Building a Hypertextual Digital Library in the Humanities: A Case Study on London

Gregory Crane, Clifford E. Wulfman, David A. Smith
Perseus Project
Eaton Hall
Tufts University
Medford, MA 02155
E-mail: {gcrane, cwulfman, dasmith}@perseus.tufts.edu

ABSTRACT

This paper describes the creation of a new humanities digital library collection: 11,000,000 words and 10,000 images representing books, images and maps on pre-twentieth century London and its environs. The London collection contained far more dense and precise information than the materials from the Greco-Roman world on which we had previously concentrated. The London collection thus allowed us to explore new problems of data structure, manipulation, and visualization. This paper contrasts our model for how humanities digital libraries are best used with the assumptions that underlie many academic digital libraries on the one hand and more literary hypertexts on the other. Since encoding guidelines such as those from the TEI provide collection designers with far more options than any one project can realize, this paper describes what structures we used to organize the collection and why. We particularly emphasize the importance of mining historical "authority lists" (encyclopedias, gazetteers, etc.) and then generating automatic "span-to-span" links within the collection.

KEYWORDS: automatic linking, collection development, document design, reading, browsing.

INTRODUCTION

Two years ago, we set out to create a new, densely hyperlinked digital library of materials pertaining to pre-twentieth century London and its environs [1]. The first results of this work are now available in the Perseus Digital Library (http://www.perseus.tufts.edu). This paper describes some of the results from our initial work on this collection.

Before the London work, we had spent more than a decade developing a collection of Greco-Roman cultural materials [2]. Although we learned a great deal about the tasks building such a resource entailed and the benefits such a resource could provide, we knew that these were in some ways unique to classical studies and did not necessarily pertain to humanities digital libraries as a whole. Consequently, we began to explore other domains of humanistic interest, like early modern English and the history of science [3], [4], in order to more concretely distinguish general from domain-specific issues. A collection housed at Tufts University intrigued us particularly. In 1922, the university had acquired a major set of books, maps, and pictures of London and its environs [5]. The collection is an important one, because its materials shed light on London when it was arguably the greatest city in the world; unfortunately, sequestered in the archives, it was accessible only to specialists who made their way to Tufts' special collections.

One difference from our Greco-Roman collection particularly intrigued us. The classical record is sparse: scholars spend a great deal of time determining who people were, where places were located, and what things may have looked like. By contrast, our data for the past few centuries of European and North American history are vast, and their organization and presentation raise different challenges and opportunities from those presented by the remains of the ancient world. We wanted to see how effectively we could use data available in printed form to create a digital collection that would have properties that built on, but were distinct from, its print sources. In particular, we wanted to see how time and space could be used as axes along which to organize the materials.

We also wished to discover how some of the technologies we had developed for classical study could be adapted to more modern texts. For Greek and Latin, we had surmounted a major technical hurdle that has bedeviled novices and experts since non-native speakers began to study these languages [6]. The morphology of Greek and Latin is far more complex than that of Western European languages such as English, French, Spanish, German or Italian: a single Greek verb can in theory appear in millions of different forms. We developed a system that could map inflected forms to their dictionary entries and were thus able to create links from inflected words to dictionary entries, a feature which has proven enormously popular among students of the languages. The same system allowed us to create much better retrieval tools to aid those conducting philological research. We wanted to see whether similar dense links might be useful in a collection in English, most of whose users were not desperate for all the linguistic help they could get.

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

This paper is aimed at two audiences. First, we hope to present one strategy of collection building for those who are themselves contemplating similar projects. We expect that many who use the London collection will be working with literary texts; nevertheless, rather than starting with major literary works (many of which were in any event already on-line), we chose to emphasize histories and descriptions of London and its environs, sources that might not occupy such a prominent position in the curriculum or public eye but that would add value to canonical literary texts. Since these reference works tend to be larger, more complex in format, and thus more expensive to digitize than literary works, such an approach was not easy, but our experiences with Greco-Roman Perseus, and now with the London collection, suggest that reference works are, in fact, a logical starting point for collection building.

