DHQ: Digital Humanities Quarterly
2017
Volume 11 Number 4

Introducing DREaM (Distant Reading Early Modernity)

Matthew Milner
Stephen Wittek <stephen_dot_wittek_at_mcgill_dot_ca>, Carnegie Mellon
Stéfan Sinclair <stefan_dot_sinclair_at_mcgill_dot_ca>, McGill University

Abstract

We provide a comprehensive introduction to DREaM (Distant Reading Early Modernity), a hybrid text analysis and text archive project that opens up new possibilities for working with the collection of early modern texts in the EEBO-TCP collection (Phases I & II). Key functionalities of DREaM include i) management of orthographic variance; ii) the ability to create specially-tailored subsets of the EEBO-TCP corpus based on criteria such as date, title keyword, or author; and iii) direct export of subsets to Voyant Tools, a multi-purpose environment for textual visualization and analysis.

DREaM[1] (Distant Reading Early Modernity) is a corpus-building interface that opens up new possibilities for working with the collection of early modern English texts transcribed thus far by the *Text Creation Partnership* (TCP), an ongoing initiative to create searchable, full-text versions of all materials available from *Early English Books Online* (EEBO).[2] To offer maximum access within the bounds of restrictions on protected materials, the interface provides access to two versions of EEBO-TCP: i) a public version, which comprises all openly accessible texts (EEBO-TCP Phase I only, 25,363 texts), and ii) a restricted-access version, which offers access to all texts in the collection (EEBO-TCP Phase I and Phase II, approximately 44,400 texts).[3] The URL for the public version of DREaM is http://dream.voyant-tools.org/dream/?corpus=dream. For a quick overview, please take a moment to view the following three-minute demonstration video:

Figure 1. DREaM demo video.

The creators of DREaM (Matthew Milner, Stéfan Sinclair, and Stephen Wittek) are members of the Digital Humanities Team for *Early Modern Conversions*, an international, five-year project that has brought together a group of more than one hundred humanities scholars, graduate students, and artists in order to study the tremendous surge of activity around conversion that followed from developments such as the Reformation, the colonization of the Americas, and increased interaction amongst European cultures.[4] Proceeding from the basic observation that conversion in early modernity was not an exclusively religious phenomenon, contributors to the project have endeavored to chart the movement and evolution of conversional thinking in the period, and to ask how the stories, spaces, and material affordances of conversion contributed to a conceptual legacy that has persisted into modernity (see [Hadot 2010, 1]; [Marcocci et al. 2015]; [Mills and Grafton 2003, xii–xv]; [Questier 1996, 40–75].[5] In order to accommodate the massive scope of this enquiry, the DREaM development team began to consider how one might apply “distant reading” techniques to a collection of texts like those made available through EEBO-TCP, with the long-term goal of lighting a way forward for similar investigations of other early modern corpora in the future [Moretti 2013, 3–4] [Jockers 2013, 48].[6] It is our intention to position our work with EEBO-TCP as a test-case for how scholars of early modernity might collect and create personalized corpora using resources and electronic archives of heterogeneous texts. In this respect, working with EEBO-TCP points to fundamental issues of corpus-building and textual analysis that become acute in an early modern context: orthographic regularity, thematic collation, and the use of metadata to allow easy assembly of new corpora for scholarly investigation. This kind of functionality manifests what scholars have long known about

archives: that they are driven by the interests and politics of users and the communities they represent. DREaM instantiates this archive-building theory into ready practice as a tool.[7]

3 The EEBO-TCP collection does not lend itself to computer-based analysis without a great deal of labour-intensive preparation and re-organization, primarily because of two distinct sets of problems: a) the relative inflexibility in the EEBO interface (see http://quod.lib.umich.edu/e/eebogroup/), which functions very well as a finding aid, but is a poor mechanism for compiling subsets of full texts; and b) orthographic irregularity. To get a sense of how these problems can impact a workflow, suppose a user wanted to analyze the 740 texts in the EEBO-TCP corpus dating from 1623 to 1625. Although identifying the desired material on EEBO requires nothing more than a simple date search, the process of bringing all the files together into a unified, searchable subset is considerably more difficult because the EEBO interface does not offer options for batch downloading, and therefore requires users to download files in a desired subset on a one-by-one basis. Once assembled, any subset created in this manner will also require significant editing because the plain text file one can download from EEBO does not come along with an accompanying file for metadata (i.e., information to indicate the title, date of publication, author, etc.). Rather, metadata appears within the file itself, a convention that can significantly complicate or contaminate the results of macro-scale analysis.

