passage…) or passage context (text before and after each reuse).

Web Interface overview

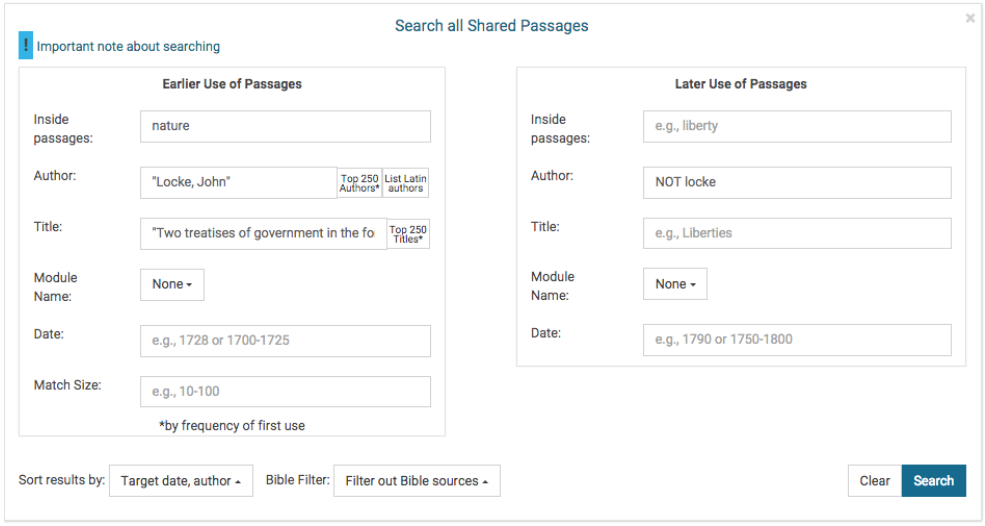
The primarily deliverable and means of dissemination of the results of this project is a web-based interface to the alignment database. This is currently hosted on:

http://rousseau.ci.uchicago.edu/

Documentation, project descriptions and other materials are available on the project web site:

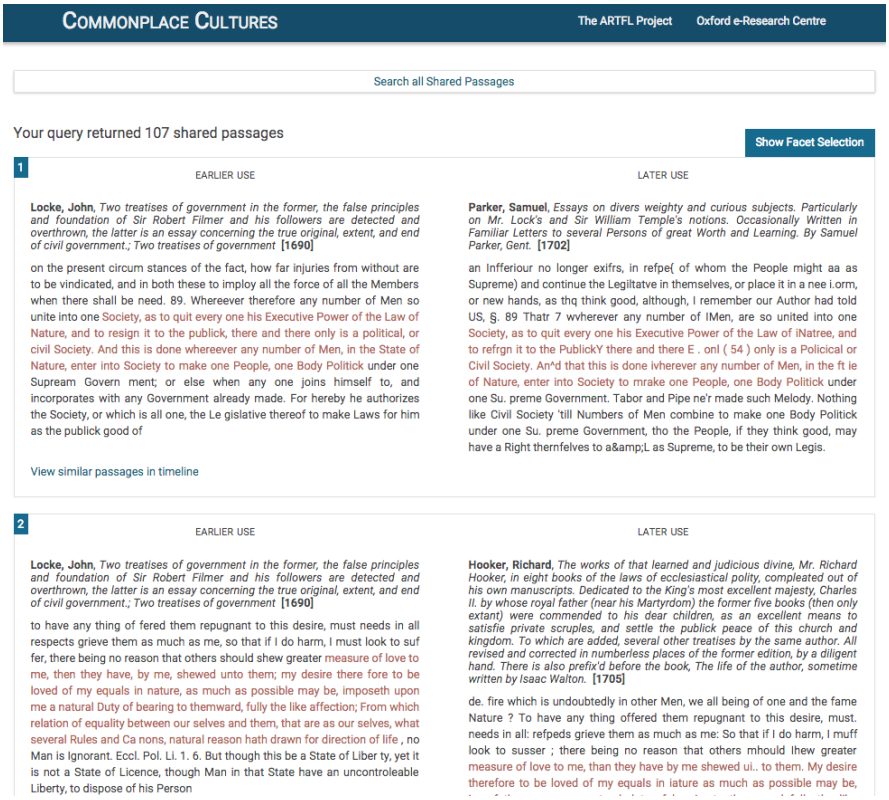
http://commonplacecultures.org/

This section will outline the main functions of the search engine by describing the query form and displaying sample search results. The main search form provides a set of criteria which allows the user to query the dataset.