We also hope to present an audience of information technologists and interface designers with a reasonably well structured test bed that is distinct in form and content from those built for the fields of science, technology, and medicine. People have been making books and reading them for thousands of years: long before the digital age, technology was shaping the organization and display of information (see, for example, a recent essay entitled "The Early Modern Search Engine: Indices, Title Pages, Marginalia and Contents," [7] part of a book called *The Renaissance Computer: Knowledge as Technology in the First Age of Print* [8]). The strategies we pursue today as we develop digital libraries build on traditions of information organization that have evolved since antiquity. The London collection provides a new historical text, primarily in English, with which to test various strategies for organizing collections. We have been particularly interested in seeing how effectively the organizational elements in these pre-twentieth century books support visualization strategies and the automatic generation of links.

Building a digital library of materials on any major city — especially on one so vast and important as London — is an open-ended task that can easily absorb decades of labor and millions of dollars. The results that we offer here constitute baseline observations after two years of work. The Greco-Roman collection in the Perseus Digital Library contains c. 35,000 images and 22 million words, of which 5 million are in classical Greek, 2.6 million are Latin and the rest primarily English. It has been evolving on the Web since 1995 and now has a substantial user base: in 2000, we served 67,000,000 pages to 6,800,000 sessions. The London collection has c. 10,000 images and 11 million words — substantially smaller but large enough to begin exhibiting problems and advantages of scale. We have only just begun to make the initial London materials available.

SUPPORTING SCHOLARLY READING

Collection design (whether the collection is digital or print) often presupposes a model for collection use. Our model requires some explanation, because it differs from those assumed by other digital library resources.

Most libraries of journal articles and monographs assume a rather utilitarian model of reading, in which the best document is the one that yields the most useful information in the shortest time with the least effort. In this model, reading is driven by explicit goals: the need to prepare a briefing on security concerns in a Latin American country, or to develop a new procedure for treating a form of hepatitis, or to find the most appropriate methodology for clustering related documents. The documents themselves are means to an end, to be absorbed and discarded. The digital support of journal-reading practice has been the object of study in its own right [9], while some worry (with good reason) about the superficiality of such "hyperextensive" reading [10].

Literary reading (insofar as there is any single practice by that name) defies (and to some extent is defined in opposition to) such utilitarian models of reading; it abjures the extraction of discrete, well-defined messages from closed works for open texts with meanings that are problematic at best. It is the most theorized and hotly contested of reading practices, and its digital formations have been written about extensively (see, for example, Landow [32], Joyce [33], Murray [34], Aarseth [11] and Douglas [12])

Those who "historicize" documents — struggling to experience them as parts of past cultures —often occupy a position that partakes of each extreme, occupying less a stable mid-point than the third point of a triangle, midway between the other two extremes but as far from each as they are from each other. On the one hand, they must immerse themselves in information: countless factoids are the raw material for larger narratives and often allow us to breathe life into the past (this is underlying argument of [35], for example). They must be, like any good researcher in any field, cold and passionate at once, able to react with delight to the dry where possible and to drag themselves through the frankly dull where necessary, ploughing through large stretches of material and retaining as much as they can.

On the other hand, many of the objects historicists study cannot be reduced to containers of information, but are objects whose meaning and interest deepen with study. Thucydides' *History of the Peloponnesian War* is not simply a historical source but an object of intensive analysis and indeed pleasure in its own right. The rise of cultural studies has brought some of the practices of literary reading to documents and objects far outside the canons of high culture. In this view, every historical text requires intense and thoughtful study if it is to be properly evaluated: we cannot even accept "objective" data sets (such as census records) unless we understand the process whereby the structuring categories are designed and the data collected.