4 On a similar note, texts in the EEBO-TCP collection also feature a high degree of orthographic irregularity, a characteristic that adds an extra layer of complication for researchers who want to conduct any sort of query that involves finding and measuring words. As anyone who has ever read an unedited early modern text will be aware, writers in the period did not have comprehensive standards for spelling, so any given term can have multiple iterations (e.g., *wife, wif, wiv, wyf, wyfe, wyff*, etc.). Without a method for managing spelling variance, one cannot reliably track the distribution of words in a corpus, or perform many of the basic functions fundamental to textual analytics.

5 DREaM addresses both of these issues directly. Although the interface is similar in some ways to the interface for EEBO, it is much more flexible, and offers different kinds of search parameters, making the process of subset creation more powerful and convenient, and allowing users of the EEBO-TCP corpus to perform macro-scale textual analysis with greater ease. Secondly, and perhaps most important for corpus text-analysis, DREaM utilizes a version of the EEBO-TCP corpus specially encoded with normalizations for orthographic variants, a feature that enables frequency tracking across multiple iterations of a given term. In other words, DREaM can search orthographically standardized versions of texts *as well as* the original texts, and can produce subsets containing orthographic variations. This innovation solves a major issue for text-analysis of pre-modern literature. Designed to work seamlessly with Voyant Tools, in many ways DREaM can be viewed as a kind of archive-engine. It comprises an interface that allows quick and easy subset building and export of new corpora from pre-processed texts of EEBO-TCP. Conceptually, however, DREaM is a prototype that facilitates rapid creation of groups of texts for analysis around user-determined parameters.

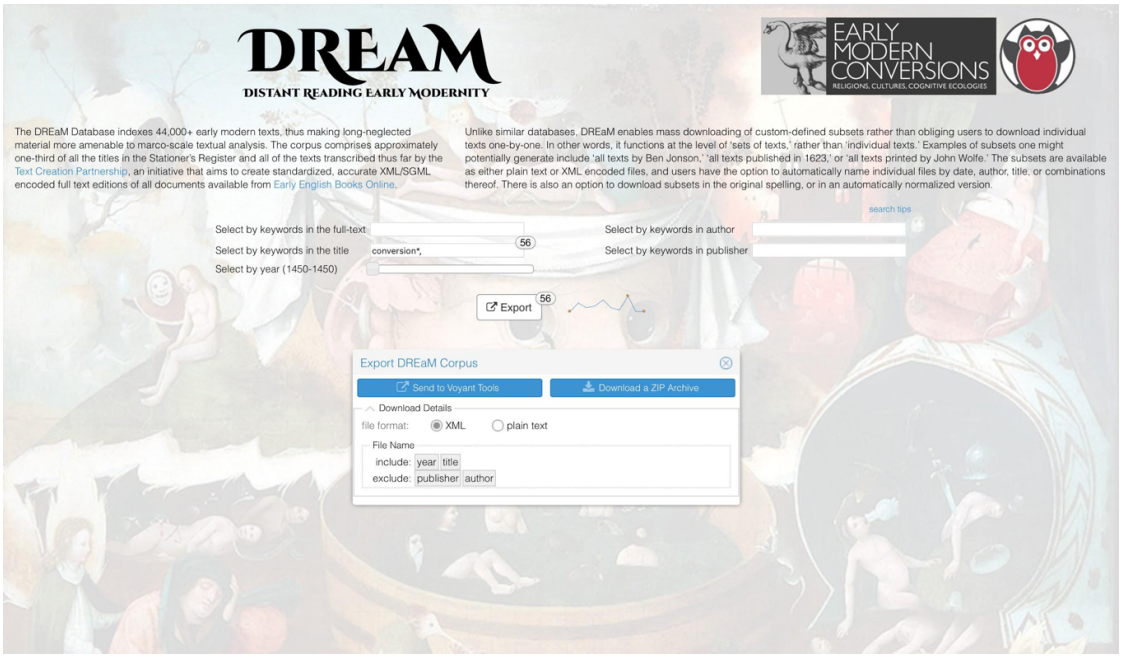

Figure 2. The interface for DREaM.

Figure 2 shows an example search on the DREaM interface. In the top half of the screen, there are four fields that one can use to define a subset in terms of keyword, title keyword, author, and publisher. Just below these fields, there is a horizontal slider with two handles that users can drag to define a date range. As one enters the subset parameters, a number appears in the top right hand corner of the Export button to indicate the number of texts the proposed subset will contain (for example, in Figure 2, the user has defined a subset of 56 texts that feature the term “conversion” in the title). To the right of the Export button, a thumbnail line graph offers a rough idea of text distribution across the date range (the graph in Figure 2 shows a significant peak toward the latter end of the range). 6

