Towards a digital carrel:
A report about corpus query tools

1 Finding aids in a predigital and digital world

This is a report about corpus query tools and more particularly about lowering the entry barriers to their use by scholars in text-centric humanities disciplines such as history, literary studies, and philosophy. It is written for an audience of such scholars, tries to avoid technical terms of art as much as possible, and provides explanations for anything that might require explanation at a meeting of an English or History department. It often talks about 'readers' and 'scholars' rather than 'users' to keep in mind that texts, however processed, in the end are there to be read and that for individuals with a scholarly cast of mind reading will remain a prime activity. The scroll, codex, or screen are tools with different 'affordances', and the experience of reading with their help will differ.[1] But there is also a way in which reading remains reading. Few scholars in the humanities will want to transcend it.

Corpus query tools of one kind or another, as well as statistical routines associated with them, have been around for almost two generations. They have served as fundamental tools in Natural Language Processing (NLP), the manipulation by machines of human language, whether spoken or written. With the explosive growth of the Internet over the past two decades NLP has moved far beyond its disciplinary home in corpus linguistics. It is a close relative of information retrieval (IR), and many of its basic routines undergird the many professional, business, and political activities for which increasingly sophisticated forms of information retrieval have become a daily necessity. The rapid pace at which library schools are refashioning themselves into "i-schools" is a telling sign of changing times.

1.1 Scale

Change of scale matters a lot. Here is a paragraph from the preface to *Corpus Concordance Collocation* by the British linguist John Sinclair:

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular.

That was written twenty years ago. The British National Corpus, a stratified sample of 100 million words of contemporary English, was released in 1994. Over the past year, Mark Davies at Brigham Young University has released two 400 million word corpora of Historical American English (1800-) and Contemporary American English (1990-2010).

The Text Creation Partnership, probably the most ambitious academic transcription enterprise, will by 2015 produce digital versions of some 70,000 public domain texts of

[1] 'Affordance' is a term of art from psychology. Google's Ngram Viewer gives a good idea of its increasing use since the sixties. It is close in meaning to 'properties' or 'characteristics' but focuses on the properties of a thing that let you do something with it. The OED cites J. J. Gibson, who coined the term and defined it as "what things furnish, for good or ill." The term will grate on some ears, but it is a useful term for discussing what corpus query tools can "furnish, for good or ill."

English texts before 1800, with a word count of somewhere between 5 and 10 billion words. Google has released its n-gram corpus, which provides limited access to 500 billion words in some five million books.[2] There are trillions of words on the Internet, with billions added by the day.

Two decades is a reasonable time for a scholar to progress from finishing a dissertation to being promoted to full professor. It is difficult to think of another period in which the conditions of access to the documentary infrastructure have changed as deeply and quickly as in the past twenty years. Working successfully within any text technology -- whether scroll, codex, printed book or digital file -- depends on a realistic calculus of the possible. Scholarly habits acquired in graduate school are not easily shed or adjusted. The best tool is often the tool you know best. What is a humanities scholar to do when even David Pogue, the technology correspondent of the *New York Times*, admits that he cannot keep up?[3]

1.2 Beyond simple searching

A good corpus query tool should do more than ‘simple’ searching. A simple search is simple from the user’s perspective: you put one or more words in a search box, which is very much like looking up words in a dictionary or the index of a book. Such a search may or not be simple in terms of how the search engine interprets the request or returns its results. It is indeed a simple search if the search engine takes the characcter string ‘love’, looks for all its occurrences as a whole word in the documents that are searched, and returns either the documents in the order in which they were searched, with the word ‘love’ highlighted or a concordance output in which each ‘hit’ or occurrence of the search term is surrounded by a few words on either side.

If you look for ‘jealousy’ in the EEBO corpus, the list of hits includes passages in which the word appears in different spellings, such as ‘ielosie’ or ‘Ielousye'. The reader’s life has been made a little easier at the cost of complicating what the machine does: it has been given a list of variant spellings that are bundled under the modern standard spelling. If you Google ‘jealousy’, you are told that it retrieved 35,900,000 records in 0.20 seconds. It gives you a top list of hits that begins with a Wikipedia entry and is followed by a list of entries from pop psychology Web sites. The procedures leading to that result are anything but simple. They are also not very transparent, but seem to follow the inscription on the gold casket in *The Merchant of Venice:*

Who chooses me shall gain what many men desire (2.7.5)

Wherever your choices follow that goal Google will provide useful answers. There are many occasions in everyday, professional, or scholarly where you want just that. It not only helps with buying cameras or refrigerators but is useful as a first orientation in a research project. A search for ‘jealousy’ in Google Scholar provides you with a list of articles and books ranked by the frequency with which they have been cited.

[2] An n-gram is a sequence of one or more words, such as a unigram, bigram, trigram, etc.

[3] “The Lessons of 10 Years of Talking Tech,” New York Times, 25 November 2010 (http://www.nytimes.com/2010/11/25/technology/personaltech/25pogue.html?_r=1&ref=davidpogue).

Simple searches of this kind -- simple from the user's end, but increasingly complex in terms of the machine's response to a request-- will always be a cardinal feature of any query environment. Far more people walk than run or go on strenuous hikes, and even runners and hikers spend more time walking than hiking or running. More complex tasks are always embedded in simpler tasks from which they emerge and to which they return.

Corpus query tools deliver more than simple searches. They support what Lou Burnard calls "robust searching." They may not play a very important role in the way in which scholars and scientists find their way around an ever-growing secondary literature. They play a critical role in the scholarly engagement with primary documents or, more accurately, documents considered as primary in a particular inquiry.

Robust searching involves requests or questions like the following:

1. What is the most frequent noun in this corpus or some specified subset of it?
2. Find a random sample of 100 instances of 'fish' followed by 'chips' within four words.
3. Find all variant spellings of 'lieutenant' or 'gentleman' in texts before 1650.
4. Find sentences beginning with a conjunction or beginning with 'well' and ending with a question mark.
5. Find all n-grams that match the syntactic structure of "handsome, clever, and rich" or "the faster the better."
6. Do men use colour vocabulary differently from women?
7. Which verbs collocate significantly with 'bosom' at different periods of history?
8. What n-grams are use significantly more often or less often in a specified subcorpus when compared with the entire corpus or some other sub-corpus?
9. What n-grams longer than three words or more are shared by two texts but do not occur or occur only rarely in the rest of the corpus?
10. Given a set of texts (sermons by Launcelot Andrewes, novels by Maria Cummins) find texts that are most or least like it and list the n-grams or other discriminating factors in order of descending discriminant power.

Answers to questions of this type depend on the degree to which the source data in the corpus are 'curated' by 'annotating' them with appropriate metadata at different levels of texts For the types of question listed above, such annotation can be applied automatically, quickly, and with tolerable accuracy to corpora running into hundreds of millions, perhaps even low billions of words. [4]

[4] This statement is based on the experience of the MONK project, where 150 million words in 2,800 texts between the early 16th and late 19th centuries were lemmatized and linguistically annotated.

1.3 Finding aids in a predigital world

Corpus query tools belong to a larger class of finding aids that help you find stuff in some body of materials.[5] While they are creatures of a digital world, in a report for humanists it is useful to begin with their precursors in a world of scrolls, codices, or printed books. A book is a container that holds content and provides access to it via the hands and eyes of human readers. The size of letters, the length of lines, the number of lines on a page, and the aggregate weight of pages, including their covers are limited by what eyes can see or hands can hold. As you move from one book to many and the shelves needed for their storage you need to consider the limits of how far arms can reach or feet can walk. Anything that stretches or overcomes these limits may be called a corpus query tool. Ladders have their place in the hierarchy of such tools --witness Spitzweg's picture *The Bookworm*.[6]

1.4 Library catalogs

Library catalogs are at the top of the hierarchy of finding aids, and in a broad sense of the term a library catalog is the corpus query tool par excellence. The Hellenistic poet Callimachus is supposed to have been the first formal bibliographer. Surviving fragments of his *pinakes* or catalogue tables testify to a crude classification scheme for the Alexandrian library that remained influential well into the nineteenth century -- a useful reminder that poetry and pedantry are ancient and perhaps necessary bed fellows. Callimachus was a geek who made a major contribution to text technology.