A different view of documents and libraries emerges from this third, "historicist" perspective. On the one hand, the documents that we produce are not disposable tubes of information that can be squeezed dry and cast aside. We can expect readers to go through documents from beginning to end (often more than once), and from end to beginning; to jump from one point to another; to race through some passages and linger over others. At the same time, we can expect them to search for those materials that can give context to the words or images before them — to find "information" that will cast the object of immediate interest in a different light. Such information can range from preliminary background information (e.g., who is a particular person? where is a given place located? what is the traditional custom being mentioned?) to more complex issues touching the culture as a whole (e.g., the relationship of mass and elite, or ways of describing space or broad ritual practices). To develop their own mental models, readers need information and lots of it.

The above outline has a number of implications for digital library system design:

- Digital libraries are not designed to generate short-term remuneration, be it massive traffic (indicating that the content is heavily used) or financial gain. We want to help individuals systematically expand not only their knowledge of a particular subject but also their ability to approach problems in general.
- Collection builders want to maximize their audiences, but interpreting cultural artifacts is inherently complex, and we defeat our own purposes if we artificially simplify our materials. Good design is crucial for the broad acceptance and sophisticated use of digital collections, and attractive and engaging presentation is as important for libraries as it is for commercial sites.
- Size matters. Digital libraries need substantial bodies of material if they are to become useful, and these bodies of material may need to contain heterogeneous categories of data (e.g., animations, statistical datasets, and geospatial data as well as texts).
- If the documents in our digital library are not simply containers for information but objects of study in their own right, we need to be able to work with them at a fine level of granularity. Document to document links are not enough: we need "span to span" links connecting arbitrary subsections of documents.
- Above all we need as many links as possible between the objects of study and related materials. While most web designers aim for a small number of highly pertinent links, we need more, rather than fewer, links. We want to support free browsing and are willing to tolerate a limited number of false leads, since false leads are inherent in all serious inquiry. In the language of information retrieval, where conventional web design stresses precision (a small number of focused links), we seek to emphasize recall (getting as many links as possible).

  Human-generated links are useful and critical editions traditionally provide rich connections to supporting materials. The *New Variorum Shakespeare* series (NVS), for example, produces editions of individual plays that collate every significant edition ever published, provide line by line commentary summarizing the important findings in scholarship, and include major source materials and essays on stage history, character studies, actors' interpretations, criticism, and other topics. A single such edition can require ten years of labor. One recent edition contained more than 10,000 bibliographic citations and 5,000 links to parts of the play, each of which was the product of substantial thought. However, the NVS cannot keep pace with ongoing Shakespeare scholarship, and its print volumes begin drifting out of date the minute the author hands the manuscript over to the copy-editor. The Shakespearean canon comprises 36 or 40 plays (depending on who decides the marginal cases). Even if the NVS could produce a new edition each year, the series would, on the average, reflect the state of Shakespearean scholarship eighteen years in the past. Thus, hand-generated links do not fully satisfy our needs, because they are too labor intensive to keep pace with rapid changes in scholarship. Furthermore, even if we had the scholarly labor to produce the equivalent of a variorum for every important document, we would still not be entirely satisfied because we would inevitably have questions that went beyond the editor's interests.

  Thus if we are to support scholarly reading, we need to connect each document to a hypertextual digital library — a digital library that is not only large enough to support serendipitous discovery but is broken up into logical chunks that can (when appropriate) be rapidly digested. Where many Web designers strive for a few well-chosen links, our goal is to provide information about as many words and phrases as possible. We strive to create an environment that encourages the widest possible browsing and searching strategies. Rather than creating a few choice links to augment a single editorial voice, we challenge readers to refine from a superfluity of data their own paths and distinctive interpretive voice.

GENERATING LINKS FROM TEXT TO TEXT

A great deal of previous work has gone into the automatic generation of semantic links — content-based links which connect different documents that are related to one another by subject ([13]; [14, 15]), and other approaches that use automated semantic analysis to link documents (e.g., [16]). We have drawn upon this research, particularly on those aspects most relevant to our collections (e.g., cross-language document comparison between Greek and Latin: [17], [18]). Much of our present work on the London collection has centered on leveraging the information encoded in print to generate hypertextual links within our digital library.