Clicking on the Export button will bring up a window where users can choose to download the subset as a ZIP archive, or send it directly in Voyant Tools, a multi-purpose environment for textual visualization and analysis. Users can also choose to download the subset as a collection of XML or plain text files. At the bottom of the Export window, a convenient drag-and-drop mechanism offers options for naming the files according to year, title, author, publisher, or combinations thereof, a functionality that facilitates custom tailoring for specific sorts of enquiries. For example, a researcher comparing works by various authors would likely want to put the author at the beginning of the file name, while a researcher tracking developments across a specific date range would likely want to begin the file name with the year. 7

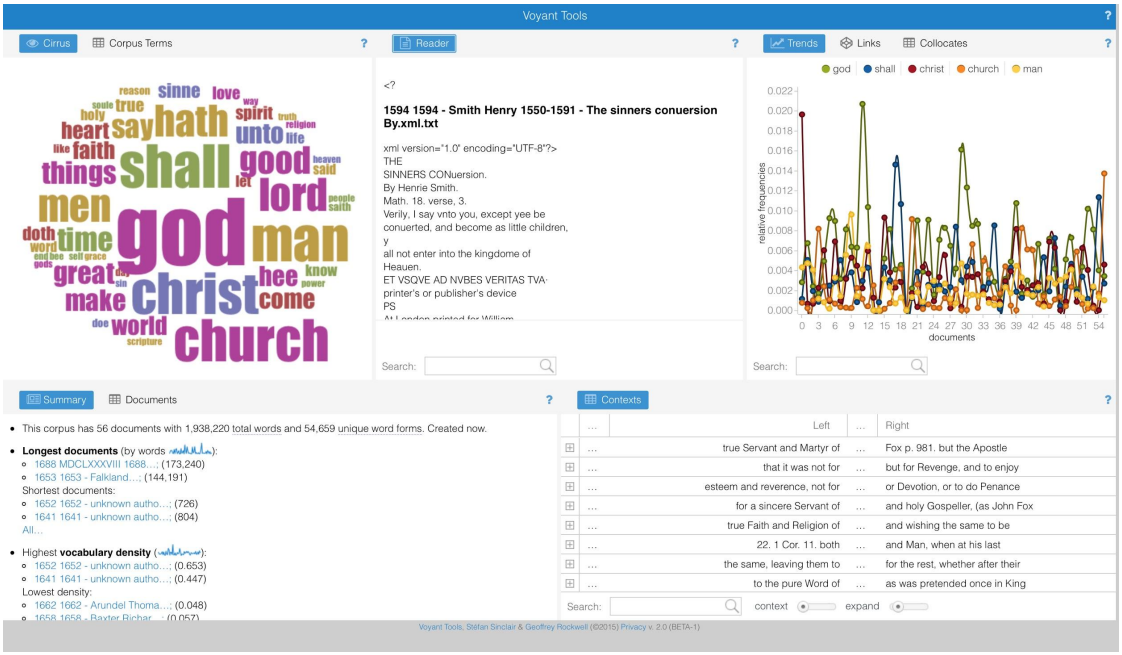

Figure 3. Voyant Tools.

Clicking on “Send to Voyant Tools” in the Export window will open up a new page that shows textual analysis from a suite of digital tools (see Figure 3). The tools displayed in the default settings are i) *Cirrus*, a visualization tool that correlates term frequency to font size (top left panel), ii) *Corpus Summary*, a précis of key frequency data (bottom left panel), iii) *Keywords in Context*, a tool that shows brief excerpts from the subset featuring a target term (bottom right panel), iv) *Trends*, a line graph that visualizes frequency data for select terms (top right panel), and v) *Reader*, a tool that enables users to scroll through texts in the subset and view highlighted instances of select terms (top center panel). Users can access further tools or adjust the arrangement of tools on the page by clicking on the Panel Selector icon in the top right corner of each panel. Notably, the tools in Voyant function inter-operably, so an action in one tool will carry over to the analysis for others. For example, if a user clicks on a term in *Cirrus*, a line graph for the term will appear in the *Trends* tool, and the *Keywords in Context* tool will provide a series of excerpts to demonstrate usage of the term throughout the corpus. This functionality enables researchers to switch back-and-forth very quickly between “distant reading” and “close reading” perspectives, and also makes it easier to follow up on unexpected discoveries, or explore specific items of interest on the fly. 8