1.5 Finding aids inside the text

Bibliographical data or data about a text as a whole are known to librarians as 'metadata'. For a text of any length some internal division helps readers find their way around it. Such divisions, which can be thought of as text-internal metadata, have a rich history and typically depend on what in computer parlance is called a "tree structure" or "ordered hierarchy of content objects" (OHCO). Divisions may be authorial or imposed by posterity. Vergil composed his *Aeneid* in twelve books, but the division of the *Iliad* and *Odyssey* into twenty-four books is the work of Hellenistic editors. The *Summa Theologiae* is built from the ground up as an "ordered hierarchy of content objects." One of the first tasks facing future Aquinas scholars is to find their way around citation paths

[5] 'Finding aid' is a term of art that normally refers to a tool that helps you find a book in a catalogue, while a corpus query tool helps you find words in books. But it is an important theme of this report that digital technology blurs the line between looking for books and looking for stuff in books. 'Finding aid' is here used in its literal and broadest meaning of anything that helps you find something, whether books or words in books.

[6] So do the Renaissance book wheels and Jefferson's Lazy Susan version, which is a useful reminder that all finding aids are Lazy Susans of a sort.

like “IIa.IIae. q.1 a.2", which describes the one and only direct path from the “root” of the tree to the “leaf node” of the passage in question.

Drama is the genre in which a particularly rigid and closely observed system of text-internal metadata is constitutive of the genre itself. In early modern drama you find a “stage Latin” with a “controlled vocabulary” (more technical jargon) of terms like *actus, scena, exeunt, manent,* or *solus*. *Exit* is the one term that has survived into the modern world.

Print technology greatly increased the complexity and granularity of devices for navigating within a text. There are tables of content with numbered sections and page references. Printed books are more likely than manuscripts to have page numbers, and they draw on the reader’s command of elementary arithmetic. In a book of 300 pages, p. 150 is in the middle and p. 297 very much towards the end. If you don’t know this, it takes a lot longer to find the page.[7] Running headers or footer provide more regional orientation.

Consistent pagination across a print run makes indexes worthwhile. Indexes of names, places, and key concepts provide metadata about the bottom level of a text. They take you to the words in the text and support various forms of discontinuous reading that conceptually are not very different from what geeks call “text mining.” In sum, the query space of any scholarly discipline in the print world has always depended on an elaborate, if fuzzy, apparatus of bibliographical, structural, and lexical metadata at the top, mid, and bottom levels of particular texts. Imagine all the printed books in Discipline X suddenly stripped of these metadata and presenting themselves to their readers as “plain text” files. The discipline would grind to a halt very quickly. This is an important point to keep in mind as we look ahead to a world in which the query space of particular disciplines will be increasingly digital.

1.6 Citation schemes

Citation schemes deserve special mention in the ecology of a print-based scholarly information space. The more canonical a text the greater the need for consistent and fine-grained reference schemes. The division of the bible into the roughly one-page chunks known as chapters dates back to the Middle Ages. The division these chapters into ‘verses’ or units more or less at the sentence level is the work of the French printer Henri Stephanus and appears first in a Bible of the 1550’s .[8] Twenty years later the same publishing house produced the citation scheme that is still used for Plato. In Stephanus’ 1578 edition each page is divided into quintiles marked with a lower case letter from ‘a’ to ‘e'. A reference such as *Republic* 484c identifies a word or phrase as occurring somewhere in the middle of page 484 of that edition. The citation scheme stuck and is replicated in any

[7] One of useful side effects of physically handling a book is that you always know where you are. A lack of comparable and tacit orientation is probably a major source of discomfort when reading on a screen.

[8] References like John 3:16 or 1Corinthians 13:12 are not only finding aids but over time shape the nature of the discourse about their referents by creating the impression that every biblical verse is a self-contained utterance.

scholarly edition so that *Republic* 484c can be readily found in any modern edition regardless of its pagination.[9]

Citation schemes of this precision allow for a much quicker move from an external reference source to textual passages. Medieval monks implemented a primitive version of a concordance. From a conceptual perspective, it contains the fundamental ingredient of a textual database, where a text is divided into an inventory of its smallest parts and each occurrence of a part is given a unique address. Consistent citation schemes are the prerequisite for ancillary scholarly tools like concordances or lexica. Such tools are the print equivalent of what in computer parlance is known as "stand-off markup," which refers to any annotation of a particular location range in a digital file that is not kept with the file (like a marginal gloss in a manuscript) but is stored separately and linked to its target through a reference that is likely to be even longer and less intuitive than the citation of a passage in Aquinas.

1.7 Back to digital finding aids

A digital environment blurs the difference between searching within or across books. A machine readable cataloging record (MARC) has about 200 distinct fields, although only four dozen are regularly used. It is a thumbnail sketch of a book. Forty years ago, it was a tremendous task to store and manipulate low millions of such thumbnails on a mainframe computer. Today, it is entirely possible to put the entire book inside its thumbnail as a separate data field. The resultant library is an OHCO (ordered hierarchy of content objects) that lets you find books by traversing a tree structure in one of two directions.[10] You can travel down from the root of the catalog to a particular title (visual representations of tree structures usually put the root at the top). You can also get to the title by moving up from the "leaf nodes" of some words in text.

Because language is a phenomenon with "large numbers of rare events" (LNRE), the odds of finding two words in the same document are quite low once you move outside the small base vocabulary of any language. The odds decline precipitously with three, four, or more words. These LNRE properties of human language account for the effectiveness of unstructured searches. You start from some combination of words you remember because the time cost of ruling out false positives is often less than the time it takes to conduct a more properly structured search. If you read Herodotus, are struck by the strange sexual customs he describes in his chapter on the Lydians, and want to find out more about them, putting 'Lydian', 'temple', and 'prostitution' in the search field of Google or Google scholar, will give you good enough results much more quickly than a structured bibliographical search.

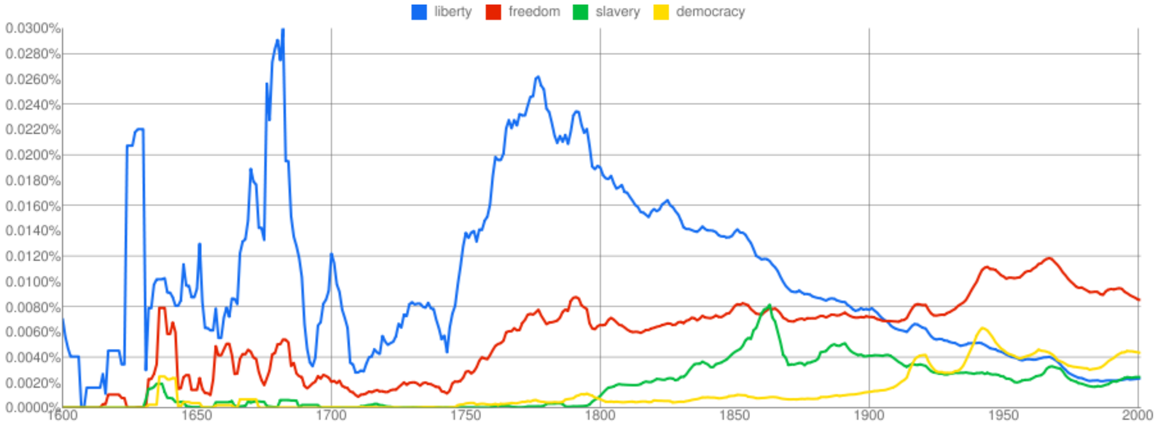
[9] In modern editions the scheme has of course lost some of its expressive power as a crude visualization scheme. References ending with 'a', 'c', or 'e' are no longer found at the top, middle, or bottom of a page, a reference may stretch across page breaks, and if the Stephanus page references are merely given as ranges in a running header, readers may have to search across an entire page rather than a range.

[10] The hierarchy below the bibliographical level will be very flat and descend from the page 'children'of the title to the word children of the page.

1.8 Working with corpus query tools

For many humanities scholars, the use of digital tools is dominated by the library catalog paradigm: you use the digital tool to find what you need, and then you read what you have found in the ordinary way. Corpus query tools are more like the concordances, lexica, or other ancillary tools that in some disciplines are your constant companions and deeply shape your manner of engagement with the primary data. Such tools help scholars explore the play of differences that are at the heart of human language -- witness the linguist's favourite sentence "the cat sat on the mat."