In viewing 19th century English texts from the perspective of a twenty-first century American reader, two things about their original readership stand out. First, they knew more

Latin and Greek than does the average reader today. Latin and Greek remained widespread in the British curriculum through the nineteenth century. The current London collection includes more than 1,500 passages in Latin — some of them fairly substantial, almost none translated into English. Second, they were familiar with many people, places, and topics that are no longer part of an average reader's general knowledge. These observations suggest two useful services a digital library could provide:

- By tagging classical languages (and not simply as italics), we could, as we do in classical Perseus, link inflected words to grammatical analyses, dictionaries and other linguistic support tools, making the embedded Latin and Greek quotes accessible to a wider audience.
- By tagging the names of people, places, and topics, we could link them to reference works that provide glosses and further information for readers unfamiliar with the period.

In the Greco-Roman Perseus, we have long added links from English words to what we optimistically termed the Perseus encyclopedia: several thousand small glossary entries and several hundred essays. The approach was simplistic: we added links from every instance of Homer to information about the poet and made no attempt to separate out references to, for example, "Winslow Homer". Our knowledge base also emphasized the fifth century; thus users would find eight Cleopatras, but not the famous one who allied herself with Marc Antony. Furthermore, precision and recall measures were hard to establish because they differed from text to text. Nevertheless, users made considerable use of the automatic links and they seemed to provide an important service.

We therefore set out to create a similar service for the new London collection, where we faced two basic problems. First, we were starting from scratch and had access to very little preexisting digital data that we could ourselves make freely available; we needed to build up a useful knowledge base in a relatively short period of time and with limited resources. Second, the form of proper names was more complicated: where most of our classical names were single words, the London collection had much more complex phrases: e.g. "St. Martin in the Fields" (with various combinations of spaces and hyphens).

In an ideal world, we would create a unified authority list for every proper name in the collection, but in practice, of course, this was not feasible. We needed to collect as many authority lists as we could and automatically create a unified resource that could create links from text spans to supplementary information. We wanted the user to be able to click on *Fleet Street* and then see pictures of Fleet Street, articles about Fleet Street, maps that included Fleet Street and any other relevant information.

We drew upon a variety of different resources. First, we collected conventional reference works and mined these for links. For information about famous people, we entered the 1903 one-volume index and summary volume of the *British Dictionary of National Biography:* this included names, dates and brief biographies for roughly 34,000 famous individuals. For places, we included Henry Wheatley's 1891 three volume *London Past and Present,* an encyclopedia with 9,800 entries. For geographically referenced street names, we entered a gazetteer to G. F. Cruchley's 1843 *New Plan of London*. This compact lists mentions 4,800 streets and locates them within 1/2 mile quadrants, providing a coarse but often effective georeference. We also acquired (from Bartholomew Mapping Solutions) a modern dataset for London, with vector data for more than twenty thousand contemporary streets. The two datasets complemented each other, because while the Bartholomew dataset provided vector data and much better information, many streets have disappeared since the mid nineteenth century either because of development or simple name changes. Of the 90,000 phrases that we have tagged as possible street names, only 53% (48,000) were in the contemporary Bartholomew dataset.

Besides these conventional authority lists, we also identified other sources of proper names for linking. Chapter and section headers, for example, tend to be discursive (e.g., "A Description of the Westminster Abbey"). Also, because many of the books we selected describe the history and topography of London, their tables of contents are often hierarchical and the section headers rich in proper names. We extracted 4,500 explicit headers. Less structured books provided other typographic clues as to the relevance of a page or paragraph. Augustus Hare's *Walks in London,* for example, uses italics to mark (among other things) significant place names, and several hours of labor allowed us to tag 1,500 italicized phrases as place names suitable for generating automatic links. Thomas Pennant's early nineteenth century *Popular London* italicizes every proper name, thereby reducing the heuristic value of italics for automatic place-name extraction, but it includes marginalia that we quickly mined for another 1,500 link phrases.