In order to clarify the intervention that DREaM aims to bring to digital humanities research on early English print, it will help to briefly review two similar projects based around the EEBO-TCP collection, and to situate them in comparison to DREaM. The first is *Early Modern Print: Text Mining Early Printed English* (EMP), a project developed by Joseph Loewenstein, Anupam Basu, Doug Knox, and Stephen Pentecost, all of whom are researchers for the Humanities Digital Workshop at Washington University in St. Louis.[8] Like DREaM, EMP features a *Keywords in Context* tool, a *Texts Counts* tool (similar in function to *Corpus Summary* in DREaM), and a version of the EEBO-TCP corpus specially encoded with normalizations for variant spellings. Other key features include a *Words Per Year* tool and, most impressively, an *EEBO N-gram Browser*, which charts frequencies of a given word or short sentence using n-gram counts for each year represented in the EEBO-TCP corpus. The second project of note is the *BYU Corpora Interface for EEBO-TCP* (BYU-EEBO), a site created by Mark Davies at Brigham Young University.[9] As with the other interfaces developed by Davies for large corpora, BYU-EEBO visualizes term frequency data by decade across the full date range of the corpus, facilitates decade-by-decade comparisons of frequency data, and shows how the collocates of a given term evolve over time.[10] 9

Despite overlap of certain functions, DREaM, EMP, and BYU-EEBO represent very different responses to a growing demand for tools that reach beyond the conventional use scenarios envisioned by the designers of the EEBO interface 10

in the late nineties. Each project has distinct strengths and weaknesses. At the risk of over-generalization, one might say that BYU-EEBO caters primarily to linguistics research, while DREaM aims for a more open-ended, exploratory style of corpus interrogation — and EMP is somewhere in between. Although all three projects bring benefits of value to a growing field, it is important to note that, because it works seamlessly with Voyant Tools, DREaM is the only one that enables full, direct access to the texts in the EEBO-TCP corpus, a feature that allow users to check the source of data very quickly, or follow up on elements of particular interest. On a similar note, DREaM is also the only platform designed to work in conjunction with other tools. For example, a user could create a subset in DREaM, download it, and conduct analysis in other platforms, such as R or Python.[11]

Transforming & Enhancing EEBO-TCP

11 Our vision of an easy-to-use interface, and quick corpus creation required transformation and enhancement of the existing EEBO-TCP collection in two major ways. First, was the normalization, or standardization, of orthographic variants which are critical for the statistical analytics that power distant reading methods. Second was the enhancement of the metadata, allowing a richer set of parameters for building corpora of texts to analyze. Both were iterative processes. The resulting texts which power DREaM are new versions of the EEBO-TCP corpus that combine the EEBO-TCP metadata header with the outputs of each process: a full normalized version of each text containing the EEBO-TCP TEI SGML, with tagged normalizations, and new metadata drawn from linked open data provided by OCLC.

12 Before we could normalize the spelling of our 44,418-document corpus the texts need some pre-processing. Although the EEBO-TCP is primarily English, it also contains texts in Latin, French, German, Dutch, Spanish, Portuguese, Italian, Hebrew, and Welsh. To make things more complicated, some EEBO-TCP texts are multi-lingual. We decided to use VARD2, a tool built specifically for Early Modern English.[12] Although its creator Alastair Baron notes in the tool's guide that customizing VARD2 for another language is possible (it has been used successfully with Portuguese), we opted to work with the English-only sections of the texts.We used xPath to identify which documents contained English text elements `<text lang="eng">`. Since TCP allows for multiple values in the language attribute, we initially identified potential texts if "eng" appeared as a value of the attribute, e.g. `<text lang="eng lat ita">`. The result was a working set of 40,170 documents which contained declared English text in some place or another.

13 In EEBO-TCP `<text>` elements can be nested. Combined with multilingual values for the language attribute of the element, this creates potentially complicated conditions for extraction. We quickly found that there were 128 possible xPath locations of `<text>` in the TCP SGML documents. Here are a few examples of these locations:

```
/EEBO/ETS/EEBO/GROUP/TEXT /EEBO/ETS/EEBO/TEXT /EEBO/ETS/EEBO/TEXT/GROUP/TEXT
/EEBO/ETS/EEBO/TEXT/BODY/DIV1/P/TEXT /EEBO/ETS/EEBO/TEXT/BODY/DIV1/Q/TEXT
/EEBO/ETS/EEBO/TEXT/BODY/DIV1/DIV2/P/Q/TEXT
/EEBO/ETS/EEBO/TEXT/FRONT/DIV1/Q/TEXT
/EEBO/ETS/EEBO/TEXT/FRONT/DIV1/P/TEXT
/EEBO/ETS/EEBO/TEXT/GROUP/TEXT/BODY/DIV1/DIV2/P/NOTE/P/TEXT
```

14 Since we were only interested in normalizing English text, we had to isolate the `<text lang="eng">` elements, and preserve their order. As noted by the following examples, where the parent was English-only, we extracted the parent; however, when `<text>` elements contained multilingual language attribute values extracted English-only children.