The drafting of this report coincides with the release of the Google Book Ngram Viewer, which is a corpus query tool of sorts, simple in some ways and powerful in others. The Ngram viewer is not a tool for finding a book in the library or a word in a text. Instead is a tool for revealing differences and the patterns that result from them. Consider the results of looking for 'liberty', 'freedom', 'slavery', 'democracy' in a minimally curated collection of one million English books:



For readers with any knowledge of Western history, the spikes in the 17th and 18th centuries offer a striking confirmation of what they already know. For readers without such knowledge the graph immediately poses questions: What is going on with 'liberty' in the 1600s and late 1700's or with 'slavery' in the mid-nineteenth century?

1.9 Replacing searches "from string to properties" with searches "from any property to any other property"

Despite its power the Ngram Viewer remains a simple search tool in the sense that it starts with the look-up of one or more words or phrases. There are, however, many questions about a body of texts that do not start from particular search terms. Consider a historical linguist who wants to trace the gradual replacement of question markers from the inversion of subject and object (Knew you not Pompey?) to the use of auxiliary 'do' (Did you not know Pompey?) You could look up all the occurrences of 'do' and classify them. But this would not help you much in tracing how and where the older form survives. Instead you need a method that lets you identify particular verbs ('love', 'know', 'do') as instances of a general category ('verb') and look for all the cases where verbs are followed closely by a pronoun as subject ('he' or 'she', but not 'him' or 'her').

This is not a research question of consuming interest beyond a very special form of historical linguistics. But it does illustrate a central property of complex corpus query tools -- and a property whose utility reaches far beyond linguistics in a narrow sense. The User Manual of the venerable Corpus Query Processor (CQP) is very helpful in this regard. It tells you that

> the most basic version of a *text corpus* ... is just a sequence of *words* (or *tokens*). An individual occurrence of a word determines a *corpus position*.
> ...
> The *corpus positions* of a corpus can be annotated with an unlimited number of *attributes*. For CQP, the word form ... which is associated with a corpus position, is just one, albeit *distinguished* kind of *positional attribute*. This means that the query
>
> "Clinton"
>
> is just an abbreviation of its more formal equivalent:
>
> [word = "Clinton"];
>
> The square brackets [,] mark the beginning and the end of a query for a single corpus position. The = -symbol marks an attribute-value pair . It is a two-place operator which takes an attribute name (e.g. word) on its left side and an attribute value on its right side.

This prose comes from the "command line" world and is indicative of a mode of human/machine interaction that will not work for tools with wide acceptance in the humanities. It does, however, do an admirable job of articulating a fundamental difference between the reader and the machine. The reader sees a word in a position. For the machine the word is an 'attribute' of a corpus position. A 'distinguished' or first attribute, to be sure, but just one of many other attributes that could be associated with that position. The most common attributes in "corpus annotation" -- the technical term for such association -- involve mapping a spelling to a lemma or dictionary entry form of a word and to a part of speech. This can be done automatically with error rates of 4% or less. Once you have associated corpus positions with attributes you are no longer limited to the first attribute of the string as the point of departure for a search. Instead you can make any attribute the starting point of your search and make any other attributes the destination. Thus our historical linguist searches for the pattern "verb + subjective personal pronoun." His search retrieves lists of verbs, and their distribution across a corpus by genre or period provides the evidence for his inquiry.

Frequency is a very important property of words and phrases, and computers are much better at it than people. They can keep track of unimaginable numbers with extraordinary precision. In Google's Ngram Viewer, you put in one or more n-grams and the machine returns a graph of their changing frequencies over time. But instead of asking "How common is n-gram X?" you may want to ask "What n-grams are disproportionately common in X?", where X is some subset of a corpus, ranging from a single work (or a part of it) to the *oeuvre* of an author, a genre or the verbal output of an entire culture in a particular time or place. Alternately, you may want to know about n-grams that occur in one work and another but nowhere else. A query tool with the ability to answer this question is very useful for intertextual studies.

In sum, the most distinctive power of a complex corpus query tool rests on two factors. First, it lets you define corpus positions as 'tokens' of different 'type': 'know' as a spelling of 'know' or as an instance of a verb. Second, it lets you change the starting point of your search. You can start from a word, but you can also look for words that match some property. More broadly speaking, you can use any combination of properties to look for any other combination of properties. But just as the positional attribute of the word retains a place of special distinction, so the look-up search retains a privileged position. Indeed complex searches will often generate lists that guide subsequent look-ups. For the historical linguist, the search for the pattern "verb + subjective pronoun" yields a list of verbs with counts. Comparing those counts with overall counts yields a list of verbs where the older pattern persists or disappears particularly rapidly. At that point, the linguist will turn to looking up cases one at a time.

In a somewhat similar research scenario, a literary scholar might look at a large corpus of American novels published 1851-3 and 1873-5, the three-year ranges separated by eight years from the beginning and end of the Civil War. She might begin by extracting keywords that are strikingly present or absent in those sub-corpora when compared with each other or with a broader corpus of American fiction. This is distantly analogous to a bibliographical search, but instead of using keywords to locate articles to read, you use comparative frequencies of unknown words to identify words to look up.

2 Corpus query in a digital carrel

Corpus query tools in academic use have been available as open source software. Such programs as the CWB workbench, PhiloLogic, XAIRA, and Poliqarp, search engines like Lucene, or more complex search tools like ANNIS are yours for the asking and can be downloaded from Source Forge or similar repositories.[11] Installations of them are maintained for private or public use by individual scholars, institutes of one kind or another, and libraries.

If humanities faculty wanted to install corpus query tools, the technical or financial obstacles would not insuperable. A quite ordinary desktop computer would accommodate a PhiloLogic, Corpus Workbench or XAIRA installation of the 25,000 current TCP-EEBO texts or of other corpora running into the low billions of words. The programming

[11] Information about these tools can be found at
http://cwb.sourceforge.net/, http://sites.google.com/site/philologic3/, http://www.oucs.ox.ac.uk/rts/xaira/, http://poliqarp.sourceforge.net/, http://lucene.apache.org/java/docs/index.html,
http://edoc.hu-berlin.de/oa/conferences/reS4Xo05sncZc/PDF/29i8VTIzYfT3M.pdf

or systems administration skills required to do so are well within range of humanities 'hackers', and a faculty member who lacks those skills but has a research fund could always hire a computer science major to install or troubleshoot an installation.

This will happen from time to time. There will be more such installations, they will provide useful "bleeding edge" experiments, and "roll-your-own" installations will have a place in the ecosystem of digital scholarship. But for a variety of reasons that is not where the main action will or should be. The reasons have do with scale and the social space of scholarship. An article or book enters a social space in which its citations and references can be checked. There are projects that depend on the scholar's privileged access to data, and many readers of scholarly work do not in practice have access to the data on which the work is based. But these cases do not challenge the principle that a scholarly work and its underlying data should exist in the kind of public space that we are familiar with from the practices of academic libraries in a print culture.[12]

It is of course possible for scholars to work with very large data sets on their own computers in their own query environments, as long as these data sets are available in the public space of scholarship. But it is one thing for a corpus with billions of words to fit on the hard drive of an office computer and another for them to be managed successfully over the course of a scholarly project. Moreover, the comparative advantage of digital projects consists in the promises of scale and agility, which involve one or more of the following:

1. work with very large data sets
2. profile smaller data sets against larger data sets
3. combine heterogeneous small data sets into large sets

In a world of print based libraries, a carrel offers you space close to library books, guarantees you continued access to a subset of those books, putting them literally at hand, and may offer some space for keeping your own data. The carrel remains a major factor in promoting the agility of researchers by giving them a little space of their own in the open stacks of a large library where most needed data are immediately accessible and manipulable within the scope of readerly eyes, hands, and feet.[13]

The agility that is required for the successful manipulation of large and heterogeneous data sets is best guaranteed by putting large corpora into libraries or similar institutions and to implement corpus query tools as key functionalities of "digital carrels." This concept is not unlike the 'workbench' of the MONK project, the "work spaces" of Project Bamboo or the project spaces you find in Microsoft Office and other software tools. For humanists, the term "digital carrel" has the advantage of relating its affordances very concretely to work habits that are deeply familiar to them from their work with books. In

[12] This public space is a different thing from the public domain.