Screen One

Screen One shows the options, some of which have been filled out. In this case we are searching for aligned passages from John Locke's *Two Treatises of government* which contain the word "nature" found in later passages which do not have Locke as an author. In this case, we want to have the results sorted by the date and author of the later use (also known as target passage) and removed probably Bible passages. The use of NOT in Late Use author and title fields is recommended because there are many subsequent editions of Locke's works in various collections which may be counted as aligned passages. For convenience, we have included pull down lists of highly frequent authors and titles as well as module names (enumerated above) to allow restrictions to specific subsets of the database. The result of this query is shown in Screen Two:



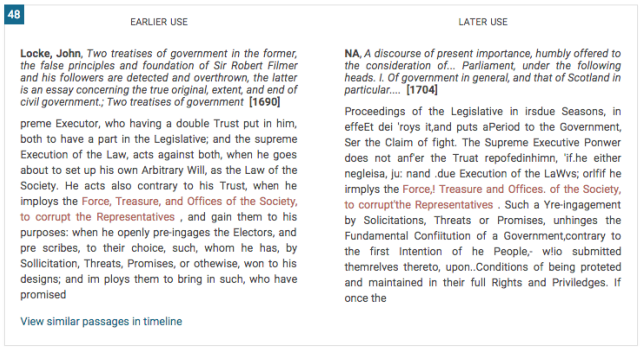
Screen Two

Screen Two shows the aligned passage pairs with some context before and after. Due to limitations by agreement with our data provider Gale Cengage, we are unable to provide links to additional contextualization at this time. The list scrolls dynamically and the user has options to follow a particular passage to see its uses displayed in a time line or to open a facet to get counts on particular metadata fields.

Screen Three

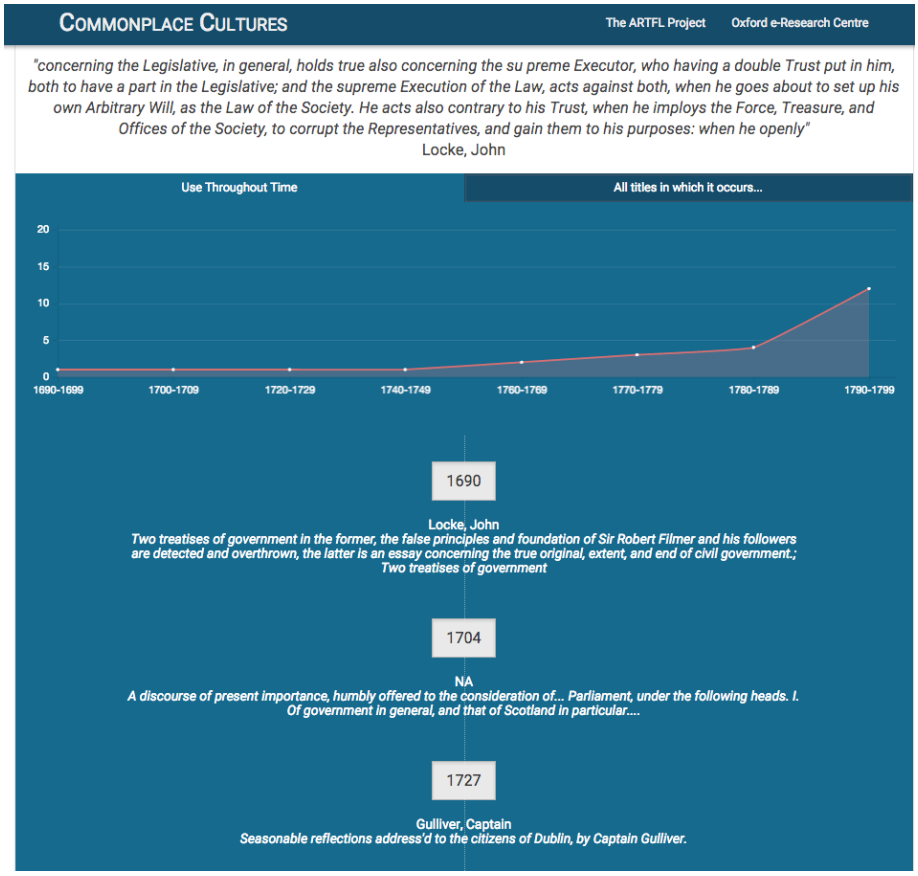
Screen Three shows an opened facet with the available frequencies and the first 5 most frequent authors of the later use aligned passages. Clicking on the author name (or other faceted items), such as Towers, Joseph, allows the user to further restrict the query in order to drill down on a particular item or set of items of interest.