`<text lang="eng"> <text lang="eng"> </text> </text>`	`<text lang="eng lat"> <text lang="eng"> </text> <text lang="lat"> </text> </text>`

Table 1.

The script we created concatenated the matching <text lang="eng"> elements, and remarried the existing EEBO-TCP metadata file (*.hdr in the EECO-TCP file dump) to the new English-only “body” to produce a corpus of 40,170 truncated texts that could stand on its own, or be used for orthographic normalization. 15

Normalizing the 40,170 English-only texts was iterative, as we optimized and tweaked our method. Prior to producing any final sets using VARD2, we ran it on the entire English corpus, and edited the dictionary by hand, catching some 373 normalizations which were problematic in one way or another. These amounted to 462,975 changes overall — only 1.03% of the total number of normalizations. Several examples are illustrative of the problem normalizations: “strawberie” became “strawy” rather than “strawberry,” and “hoouering” became “hoovering” rather than “hovering”, effectively creating a word that simply didn’t exist in period English. VARD’s statistical method is not robust enough to analyze the particular context that would allow it to discern which was appropriate: more ambiguous words like “peece” could either be “piece” or “peace”. In such cases, we decided normalization was too problematic, and thus set VARD to ignore “peece” entirely. 16

Our pre-processing script not only extracted the <text> elements, it also altered them prior to loading the texts into VARD2. We experimented with both the texts we input into VARD2, and with its orthographic normalization parameters, creating twelve overall versions of the EEBO-TCP English corpus. The purpose was straightforward: to ascertain what degree of normalization best balanced the contextual ambiguities like “peece” but achieved a level of optimal normalization for text-analysis tools like Voyant, and to see whether limited pre-processing of the EEBO-TCP texts themselves, prior to VARD2 processing, improved the results. We ran each degree of normalization on three versions of the English-only <text> elements: a “Regular” unedited version acted as our control; a “Cleaned” version which removed characters and tags that impeded VARD’s normalization process by splitting words; and an “Expanded” version that expanded the macron diacritic, typically used in early modern English to represent a compressed “m” or “n” on the preceding vowel (e.g. “comitted” became “committed”).The “Cleaned” version removed pipe characters | and editorial square brackets [], but also <supr> and <subs> tags which broke up words. We opted not to remove any <GAP DESC="illegible"...> elements as it quickly became apparent that doing so would create more problems than it might resolve. Although VARD could likely handle a single missing character, the highly variable nature of a text gap made it questionable what gaps we should let VARD “patch”, and which it could not: two characters, or only a single character? Is that character a word on its own? We felt that the subjective nature of matching whether legibility was accurately assessed or not, and where to set the bar for VARD normalization, made the removal of these elements unpredictable enough that it would confound later analytical interests. 17

We ran each of these versions in VARD2, in turn, at four distinct “match” levels in order to assess what level seems best for the overall corpus. VARD normalizes words if it finds a match that is +0.01% higher than the match parameter a user sets: a 50% setting will only alter a variant with a statistical match of 50.01%. We first ran VARD at 50%, it was evident this excluded a large number of obvious variants we needed normalized which appeared between 45% and 50%. We re-ran the normalization at 45%. We also ran VARD2 at 35% and 65% both as controls, with a mind to producing sets of the English-only corpus that might help us determine whether normalization levels were best adjusted as we proceeded chronologically in the corpus. In each case we set the output to “XML” so that VARD2 would tag the normalizations, retaining the original spelling as an attribute in the surrounding <normalized> tag. These tags were critical later on for indexing the texts in DREaM itself. In the end we decided to use the 45% match “Cleaned” version of the English only EEBO-TCP corpus for DREaM because it seems to have the best balanced normalization following the removal of tags that prevented VARD2 from working correctly, but without any expansions or “guess work” resulting from replacement of <GAP> elements. When collating the final file, we noted the set name, as well as the date and match level, as attributes in a wrapper element that enclosed the normalized text. We then coupled the entire XML text to the EEBO-TCP TEI header, and enclosed everything once again in an <EEBO> element. 18

VARD2 itself was fairly easy to use. Even so, it took some doing to ensure it operated smoothly with the variable sizes of EEBO-TCP texts. Although it has a command line batch mode, we quickly ran into trouble as VARD2 would crash handling 5MB texts on the recommended memory settings of “–Xms256M –Xmx512M”. The crashes occurred frequently enough, despite raising the memory settings to over 1GB, that we decided to run VARD2 through our own php script, executing the program once for each of the 40,170 input files, with the appropriate setting. Even then, 1-3GB 19

memory settings were insufficient to handle the ~100 or so EEBO-TCP files that are over 10MB. We ended up running our VARD2 processing script over a four-day period at "-Xms6000M -Xmx7000M" – or, significantly higher than the recommended settings. Undoubtedly this took longer than batch mode, but it was more stable.