[13] This slightly idealized account of the ease of movement in a large open-stack library in a world before digital catalogues ignores the many bumps in the road scholars would encounter. On the other hand, digital interoperability is full of bumps too.

particular, it draws attention to the increasingly close connection between searching for books in a catalogue and searching within those books.

A digital carrel is an account that upon proper authentication gives you access to a virtual space with distinct affordances and privileges. A very simple version of a digital carrel is found in the Hathi Trust Digital Library. You can assemble named collections, keep them private or make them public, browse through all public collections, and search for words within these collections.

2.1 Affordances of a digital carrel at the catalogue level

The query potential of a digital document space depends on the interaction of the powers of the search tool with a corpus and its degree of curation. No query tool can extract data from a corpus that does not contain them in the first place, whether explicitly or by implication. Digitized corpora of printed books inevitably depend on bibliographical data about the source text. But bibliographical data describe the physical book, not its content. While they are essential for describing the provenance of the source they may be misleading indicators of its content.[14] In a corpus-based analysis of many texts you may want to have answers to questions like

1. When was the text composed?
2. Is the author male or female?
3. Where is the author from?

The answers would guide the definition of subsets for different forms of analysis. Library catalogs were not designed for those purposes and may require some retrofitting if they are to serve as good control tools for digitally assisted text analysis. Indeed, the concepts of 'text', 'work', and 'book' need rethinking in such a context. [15]

A software developer is likely to think of this problem as "out of scope" for the purpose of designing a query tool. From the scholar's perspective it is very much "in scope" when thinking about the affordances of a digital query space. Obstacles to the successful execution of a corpus query may not only arise from the fact that the search engine is too slow or too coarse. They are just as likely to arise from the fact that critical hooks for engaging the data are not kept in the appropriate metadata or are kept in a form that is not easily processed by a machine.[16]

[14] For instance, Marlowe's *Jew of Malta* is a work of the late 1580's but was first published in 1633. In the English Poetry database, Browning's poems, written over the space of half a century from 1833 on, are all referred to 1888, the date of his *Collected Poems*. On the other hand, if you are studying the reception of a work, successive publication dates are just as important as dates of composition or initial circulation.

[15] Such rethinking leads you to the acronym FRBR or Functional Requirements for Bibliographical Records (http://www.ifla.org/publications/functional-requirements-for-bibliographic-records)

[16] A good example of the latter is the way the name and birth date of an author are kept in a library catalog or TEI header.

2.2 Managing search results: the Achilles heel of corpus query tools

In the digital carrel as many operations as possible should be thought of as digital objects that can be saved, stored, reused, and modified. You want to move away from the "stateless" environment of the Web, where every call to the machine is an entirely new transaction with no memory of what preceded. In the Hathi carrel, you can save one or more sets of bibliographical items as collections to which you can return at a later. But it does not offer any query power beyond simple look-ups. The CQP website at Lancaster University, while much simpler and smaller in scale, offers a better model of a digital carrel as a work space for fairly complex forms of text analysis. The PhiloLogic environment supports quite powerful searches across very large collections, but it is "stateless" and does not give its users the little room of their own that is the fundamental feature of a carrel.

In addition to executing the "first order" search of retrieving a list of results, the search tool should also help users with the "second order search" of making sense of the result set by manipulating it in a variety of ways. By and large the management of search results is the weakest feature of search engines. The interfaces are built for intermittent use and do not support sustained work with textual data. All too often the machine returns a list that follows some default order or is "relevance ranked" according to algorithms that are opaque to the reader and may or may not be appropriate to the purposes of a particular search.

A short list can be 'eyeballed' easily enough. If the list is too long users are sometimes encouraged to 'refine' their search. Such advice assumes that the proper outcome of a search is a short list. If the list is too long, refine the search until it produces a list of the right length with the right answers at the top. But this is not always what you want. The proper outcome of a search may be a very long list that is then manipulated and displays its various properties depending on how it is grouped or sorted. Resorting a long list by hand is a very tedious business, but computers are very good at this task if the result sets are conceived of as machine actionable items. [17]

In library based search engines, the default search scenario ends with a list from which readers choose one or more for look-ups. In the Hathi library, the result list presents the reader with snippets of the text in the order of the occurrence of the search terms. A single click takes you to the page image. Looking at passages in books in this manner is certainly faster than using your hands and feet in a library. You can also do it in your pyjamas at 2 am. But considered as a mode of exploring a large result set it does not scale well.

Something similar true of Google's Ngram Viewer, which presents the reader with a very elegant and informative representation of the lay of the land from a bird's eye perspective. But the move from the overview to particular cases is coarse, clumsy, and does

[17] Business is way ahead of scholars in this regard. There is a large arsenal of tools that will help analyze the sales of a company by product, regions, salesman, special incentives, or any combination of these.

not scale. You can activate a link that will take you a list of books from a time slice of that search. The list does not appear to follow any order. You can then scroll through the page images of particular books. The search terms are highlighted in the digital facsimile based on the underlying OCR. This is very striking the first time, but you quickly tire of it.

In an ideal corpus query tool, the machine assumes that a search will return result sets of varying length, and that the initial result set may undergo various stages of post-processing before it yields results to the researcher. However it is presented, a result set is by definition a list of some sort. The rawest form of that list should be modeled in what Harald Baayen in *Analyzing Linguistic Data* calls a "long data format," a table that is savable as a distinct digital object and whose rows can be exported into a spreadsheet, become the 'observations' of a statistical program, serve as the input for a visualization program, or be variously displayed and manipulated within the post-processing capabilities of the corpus query tool itself. [18]

2.2.1 Differently scaled result sets

Google Earth provides useful guides for how to think about displaying result sets. You can zoom in and out of a particular map, and depending on your distance, various properties of a terrain move into focus. From one distance the most salient property of Chicago is its location at the Southern tip of Lake Michigan. From another distance, you notice the presence of very long and unbroken North-South streets, which you cannot see if you look just at the Loop. In a polemical opposition to "close reading" Franco Moretti has coined the term "distant reading" as a method for exploring digital corpora. "Scalable reading" may be a better to describe the distinct affordances of the digital medium. You can vary the distance from which you look at the phenomena.

The first important property of a result list is its length. A short list can be 'eyeballed'. For results ranging from two dozen into the low or mid thousands, you want to begin with some grouped representation of the data. You also want to let readers group, summarize, and sort the return by any combination of the 'factors' that support grouping, such as work, author, date by decade, quarter century etc, the preceding or following word., the sex or regional origin of the author, and the text category.[19] The post-processing of a data set of this middle size can probably be done inside the browser.

[18] Any 'hit' or 'observation' in this list is uniquely identified by its "corpus position" and is associated with unambiguously extractable data that are either stored as the "positional attributes" of the corpus position or are 'inherited' from the fact that any corpus position is part of a larger "corpus interval," which may be a section of a text as defined by its XML encoding or may represent the whole text, which is after all just an interval or range of positions in a multi-text corpus. If printed out or displayed on a screen, the redundance and verbosity of such a list will be repulsive to the human reader, but it is the basis for agile manipulation by the machine, and readers will have rarely have a need to look at the list in its rawest form.

[19] Within the small scale of the corpora for which it was designed (~10 million words), WordHoard has very sophisticated affordances for "on the fly" group and sort operations.

Grouping operations produce summary counts of various kinds. If the downloaded data set includes the word count of a document and the count of a word in that document, sort operations can display data by ascending or descending relative frequency, which is nearly always a useful way of sorting returns.

2.2.2 Snippets or keyword in context?

The Hathi trust environment returns ‘snippets' that are about the length of a biblical verse. PhiloLogic by default returns a slightly larger context, but also has a single-line concordance return (as does CWB), with the keyword centered and something like five words on either side. For linguistic or stylistic purposes, concordance output is often more efficient. For other inquiries, the snippet approach works better. The choice should be the user's.[20] There do not appear to be search engines that support the retrieval of results by such universally recognized categories as the sentence or the line of verse. But this has more to do with the encoding of textual data than with the powers of search engines.