The other option from the default result representation is to drill down to the use of a particular passage. Scrolling down the results displayed on Screen Two, the user may find that result number 48 looks interesting.



Screen Four

Clicking on “View similar passages in time line” switches the view from a pairwise display to a timeline display with longest reference displayed at the top associated with the earliest author and text. As shown in Screen Five:



Screen Five

the longest variant of the passage in question is displayed with the earliest author associated with that passage, in this case John Locke. A simple graph shows the use of this passage over time. Clicking on the titles of any of the references will be the passage found in the (first) aligned pair. In this case, the passage is found in 25 titles with unique authors in the period following its publication. Overall, this passage is found in 36 different titles including editions of Locke where it may be used more than once by an author. In this way we attempt to reduce the impact of the reverberation of multiple reuses of the same passage in various reprints and other less interesting repetitions.

While this overview touches on the various functions built in to the search engine and reporting system, please consult Charles Cooney’s “Text reuse and the reception of Lucretius in 18th-century England” (unpublished working paper) included as Appendix One for a more detailed examination of the new kinds of research that the database facilitates as well as a case study of the kinds of results we don’t think can be addressed using previous digital humanities tools.

Non-WWW dissemination

The project team has been very active in giving talks and writing papers on the various components of this work over the past 2 years. Members of the group have have presented to a number of conferences, including Digital Humanities, DHCS, and an invited colloquium in Paris. We already have a group paper published in *Computer Graphics Forum* based on our collaboration with Oxford and we have have submitted a general paper to a major French literary history journal that outlines the intellectual framework of the project. As noted below, this work has already spun-off one additional grant funded project. We are working with scholars at the University of Chicago to develop similar efforts text reuse projects in 19th and 20th century American literature and in a significant collection of English language works in economics.

Long Term Impact

We are optimistic that the database will serve the global scholarly community in a variety of ways. This is a novel approach to the examination of intertextual relations that has not been attempted before at this scale. There are three complementary approaches offered by the database that will allow scholars to examine questions at an unprecedented scale and level of detail at the same time:

- identification and examination of the reuses of a single text or author over time, as is exemplified in Cooney's case study on reuse of Lucretius in the early modern period;
- identification and examination of the texts reused by an author or collection of texts, such as the sources of Paine's *Common Sense* (http://goo.gl/LHo8Bx) and;
- The ability to track and graph the instances of a particular passage, with a significant number of variations, over time to an original source if it is available in the current dataset, such as Paine's use of a passage by Milton also found in a number of other texts (http://goo.gl/2gprRX).