Enhancing the EEBO-TCP metadata was also iterative, but did not use a specific tool; rather, we processed the metadata using a combination of php, mysql, and text files to build gazetteer data. The existing metadata files of the EEBO-TCP (*.hdr) contain a wealth of information about the file encoding process, as well as several types of identifiers. They also employ a standardized or canonized list of authors and publication places, as well as publication dates. Though the metadata contains the critical identifiers from the Short Title Catalogue, it lacks data that is now available via Open Data resources like OCLC and VIAF that is often of interest to scholars using the EEBO text, such as gender of authors, or better dates of birth and death. Moreover, in both OCLC and EEBO-TCP metadata, the <publisher> remains an unparsed string despite containing a wealth of information on publication such as publishers, dates, and historic addresses. 20

The enhancement of the metadata was two-fold. First, it was to find possible matching OCLC records and IDs for EEBO-TCP files, and pull in OCLC data to create a new metadata header containing authority dates (some <date> elements in EEBO-TCP contained artifacts like the letter L for the number 1 and the letter O for the number 0), places of publication, and titles, as well as data like gender, and birth and death dates from VIAF for EEBO-TCP authors, and referencing the OCLC and VIAF IDs as online resources in the new XML. Second, to identify individuals, places, and addresses in the unparsed <publisher> element. 21

Matching EEBO-TCP texts to OCLC records was not as straightforward as it might appear. EEBO-TCP comprises textual witnesses or instances, while OCLC collates records from its partner libraries in order to build records that are manifestations or "works". This conceptual distinction is critical as it means that there might well be several OCLC IDs for an individual EEBO-TCP text, potentially with variable metadata. Initially we had thought it possible to obtain a dump of the OCLC using EEBO as a "series" in our own McGill Library catalogue, allowing quick matching of the OCLC IDs with the EEBO texts. It turns out this was not possible, and so we opted to employ a combination of OCLC's WorldCat Search API and web page searching using the titles of each EEBO text to create lists of possible matching OCLC IDs. Using OCLC's xID service, we compared possible OCLC matches with the EEBO-TCP metadata using the title, dates of publication, and authors. Matching was a matter of confidence: titles were compared using both metaphone and levenshtein distances, to create a confidence level. We did the same with publication dates, as well as places of publication, where present. In the case of authors, we employed the same method (in order to account for spelling variants like smith vs. smythe), but also tallied the resulting matches to ensure that when EEBO-TCP noted four authors, an OCLC match did the same. We created a strict scoring system based on levenshtein distances for authors and titles, and exact matching for dates and places of publication (if they were noted). The same parameters were used to create a score for both the EEBO-TCP metadata and a possible OCLC match: we considered a high confidence match an equal score, or within 1 of the original EEBO-TCP. Inevitably this excluded some possible OCLC candidates, but it resulted in high levels of confidence matching of OCLC IDs for c. 39,000 of the 44,418 texts in the full EEBO-TCP corpus. With these OCLC IDs, we produced the first revised version of the metadata, pulling in information from the linked VIAF records to flesh out authorial data like dates of birth and death, and gender (EEBO uses TEI, which employs a <sex> element, rather than gender, to describe this information despite the problems inherent with sex / gender distinction). This version was coupled to the English only corpus we processed with VARD2, along with a truncated version of the original EEBO-TCP metadata. Combining the new metadata, and the original, in a larger metadata header for the files, gave us the ability to present users with the option of searching using the original EEBO-TCP metadata, or the hybrid OCLC-EEBO-TCP metadata. 22

The second version of the DREaM metadata, which we produced in early fall 2015, took up the challenge of parsing the <publisher> element which remains unparsed in both OCLC and EEBO-TCP metadata. EEBO-TCP A00257 provides a good example of this string: <publisher>By Iohn Allde and Richarde Iohnes and are to be solde at the long shop adioining vnto S. Mildreds Churche in the Pultrie and at the litle shop adioining to the northwest doore of Paules Churche,</publisher> John Allde and Richard Jones are not mentioned anywhere in the original EEBO-TCP metadata, only the author <author>H. B., fl. 23