2.2.3 Very large result sets

If result sets are very large, it is unlikely that researchers would turn to individual cases until after they have done a fair amount of summary analysis. It is also unlikely that large data sets could be comfortably manipulated inside a browser environment.

Consider the word ‘liberty', which occurs more than 100,000 times in more than 25,000 texts before 1700. Google's Ngram Viewer for its corpus of a million English books shows a spike in the 1680s, which is confirmed by the summarizing features of the PhiloLogic search engine. No scholar could or would want to look at all of the hits, and most scholars would not want to look at any until they had surveyed the distribution of the word by author, title, text genre, or date. In a case of this kind you would want the query engine to deliver a manipulable tabulation of summary data and request raw text data in a second stage on a more targeted or sampling basis.

2.3 Collocation

The British linguist J. B. Firth is best known for his statement “You shall know a word by the company it keeps.” Quite sophisticated techniques for establishing the company or “collocations” of a word have been developed since Firth's days. Because they are computationally expensive they tend not to deliver results in a split second, but they are informative. For instance, the expressive power of a frequency chart for liberty over time would be much enhanced if for every quarter century it displayed a tag cloud with the most common collocates. A modern reader unfamiliar with the 17th century might be surprised to see that ‘Christian' is by far the most common collocate of liberty in 17th century texts. Is that still the case for the late18th century spike of the word? In a similar chart for ‘gay' in the twentieth century, the changing collocates would immediately draw attention to the recent shift in the word's central use.

[20] The CQPWeb implementation at Lancaster is quite elegant in its flexibility. Its default setting is a single-line concordance output. If you “mouse over” the keyword, you see the raw output of the concordance hit with its linguistic annotation. If you click on the keyword, you are taken to a 200 word display of “extended context.”

Tracking differences in collocation by geographical location is another promising approach and lends itself to map -based visualizations that have become very popular and easy to do. For social and cultural historians, digitized newspapers will become an increasingly important source of very granular information. Their very ephemerality makes them valuable in the aggregate, and because ever page belongs to a particular time and place, newspapers are superb sources for the diachronic and spatial mapping of words. Corpus query tools may or may not help much with the analysis of sonnets, but they do offer new and exciting ways of 'reading' old newspapers.[21]

2.4 Pattern matching and regular expressions

Simple searches start from a string. Slightly less simple searches start from a partial string. Users can look for a word that begins with, ends with, or contains some sequence of characters. Such pattern matching is ubiquitous on the Internet. "Regular expressions" offer a powerful but tricky form of pattern matching. While full implementations have never found their way into word processors they are a staple of text editors and are typically found in corpus query tools of any complexity.

Users find regular expressions difficult because in the typical implementation it is hard to distinguish between two kinds of error. The first error is of a logical kind: the expression you construct is not sufficiently precise in what you want to include or exclude. The second kind of error has to do with the form of notation. The technology of regular expressions goes back to a time when the IBM keyboard was the Adamic language of the computer. The letters and numbers on that keyboard are literals and always stand for themselves. Some other characters are "metacharacters" and work as "operators" that do something. If you want to use such characters literally -- e.g. *()[]./?+\ -- you have to "escape" them by prefixing them with the backward slash, whose main function is to say "the next character is a literal."

The result is that regular expressions often look like jumbled toothpicks. They are difficult to read to begin with, and they practically unreadable if they do not fit into the typical small search box, as they often do not. Things brighten up considerably if you make the text field for regular expressions larger and use color coding to distinguish between metacharacters and literals. Bbedit does this for its Macintosh text editor. It reduces the time cost and error rate of regular expressions by an order of magnitude, largely because it makes it easier to identify the type of error.

This is a very primitive point about the ergonomic aspects of a user interface, but it carries the lesson that the usability of a tool as a whole may be greatly impeded or facilitated by the presence or absence of simple low-level obstacles. Few things will enhance the overall performance of a corpus query tool more than making it as easy as possible to teach novice users to craft regular expressions of moderate complexity.

[21] The underlying mathematics of collocation algorithms are quite complicated, but even if you do not understand the math you can learn how to interpret the results with confidence and some precision.

2.5 Handsome, clever, and rich

Consider the difference between the search expressions “bad king”, “adjective + noun”, “adjective + king”. In all these searches you look for two consecutive corpus positions. In the first and second, you look for “positional attributes” of the same kind, the word attribute in the first and the part-of-speech attribute in the second. In the third search you look for the part-of-speech attribute of the first position and the word attribute of the second position.

You could of course get the result of the third search by looking for ‘king’, ordering the return list by the preceding word, and selecting all the returns where the previous word is an adjective. But that takes a lot more time than letting the machine do that work. Because the classification of words as adjectives was done by a machine it will include false positives and false negatives. But a manually compiled list would also have errors, and in a very long list the machine might well do better than humans.

For many forms of stylistic inquiry, the ability to look for syntactic fragments is very helpful. Human readers will easily recognize “handsome, clever, and rich” as a type of phrase, but they will find it practically impossible to hunt for instances of it across texts of any length. A corpus query tool can in a matter of seconds retrieve all occurrences across several hundred 19th century novels.[22]

2.6 Frequency based points of departure

It is helpful to think of frequency-based searches, and indeed of all quantitatively based inquiries as “second-order” searches. They are based explicitly or implicitly on looking up words, often a lot of them, counting them, and performing some operations on those counts. The underlying math can get daunting very fast, but the fundamental steps are very simple, and keeping them in mind helps to demystify words like “text mining.”

Simple arithmetic or statistical operations are often a useful form of text profiling, especially if they proceed on a comparative basis. The bubble plots and tag clouds that in recent years have come to accompany reports of State of the Union addresses or other political speeches testify to the ease of extracting frequency-based patterns from texts and displaying them in striking visualizations. While the visualization often depend on raw counts, the crudeness of that approach is mitigated by the fact political speeches do not differ much in length. In a large collection with texts of greatly varying length, raw counts are useless or misleading. Relative frequencies and their ratios take you further and are very easy to compute over large corpora.

2.6.1 The G-test

Even more helpful is a well-established comparative statistic known as the G-test or Dunning’s log likelihood ratio. Given A and B as mixtures of similar parts, it identifies the differences in the distribution of parts that lie outside the expected range of random difference. With the help of the G-test you can identify words that Charles Dickens pre-

[22] A very casual scan of the results will show you that Charlotte Bronte is very fond of this type of phrase, which in her late novels becomes almost a writerly tic.

fers or avoids when compared to other novels written during his life time. The ability to measure what is less frequent is particularly helpful, because “the less” is not as easily seen by the reader’s eye as “the more.”

The G-test returns its results in a manner that are intelligible to anybody with a little exposure to formal statistics. It is sensitive to the overall count of words in a corpus. Thus it assigns lower probabilities to smaller differences in more common words than to larger differences in less common words. In making sense of it you need to recognize that texts are full of low probability events and that the standard thresholds of ‘significance’ do not work well with them. But once you recognize this, the G-test does a very reliable job of identifying lexical differences between texts and measuring the degree of that difference.

2.6.2 Intertextuality and sequence alignment

Much genomic research turns on identifying sequences in the four-letter alphabet of DNA, tracking those sequences, aligning them in various ways and deriving lines of descent from. A textual editor could think of this as the philology of life. Much scholarship in Literary Studies and the history of ideas turns on tracing the filiations of texts over time. Allusive practices, subconscious echoes, deliberate imitation, or plain theft provide the evidence for such filiations. Their study is a form of cultural “sequence alignment.”

If you suspect that a phrase in some text is echoed from some other text you can make it the target of a search, whether as a literal string or as a regular expression that allows for textual variance. If you suspect that two texts share long phrases, but do not know which, you are in the world of queries such as “What n-grams are shared between text A and text B but do not recur elsewhere or are rare in other texts?” Questions of this type have not been standard fare in corpus query tools, and there is not at the moment a tool that makes it easy for non-technical scholars to get answers to these questions without considerable technical support. The relevance, however, of such questions to many types of inquiry, especially in Literary Studies, is obvious -- witness such terms of arts as intertextuality, subtext, allusion, imitation, etc.