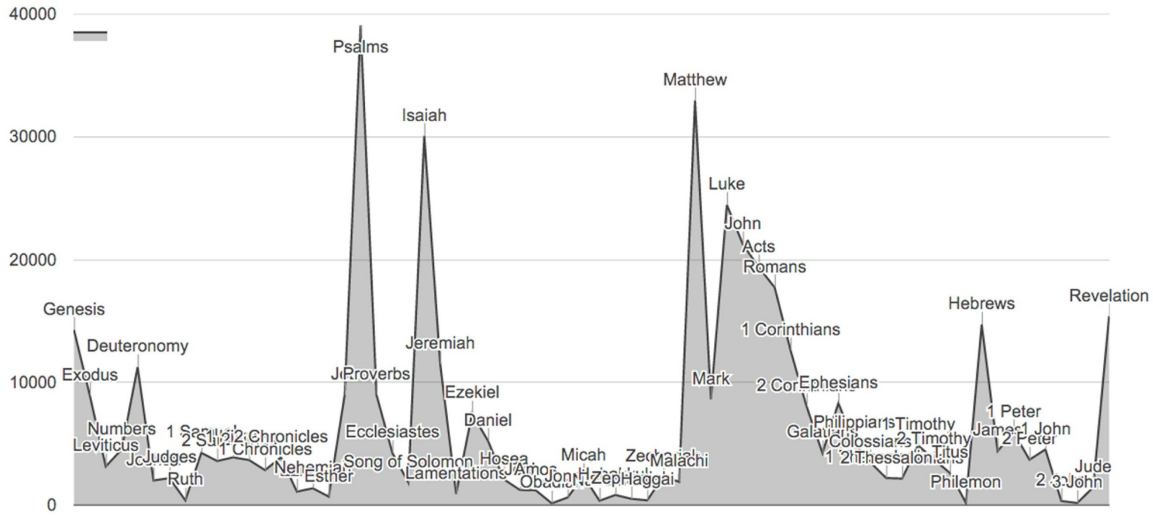
This is an unprecedented and powerful approach that we believe will be widely used as deployed in this first version. We are looking forward to getting feedback from the global community of scholars to whom this resource will be open. Our optimism on the utility of this approach and the value of user feedback comes from our experience in deploying smaller scale alignment database systems, most notably the identification of borrowed passages in the *Encyclopédie* of Diderot and d'Alembert as well as cross alignment of our own French language databases.

Future work

As indicated in the proposal, the ARTFL Project will maintain the alignment database as one our open access products for at least 5 years. The ARTFL Project has significant expertise in maintaining production databases going back to the early 1980s and can commit to ongoing support of this effort. The database is the primary mode of dissemination of this work and will, in many ways, help set the direction for future work.

We anticipate collecting comments from the scholarly user community. Already, in pre-release testing and demonstrations, we have observed a set of additions and modifications that we feel will greatly enhance the analytic utility of the alignment database. As noted above, reuses of the Bible in the early modern period form a substantial portion of all text borrowing. Since our original goal was the creation of a single large database that would would allow global examination of text reuse based on the sources as they are encoded, we did not address the

question of more refined implementation of specialized subsets of important texts aligned against the entire collection. In experimental work, we aligned a version of the King James Bible broken down by book. The graph below,



shows the frequencies of passages drawn from the various books of the Bible. We have done further preliminary work which suggests that a specific alignment of a selection of Bibles, broken down by chapter and verse, as the "source document" would allow systematic examination of the changing uses of Bible over time and comparatively between authors and in specific genres.

This approach could be expanded by developing a set of specific databases that would include curated editions of the works of specific authors as source documents, allowing for a more focussed and controlled examination of the reuse of important authors such as Shakespeare and Milton and specific corpora, such as works of classical Latin in the original and in contemporary translations or inclusion of Latin texts drawn from the Patristic Fathers (PLD). This size and coverage of the text collections we are working with is both a strength of the project, but also introduces problems of source control and closer grained analysis that we think can be address in more specific builds.

This work has already facilitated one major new research initiative: *Use and Reuse: Exploring the Practices and Legacy of Eighteenth Century Culture*. The *Use and Reuse* project is a collaboration, funded by the La Fondation Maison des Sciences de l'Homme, between organizations at the Sorbonne Universities (OBVIL) and the University of Chicago (ARTFL) to identify Enlightenment patterns of textual reuse during the long 18th century and following that tradition through the 19th and, perhaps, into the early 20th century. This project will be primarily in French language materials, in conjunction with the BNF, and has as a primary goal to develop systems that will allow identification of less literal borrowings, such as allusions, and mechanisms to show differences in large similar texts, which is currently not part of the PhiloLine system.

In a related application, the ARTFL Project is a committed participant in a proposed project titled *Lights and Bytes: Using digital tools to foster scholarly collaboration on the French*

Enlightenment. The effort is a collaboration among a wide-ranging team of experienced DH practitioners with specific focus on the French Enlightenment to integrate a wide variety of applications and databases which range from geo-spatial, temporal and content analysis to digitally captured, searchable data sets. One of the key elements of this effort will be the identification of text reuse between different kinds of corpora, such as the correspondence archived in the Electronic Enlightenment at the University of Oxford's Bodleian Libraries and the ARTFL Project holdings of a wide array of 18th century collections such as the *Encyclopédie* of Diderot and d'Alembert.