1566</author>. Isolating these individuals required creation of two gazetteers: first, for place names, and second, one for known agents (from the EEBO-TCP author list). Rather than working with 44,418 entries, we ran our script iteratively over the 23,644 distinct <publisher> strings, adding the possible places and individuals to the growing gazetteers, and pulling in variants for authors' names from VIAF's RDF XML files (<schema:alternateName>) retrieved by using the results from VIAF's AutoSuggest API (http://viaf.org/viaf/AutoSuggest?query={searchterms}). We also created a short list to translate common early modern first names and abbreviations into modern versions, such as Io. for John, or Wyliam for William. After some 20 passes over the data, including manual editing of the agents gazetteer, patching for new scenarios, and comparing possible matches to the dates of publications in EEBO-TCP texts, we were left with a gazetteer of 195,213 variants representing 24,076 distinct possible names for 19,836 distinct VIAF IDs. Some, like "F.M.", were too imprecise to resolve. Our work also alerted OCLC to problems with their AutoSuggest API when it returned the canonical names of an author with another individual's VIAF ID, usually for co-author (e.g. Nicholas Bourne's VIAF ID appeared for Thomas Goodwill). This required additional processing of incoming data to see if the names it returned matched those we wanted to query. Equally problematic were dates, which appeared inconsistently throughout the VIAF open data. RDF XML might lack any dates, while the VIAF "cluster" XML would contain it under <ns2:birthDate> and <ns2:deathDate> or as part of a MARC individual agent record element <ns2:subfield code="d">. Many values in these locations were "0" despite the presence of birth and death years as part of the canonical name; this became a third option for checking whether someone could be the author of the text in question. Last resort was using the decades of publishing activity indicated in the Cluster XML as <ns2:dates max="201" min="150">. Consequently, all of these matches have a confidence marker.[13] The latest version of the DREaM metadata for A00257, from EEBO-TCP Phase 1, is available as part of our github GIST. The metadata of the new <dreamheader> is richer than the original EECO-TCP metadata, containing not only the biographical data of VIAF records for the author, H. B., but also similar data for the individuals found in the <publisher> element.

24 The addition of confidence markers in our new DREaM metadata indicates a different approach to open metadata. Rather than seeing metadata as an authoritative or concrete accounting of actual attribution and representation (as is the practice of archivists and cataloguers), the DREaM metadata (especially in regards to the parsed <publisher> data) should be viewed more as a kind of contingent scholarly assertion. DREaM does not have the resources to double check all some 40,000 texts to ensure exact accuracy of the matching of EECO-TCP metadata with VIAF and OCLC identifiers. In many cases, doing so requires expert domain knowledge, and means to accurately resolve entities. A good example of this is "B. Alsop": is it Bernard or Benjamin Alsop? The two printer publishers were most likely related, but in some instances it isn't possible to distinguish between the two, as in the case of EEBO-TCP text A00012, Robert Aylett, *Ioseph, or Pharoah's Favourite*, printed by B. Alsop for Matthew Law … (1623), because VIAF lacks birth and death dates for both. Experts know it is Bernard Alsop, but without a corroborating data source there is no method, programmatically, to distinguish between the two as matches for "B. Alsop". By documenting both, and marking a confidence level, we're asserting that metadata is very much the product of ongoing research: it should not be seen as definitive. While DREaM allows researchers the ability to create corpora based on either exact or fuzzy searches, we're also publishing this metadata separately, and offering it to EEBO-TCP so that the wider scholarly community can refine and critique it.[14]

DREaM as Archive Engine: Enhancements to Voyant Tools

25 DREaM has led to the enrichment and enhancement of the EEBO-TCP corpus, but the project has also led to some significant architectural and functionality improvements in Voyant Tools.

26 In particular:

- a skin designed specifically for subsetting of a very large corpus based on full-text and metadata searches (the current DREaM skin is specific to EEBO-TCP but the underlying design and functionality can be reused);
- efficient querying of a corpus to determine the number of matching documents, much like a search engine (previously functionality was limited to term frequencies);
- support for NOT operators to filter out documents that match a query

- additional native metadata indexing (e.g. for publisher and publication location) as well as the ability for user-defined metadata fields that can be included in subsequent queries;
- exporting full texts from Voyant in compressed archives of plain text or XML, with user-defined file-naming protocols;
- efficiency-optimized creation of a new corpus subsetted from an existing corpus.

In addition, work on DREaM helped prioritized other planned functionality, such as reordering documents in a corpus, editing document metadata, access management for corpora (to reflect restrictions for EEBO-TCP), and overall scalability improvements. We had not yet worked on a single corpus with more than 44,000 texts of variable lengths (the input XML for DREaM weighs in at more than 10GB). In short, DREaM provided an ideal test bed for efforts to enhance the scalability of Voyant Tools. This use-case driven development seems to us an ideal scenario for building new generations of resources in the Digital Humanities.

Notes

[1] Support for DREaM and the Early Modern Conversions Project comes from the Social Sciences and Humanities Research Council of Canada (SSHRC), the Canada Foundation for Innovation, McGill University, and the Institute for the Public Life of Arts and Ideas (McGill).