From a quantitative perspective, the specific domain of these questions is the rare rather than the common. Leaving aside a few common fixed phrases, anything longer than a hexagram is likely to be a proverb or a direct appropriation of a unique text fragment. In a static corpus it is possible to use a brute force approach, generate n-grams of increasing length, store repeated n-grams (whether literal or fuzzy matches), discard singletons, and stop at the longest repeated n-gram. This is computationally expensive, but it works and produces phrases of varying length that can be treated like any other lexical item.

If you conceive of a corpus as growing over time, perhaps through user-contributed additions, you cannot ignore long strings that occur only once. You need an algorithm that can look at an n-gram of any length in the next new text and determine whether that n-gram exists in the current corpus and through the addition of a new text becomes a repetition. If algorithms of this sort can be made to work “at scale” and deployed by scholars without expensive handholding by technical staff, the utility of this capability for any form of intertextual inquiry is obvious.[23]

2.6.3 Text mining approaches

Text mining is a term for a set of statistical algorithms that can extract patterns from very large data sets. The natural domain for text mining is the flood of born-digital documents about which you know nothing but must do something before the next day's tide, whether you work in a business or some intelligence agency. This is not a very common research scenario in humanities disciplines where inquiry clusters around a body of works about which much is already known. Philology, as August Boeckh defined it in the early nineteenth century, is about "die Erkenntnis des Erkannten" or the (further) knowing of the (already) known. Text mining may become more important in fields where digitization provides access to materials, e.g. newspapers, that scholars previously could not absorb *en masse* or "at scale." But its general utility for humanities scholarship remains an open question, with attitudes ranging from keen enthusiasm to fierce rejection or deep skepticism. It is partly a matter of how you describe the phenomenon. "Text analysis" may be a better term than "text mining" with its overtones of extractive rapacity.

With the various corpus query techniques discussed so far, it is easy to relate each technique to questions, problems and practices that are familiar to scholars from their engagement with primary texts in a world of print and manuscripts. This is harder to do with text mining routines that put scholars at uncomfortable distances from their data or mediate them through complex algorithmic filters that the scholar may not really understand and therefore is reluctant to trust.

Some enthusiasts argue that these routines can perform almost magical acts of text summarization along the lines of Christian Morgenstern's prophetic poem *Die Brille* or *The Spectacles*.

Korf liest gerne schnell und viel;
darum widert ihn das Spiel
all des zwölfmal unerbetnen
Ausgewalzten, Breitgetretnen.

Meistens ist in sechs bis acht
Wörtern völlig abgemacht,
und in ebensoviel Sätzen
läßt sich Bandwurmweisheit schwätzen.

Es erfindet drum sein Geist
etwas, was ihn dem entreißt:
Brillen, deren Energieen

Korf reads avidly and fast.
Therefore he detests the vast
bombast of the repetitious,
twelvefold needless, injudicious.

Most affairs are settled straight
just in seven words or eight;
in as many tapeworm phrases
one can prattle on like blazes.

Hence he lets his mind invent
a corrective instrument:
Spectacles whose focal strength

[23] One might observe in a melancholy mood that universities spend lots of money on this technology when it comes to detecting plagiarism but are much are less apt to support its deployment by scholars who want to trace the filiations of textual traditions.

ihm den Text - zusammenziehen!	shortens texts of any length.
Beispielsweise dies Gedicht	Thus, a poem such as this,
läse, so bebrillt, man - nicht!	so beglassed one would just -- miss.
Dreiunddreißig seinesgleichen	Thirty-three of them will spark
gäben erst - Ein - - Fragezeichen!!	nothing but a question mark.

But many humanists will counter that such literally reductive approaches are precisely not what they are looking for, at least not in their engagement with primary texts. Humanists are also much more likely to take to text analysis procedures that are informal, interactive, and iterative in the spirit of John Tukey's "exploratory data analysis."[24]

That said, some inquiries in the humanities will find a use for statistically inflected forms of text analysis, whether at the macro level of charting unknown territory or at the micro-level of exploring known texts in more granular ways. Forms of "unsupervised classification" offer particular promising techniques of exploratory data analysis. An algorithm works its way through hundreds or thousands of texts, keeps track of various unspecified phenomena along the way, uses these phenomena as the basis for successive divisions, and produces a 'dendrogram' or genealogical chart of texts that groups in terms of the number of branchings by which they are divided. In an analysis of Early Modern drama it would be gratifying to see all of Shakespeare's plays neatly sitting on the same branch, with the tragedies, histories, and comedies on twigs of their own. But what if the classifier grouped all those plays together because in the underlying text you always find 'to-morrow' or 'to-day', while in the non-Shakespearean plays you find 'to morrow' and 'to day'?

The successful integration of complex text mining routines into humanities scholarship is likely to involve a good deal more than corpus query software with approachable user interfaces. It will require research teams with complementary scholarly, statistical, and technical competencies. It would not be an easy task to fit the development of such teams into the budgetary framework and institutional habits of such humanities departments as English, History, or Philosophy.

2.7 XML and TEI: the model of the muddle in the middle

A corpus query typically looks for "content objects" in some "ordered hierarchy" and retrieves branches or "leaf nodes" of a tree. A search in a library catalog leads to the leaf node of a title. In a full-text search that leaf node becomes a branch, and the words become the leaf nodes, whether immediately or through the intermediate branches of pages.

[24] In his book *Exploratory Data Analysis* (1977) the statistician John Tukey opposed 'exploratory' and 'confirmatory' data analysis, emphasizing the virtues of the former. To a humanist's ears there is something very attractive about a statistician who tells you on the first page that "one needs quite different sorts of detailed understanding to detect criminals in London's slums, ... among Parisian aristocrats, or in the Australian outback." Tukey invented the now ubiquitous box plot, a tool that goes well with his observation: "Far better an approximate answer to the right question, than the exact answer to the wrong question" (cited from http://en.wikipedia.org/wiki/John_Tukey).

The search for a word in a digital text derived from optical character recognition (OCR) goes through a document "path" of the type title/page/word.

Full-text transcriptions of primary documents in the humanities, however, are encoded in a more complex manner. Virtually all digital editions of primary texts with any claim to scholarly standards use the encoding protocols of the Text Encoding Initiative (TEI), which are implemented in XML. TEI is the lingua franca of digital scholarly editing on a global basis. You find it in editions of Buddhist sutras, New Zealand and Pacific island texts, Greek inscriptions, French manuscripts of the *Roman de la rose*, the Hengwrt manuscript of the *Canterbury Tales*, slave narratives of the American South, or the official records of the State Department. TEI has been used in all the large-scale university- library-based digitization projects of primary texts at Indiana, Michigan, North Carolina, Virginia, and the Library of Congress. The same is true of European encoding projects.

In XML based encoding the content of a document is, if you will, "containerized" in a nested structure that reflects the logical structure of the document rather than its physical layout as a sequence of words, lines, and pages. The containers, called "elements" and marked by the opening and closing tag names in the "pointy brackets" familiar from HTML, range from the general and large-scale (front, body, back) to the small and precisely defined: you can wrap notes, stage directions, speaker labels, or lines of verse in appropriate elements. The purpose of such "containerization" is to enable users to extract data from particular containers or elements. An "XML-aware" search tool lets you restrict searches to the words in a particular XML element or all the 'children' of such elements.

In theory, in an XML-aware search of TEI collections, the "ordered hierarchy of content objects" extends all the way from the structure of the catalogue through the structure of the documents to the leaf nodes of the words, promising seamless and fine-grained searching across collections. In practice, things do not quite work that way. Similar structural articulations in the source texts may be encoded quite differently in two TEI documents. This applies not only to documents from different encoding projects, but even to different documents from the same project. Thus the reality of searching TEI documents falls considerably short of the dream of readily cross-searchable collections.

Despite these shortfalls, the potential of such searches is far from trivial. You can exclude or focus on front matter across a large number of titles. Barring gross encoding errors, poetry in a TEI document is always encoded in an <l> element and only in that element. Where the printed text follows strict conventions, as is the case with drama, TEI encoding is likely to support quite granular searching by elements across collections.