[2] Early English Books Online brings together page images — but not necessarily transcriptions — of all English printed matter from 1473 to 1700. Much of the collection derives from early microfilm photographs created by Eugene Powers in the 1930s. There are approximately 125,000 titles in the collection. The Text Creation Partnership has completed transcription work on approximately 44,000, or one-third of all texts currently available.

[3] The URL for Early English Books Online (EEBO) is: www.eebo.chadwyck.com/home. For information about the Text Creation Partnership (TCP), see: www.textcreationpartnership.org.

[4] Early Modern Conversions: earlymodernconversions.com

Early Modern Conversions Digital Humanities Team:http://www.earlymodernconversions.com/people/digital-humanities-team/.

[5] Pierre Hadot wrote that conversion is “one of the constitutive notions of Western consciousness and conscience,” arguing that, “in effect, one can represent the whole history of the West as a ceaseless effort at renewal by perfecting the techniques of ‘conversion,’ which is to say the techniques intended to transform human reality, either by bringing it back to its original essence (conversion-return) or by radically modifying it (conversion-mutation).”

[6] As Franco Moretti has argued, distant reading creates possibilities for analysis where no other option exists: “A canon of two hundred novels, for instance, sounds very large for nineteenth-century Britain (and is much larger than the current one), but is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows — and close reading won’t help here, a novel a day every day of the year would take a century or so.” Matt Jockers makes a similar point: “Macroanalysis is not a competitor pitted against close reading. Both the theory and the methodology are aimed at the discovery and delivery of evidence. This evidence is different from what is derived through close reading, but it is evidence, important evidence. At times, the new evidence will confirm what we have already gathered through anecdotal study. At other times, the evidence will alter our sense of what we thought we knew. Either way the result is a more accurate picture of our subject. This is not the stuff of radical campaigns or individual efforts to ‘conquer’ and lay waste to traditional modes of scholarship.”

[7] This view of archives as collections of interest, connected to deeply embedded epistemological positions and cultural politics, as much as what is archivable, owes much to the principles of Michel Foucault’s *Archaeology of Knowledge* and Jacques Derrida’s *Archive Fever*. [Manoff 2004] provides a useful overview of the field. See also [Parikka 2012].

[8] See http://earlyprint.wustl.edu.

[9] See http://corpus.byu.edu/eebo.

[10] For Davies’ other corpus interfaces see http://corpus.byu.edu/overview.asp.

[11] For R, see https://www.r-project.org; for Python, see https://www.python.org.

[12] http://ucrel.lancs.ac.uk/vard.

[13] Confidence was a matter of dating. “0” denotes an exact match or a match with a EEBO-TCP author, with a publication date which falls in between an individual’s birth and death dates; “1” lacked either birth or death dates, or used publishing activity dates; and lastly “2” lacked any dates.

[14] http://www.matthewmilner.name/2016/05/06/EEBO-TCP-Phase-I-Metadata-Mashup-revision-II/

Works Cited

Derrida 1996 Derrida, Jacques. *Archive Fever.* Chicago: University of Chicago Press, 1996.

Foucault 2002 Foucault, Michel. *Archaeology of Knowledge*. London and New York: Routledge, 2002.

Hadot 2010 Hadot, Pierre. “Conversion.” Translated by Andrew B. Irvine. Accessed September 21, 2015. https://aioz.wordpress.com/2010/05/17/pierre-hadot-conversion-translated-by-andrew-irvine/. Originally published in *Encyclopaedia Universalis*, vol. 4 (Paris: Universalis France), 979-981.

Jockers 2013 Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. Chicago: University of Illinois Press, 2013.

Manoff 2004 Manoff, Marlene. “Theories of the Archive from Across the Disciplines”, *portal: Libraries and the Academy*, vol. 4, no. 1 (2004), 9–25.

Marcocci et al. 2015 Marcocci, Giuseppe, Wietse de Boer, Aliocha Maldavsky, and Ilaria Pavan, eds. *Space and Conversion in Global Perspective.* Leiden: Brill, 2015.

Mills and Grafton 2003 Mills, Kenneth, and Anthony Grafton, eds. *Conversion: Old Worlds and New*. Rochester, N.Y.: University of Rochester Press, 2003.

Moretti 2013 Moretti, Franco. *Distant Reading*. London: Verso, 2013.

Parikka 2012 Parikka, Jussi. “Archives in Media Theory: Material Media Archaeology and Digital Humanities,” in *Understanding Digital Humanities*, ed. David M. Berry. Basingstoke: Palgrave MacMillan, 2012: 85-104.

Questier 1996 Questier, Michael. *Conversion, Politics and Religion in England, 1580-1625*. Cambridge, U.K.: Cambridge University Press, 1996.