Three years ago the TEI released a new version of its standard called P5, which is no longer "backwardly compatible" but aligns it more closely with standards shared across the XML world. At some point, the bulk of existing collections will be converted to P5. The potential for searching across collections will be improved considerably if projects see that conversion as an opportunity to make their collections interoperable at a "highest common factor." In such an environment, collections share what the German TextGrid project calls a “baseline encoding” (*Kernkodierung*). Where differentiations for special purposes rests on a shared basis, it is possible to keep collections interoperable to the highest point of their shared encoding. Where differentiation is built into the bottom, interoperability is bound to be more limited. This is another illustration of the fact that a digital query space is as much shaped by the structure of the data as by the capabilities of the search tools.

Query environment have not yet taken full advantage of the affordances of TEI-based encoding. Some search engines have been built for element based searching from the ground up. This is true of SARA, the search engine for the British National Corpus, and its XML successor, XAIRA. General purpose search engines like Lucene, can incorporate some aspects of a document’s element hierarchy into their indexes. The Corpus Workbench has partial support for SGML, the predecessor of XML, but no support yet for XML. In PhiloLogic, it is possible to restrict searches to elements, but the interface for those functionalities is not very user-friendly, and few users are aware of it.

Xquery is a recent programming language designed for the special purpose of manipulating XML documents. In the XML world it plays the role that SQL plays in relational database environments. Xquery looks for elements rather than words, which in most texts are not wrapped in elements. As a search process this is at the other end from CQP, which ‘knows’ about strings of characters regardless of elements. However, several implementations of Xquery, notably the eXist database, have added extensions to handle text content within XML structure, and the W3C -- the consortium that oversees Web standards -- has recently published a set of recommended extension to the Xquery specification to provide the kind of features characteristic of text-searching software such as CQP or PhiloLogic.

It is not yet clear whether or how Xquery can scale to very large and deeply structured text archives. But Xquery is unlikely to differ very much from other digital technologies, where yesterday’s ‘large’ is tomorrow’s ‘small’.

Element based searches faces a fundamental problem for which there is no obvious solution: if you don’t know how a document structure is mapped to a hierarchy of elements you cannot make good use of the elements. You must have at least some knowledge of TEI to make good use of the affordances offered by TEI-encoded documents. But despite the universal acceptance of TEI as the appropriate encoding standard for primary texts in the humanities, a knowledge of the standard and appreciation of its affordances has not yet spread to the wider community of scholarly users. But this may be just a particular instance of the general fact that the affordances of digital textuality have not yet found their way into the calculus of the possible that governs scholarly work habits in the humanities. Most biologists appreciate the power of sequence alignment tools for their work. Few literary scholars do, although the underlying analytical techniques help with solving structurally very similar problems in both domains.

XML is still a quite young technology, and we are in an early phase of moving from the difficult task of developing the functionalities of a technology to the even harder task of building gently graded access ramps that will allow ordinary scholars to take advantage of the technology.

2.8 Querying syntactically annotated texts

The search routines described so far all rest on a very simple model of human language as a sequence of "corpus positions" that can be searched in terms of lexical, grammatical, or other "attributes." Queries often adopt an even simpler model. The sequence of tokens is ignored, and the tokens of a particular corpus or its sub-corpora are considered as "bags of words." This bag-of-words model of a text ought to be deeply repulsive to any writer who has struggled with putting words in the right order. It is, however, quite effective and for many analytical purposes works better than it should.

A corpus query tool that can process a text as sequences of parts of speech has rudimentary syntactic capabilities. Syntactic n-grams offer a very fragmented view of higher-order syntactic structures, but one can often draw reliable inferences about higher-order syntactic differences between two texts by observing differences in the distribution of syntactic trigrams. On the other hand, full syntactic analysis requires much more carefully annotated texts in which the dependencies of phrases and clauses are explicitly marked.

ANNIS, an acronym for ANNotation of Information Structure, provides a search environment in which the information extractable from such deeply annotated texts is mediated through a relational database environment. A system of this kind is an important reminder of the interaction of scale, complexity, and curation. If you want to get more out of your data, you have to put more into them.[25] Much time and effort goes into the curation, and the complexity of the resultant data set in turn imposes constraints on the size of the data set that can be searched with an appropriate time cost.

The calculus of scale changes rapidly with advances in technology. The Brown corpus of 1967, the first and very lightly annotated corpus, consisted of a million words (the equivalent of half a dozen novels). Today a very deeply annotated corpus of that scale can be queried by ANNIS. A few years from now you will be able to do it with a ten-million word corpus, at which point the cost of data curation is likely to be the limiting factor.

3 Shared instrumentation in the digital carrel

The National Center for Research Resources (NCRR), an arm of the NIH, recently renewed its Shared Instrumentation Grant (SIG) program, whose objective is "to make available to institutions expensive research instruments that can only be justified on a shared-use basis." Initiatives of this type are more common now than they were a generation ago. Fancy tools do more but also cost more. Financial pressures make it necessary to share. At the same time technology makes it easier to share. In well-designed digital

[25] ANNIS is not limited to syntactic annotation but also supports the annotation and analysis of higher forms of discursive or rhetorical structure.

environments it happens less often that "I" cannot use the shared tool because "you" are hogging it. In fact, sharing may lead to better data and better tools.

A corpus query environment of the scale and complexity described in this report goes well beyond what is currently available in any single installation. Its creation is clearly a task that requires financial, scholarly, and technical collaboration. Think of it as a form of "shared instrumentation" readily accessible to scholars from their digital carrels.

"Instrumentation" is a complex word in this context. It is natural to think of a query tool as piece of software that does something with some text. But consider the difference between a query tool that performs linguistic analysis on the fly and a tool that operates on a text that has previously been annotated with some other tool. In the second case, the query tool operates on a pre-processed text. Something similar happens with "named entity extraction," a very common procedure where names of people, places, institutions, etc. are identified and labeled by a machine so that they can be subsequently retrieved with greater speed and accuracy.

In fact, a user-generated query is better seen as the last stage in a multi-phase process in the text has been enriched with the output of various tools. 'Text' and 'tool' do not stand in simple opposition. But neither do 'book' and 'reader'. Yeats' question about telling "the dancer from the dance" is relevant to both of them.

3.1 The corpus as corpus query tool

Emily Dickinson's line "We see -- comparatively" is a good way of expressing what is at stake in any move beyond "simple searching." If your goal is to find needles in haystacks, more haystacks are always better: the missing needle may after all be in the next haystack. But while extraction of some defined thing from some undifferentiated mass is an important property of a finding aid, it is equally or even more important to define a found thing by seeing it in the contexts from which it derives its shape.

For such contextualizations, there comes a point where more highly curated data add more value than more raw data do. Return for a moment to Google's Ngram Viewer. When it comes to the distribution of words that occur with any frequency, there is little difference between graphs based on five million or one million books. Now imagine a corpus of 100, 000 books (~10 billion words) with these properties:

1. The texts are classified by broad categories of genre or subject.
2. Within each text, regions of prose, verse, and direct speech are distinguished.

The overall frequency charts for this hypothetical ten billion word corpus would differ trivially from the larger corpora. But the sub charts with their different proportions and differing trend lines would tell much more complex stories. On the other hand, it may be more labour to create this search granularity in a corpus of 100,000 books than to add another million to the five million Google books. [26]

Linguists are not always very interested in the boundaries of texts or authors. Their relevant speaker is Language itself or a particular register of it. For this reason, linguistic inquiry does not suffer much if texts are accessible only as snippets. While this may also be true for some inquiries in the humanities, most inquiries benefit from or require full-text access.[27] But within current constraints of scale it is entirely possible to envisage a corpus query environment that combines search tools with a "reference corpus" that serves a double purpose. On the one hand, it provides various base line measures that help in the exploration of arbitrary text collections that scholars assemble for this or that purpose.[28] On the other hand, it is a sufficiently large collection of texts to serve as the major source of primary documents for those disciplines that revolve around the study of more or less canonical texts. For scholars in those disciplines, the reference collection provides not only a convenient supply of 'citable' texts but also enhances their potential for new forms of micro- or macro- analysis.[29] Such a corpus would need to be constructed in a very ecumenical spirit and should allow for user-driven curation, modification, or expansion over time. [30]

This example shows once more that data curation and data exploration are the Siamese twins of text analysis in a digital world. The most powerful tools in the world cannot get out of texts what somebody did not put into them, and there are limits to what one get out of "plain text" versions. On the other hand, there is little point in putting things into texts unless there are decent tools for getting them out at an affordable cost of time or effort.

3.2 The digital carrel as an instrumentarium built like a pyramid

The digital carrel is a space where scholars work with different tools and corpora. There neither will nor should be a Corpus of Everything, nor is it promising to think of a single and monolithic query tool that will do everything for everybody. To borrow a med-

[26] Mark Davies made a similar point in his critique of the Ngram Viewer. His Corpus of Historical American English can be thought of as a corpus query tool that can be used in any analysis of texts since 1800 where it helps to know something about the frequency or distribution of particular words. Think of it as a yardstick that lets you compare the occurrence of words in your texts with the lexical data for the appropriate genres or periods.

[27] It may well be the case that most texts are read in snippets. But the snippets and their length are chosen by the reader .

[28] For example, an obscure set of mid-18th century sermons is an important source for your analysis of some religious or political conflict. You profile those sermons against summary data from the several thousand sermons in the reference collections and form an idea of where they sit on a broader map.

[29] Citability is the scholarly version of social respectability. A citable digital text needs to have a respectable provenance, be explicit about it, be a faithful encoding of its source, and provide a clear way of referring to its parts -- just like the Stephanus editions of Plato and the Bible.

[30] The best theoretical account of such a corpus is found in a 2004 paper about Deutsch.Diachron.Digital (DDD) in "Challenges in Modelling a Richly Annotated Diachronic Corpus of German" by Stefanie Dipper, Lukas Faulstich, Ulf Leser, and Anke Lüdeling (http://www.deutschdiachrondigital.de/publikationen/index.php). From it one can learn useful lessons about something like EEE or English Epochs Eletronically.

ical term of art, we need an 'instrumentarium' rather than an instrument. *Vive la différence* is useful advice when thinking about corpora and corpus query tools. But there are differences that get in the way of readily using one tool for different corpora or different tools with the same corpus. It helps to plug a coffee machine, a hair dryer, or a radio into the same outlet. It is less helpful to have a separate charger with its own and differently shaped plugs for each of your seven electronic gadgets.

Such troublesome differentiation can be avoided by a pyramidal model where complex operations sit on shared simple functions, differentiation occurs at the highest possible level, and complexity can both be added to a lower level or stripped from a higher level. The *Kernkodierung* or "baseline encoding" of the TextGrid Project, mentioned earlier in this report, is one example of this approach.

In any search environment there will be trade-offs between size, speed, and complexity: the kinds of complex searches possible with something like ANNIS are limited to corpora whose size is two orders of magnitude below the corpora than can be comfortably managed by CQP, PhiloLogic, or XAIRA. The absolute limits for each of these search tools will certainly go up. On the other hand, the ratios may stay the same. Nor will there be a change in a spectrum of search scenarios where the size of the corpus, the number of users, the complexity of queries, and the speed of performance move across spectra of larger|smaller, lower|higher, and faster|slower. There are no Goldilocks points in those spectra, and a single project by one scholar may involve shuttling between different end points. There also will always be a tension between the needs for differentiation and integration or interoperability. A pyramidal model will help manage those tensions. So will the digital equivalents of adaptor plugs or voltage converters that help you shave or dry your hair in different countries -- programs like SaltnPepper that take data from one project, e.g. a Perseus Treebank, convert them to a 'metamodel' and then convert them into a format that allows their 'ingestion' by a third program, e.g. ANNIS.[31]

3.2.1 An ISO standard for a corpus query lingua franca

There was some enthusiasm and some skepticism about the idea of putting together an ISO proposal for a corpus query *lingua franca* that would allow different query engines to interoperate by defining such concepts as 'text', 'fragment', 'property' and making recommendations for representing queries and result sets. On balance, it seems worth doing. If it leads to a standard it will make development easier. If it does not lead to a standard it may at least help articulate the points where interoperability becomes difficult or breaks down.

3.2.2 The pyramid of the digital carrel

At the base level of this pyramid there are functionalities of the kind that are well implemented in the Hathi Digital Library:

[31] https://korpling.german.hu-berlin.de/trac/saltnpepper. The odds are that programs of this type will always require professional support. Scholars in their digital carrels will benefit from but are unlikely to use them directly.

1. an account that authenticates you and provides access to all the digital materials that you are permitted to use
2. assemble one or more sub-collections that you can share with others
3. perform simple full-text searches over the entire collection or any set of sub collections

At the second level of the pyramid there is a sizable collection of well-curated primary documents that have been of recurring interest to scholars in various humanities disciplines.[32] This collection is a structured corpus of texts with enhanced bibliographical data and light, but consistent, linguistic and structural annotation. The enhancement of bibliographical data involves the reshaping and supplementing of existing catalog data to support text analysis at a macro- and micro-level. This corpus should not be thought of, in Thucydides' words as "a possession forever," but as a work in progress and to be shaped over time by user contributors.

The digital carrel also includes space for users to store their own data and for sharing them with others, subject to standards of interoperability.

At the level of tools, in an ideal implementation, scholars can operate the tools not only on the standard corpus or any part of it but also on their own data or combinations of their own data with the standard data. In practice, indexing schemes are likely to impose serious constraints. Searches do not operate on data but on indexed versions of these data. A search that returns results in a split second may depend on an indexing operation that took days to complete. There are not yet any systems that operate with a version of incremental indexing that would make it a trivial operation to add tests to or subtract them from a corpus.

The query tools available from the digital carrel should support the functionalities discussed in various sections of this report:

1. Limit a search by bibliographical criteria
2. Refine bibliographical criteria to articulate differences of text category and chronological or geographical properties associated with texts
3. Limit a search to particular XML elements of the searched text(s)
4. Simple and regular searches for the string values of words or phrases
5. Retrieval of sentences
6. Searches for part-of-speech (POS) tags or other positional attributes of a word location
7. Searches that combine string values with POS tags or other positional attributes
8. Define a search in terms of the frequency properties of the search term(s)
9. Look for collocates of a word

[32] This proposal assumes that scholarship in the humanities, like many other walks of life, is governed by some version of the Pareto distribution or 80/20 rule. If you look at all the primary documents used by all the scholars you will find that a relatively small percentage of documents will account for a very high percentage of use.

10. Identify unknown phrases shared by two or more works (sequence alignment)
11. Compare frequencies of words (or other positional attributes) in two arbitrary sub-corpora by means of a log likelihood test
12. Perform supervised and unsupervised forms of text classification on arbitrary subsets of a corpus
13. Support flexible forms of text display ranging from single-line concordance output to to sentences, lines of verse, paragraph length context and full text
14. Support grouping and sorting of search results as well as their export to other software programs

4 About this report

This report is based on two days of conversation among a group of humanities faculty, librarians, and information technologists on November 22-23, 2010 in Evanston, Illinois. The participants are grateful to the Andrew W. Mellon Foundation for its generous support of the event.

The participants in the conversation were:

- Bridget Almas, Senior Software Developer, Tufts University
- Lou Burnard, Assistant Director Oxford University Computing Services (retired)
- Philip R. Burns, Senior Software Developer, Academic and Research Technologies, Northwestern University
- Helma Dik, Associate Professor of Classics, University of Chicago
- Andrew Hardie, Lecturer, Corpus Linguistics, Lancaster University
- Catherine Mardikes, Senior Humanities Bibliographer, Classics, Ancient Near East and General Humanities, University of Chicago
- Martin Mueller, Professor of English and Classics, Northwestern University
- Mark Olsen, Assistant Director, ARTFL, University of Chicago
- John L. Norstad, Software Developer Lead, Academic and Research Technologies, Northwestern University
- Laurent Romary, Senior Researcher, INRIA (Institut National de Recherche en Informatique et Automatique)
- Robert Morrissey, Benjamin Franklin Professor, Department of Romance Languages and Literatures, University of Chicago
- Richard Whaling, database analyst, ARTFL, University of Chicago
- Amir Zeldes, Researcher, Collaborative Research Center 632 (Information Structure), Humboldt University, Berlin

The report was drafted by Martin Mueller, circulated among all the participants, and revised in the light of their suggestions.