We also turned to image captions as an additional source of links. While we were able to collect a modest number (at present less than 1,000) of new and newly captioned color images, we drew heavily upon engravings from books in the collection. We now have more than 10,000 captions linking the user to images. If the user clicks on a link from Fleet Street, for example, she will be discover that there are 98 images whose captions contain "Fleet Street."

The aggregate authority lists described above allow us to generate 284,000 automatic links for the 11,000,000 words in the collection — roughly one word in forty has an automatically generated link. We generate most of these links at runtime, using a fairly efficient algorithm to compare each text against a large list (> 200,000) of multiword matches. To minimize false hits, we prevent common words from initiating matches.

The following reproduces the output from a paragraph on Fleet Street:

> There were certainly rough doings in Fleet Street in the Middle Ages, for the City chronicles tell us of much blood spilt there and of many deeds of violence. In 1228 (Henry III.) we find, for instance, one Henry de Buke slaying a man named Le Ireis, or Le Tylor, of Fleet Bridge, then fleeing to the church of St. Mary, Southwark, and there claiming sanctuary. In 1311 (Edward II.) five of the king's not very respectable or law-fearing household were arrested in Fleet Street for a burglary; and though the weak king demanded them (they were perhaps servants of his Gascon favourite, Piers Gaveston, whom the barons afterwards killed), the City refused to give them up, and they probably had short shrive. In the same reign, when the Strand was full of bushes and thickets, Fleet Street could hardly have been continuous. Still, some shops in Fleet Street were, no doubt, even in Edward II.'s reign, of importance, for we find, in 1321, a Fleet Street bootmaker supplying the luxurious king with "six pairs of boots, with tassels of silk and drops of silver-gilt, the price of each pair being 5s." In Richard II.'s reign it is especially mentioned that Wat Tyler's fierce Kentish men sacked the Savoy church, and part of the Temple, and destroyed two forges which had been originally erected on each side of St. Dunstan's Church by the Knight Templars. The Priory of St. John of Jerusalem had paid a rent of 15s. for these forges, which same rent was given for more than a century after their destruction.

Proper names such as *Piers Gaveston*, *Fleet Street*, and *church of St. Mary* are automatically recognized. The authority list does not include *Priory of St. John of Jerusalem* but it contains *Priory of St. John* and *Jerusalem*. Although (or because) Wat Tyler is a famous historical figure, the header with his name in the DNB is unusually complex and the phrase *Wat Tyler* was not generated. Nevertheless, a reader clicking on *Tyler* would find the appropriate Tyler among the six Tyler entries in the DNB.

Clicking on a link (e.g., *Fleet Street)* calls up a list of resources whose headers, marginalia, and entry keywords are related.

Fleet Street" is in descriptions of...
1 Hare Chapter
By Fleet Street to St. Paul's., Fleet Street
95 Images
2 London sites
1.Fleet Street
2.Fleet Street Hill
10 Thornbury chapters
1.Fleet Street (Northern Tributaries--Shire Lane and Bell Yard).
2.Fleet Street (continued).
3.Fleet Street (continued).
4.Fleet Street (Northern Tributaries--Chancery Lane).
5.Fleet Street (Northern Tributaries--continued).
6.Fleet Street--General Introduction.
7.Fleet Street Tributaries--Shoe lane.
8.Fleet Street Tributaries--South.
9.Fleet Street Tributaries.
10.Fleet Street (Tributaries--Crane Court, Johnson's Court, Bolt Court).
Wheatley, London Past and Present (1891)
1.Fleet Street,

In practice this method of secondary link generation works much better than we had hoped. Developing useful precision and recall measures is problematic because the applicability of this strategy varies from document to document; furthermore, judgments of relevance vary depending upon the reader's purposes. Nevertheless, in practice it is not difficult to recognize dubious links. The casual user accustomed to carefully edited links may find the system off-putting, but the active reader who is eager to find out more about James Barry, for example, will welcome the ability to find a picture of the artist and will be willing to determine which of three James Barrys in the Dictionary of National Biography is the appropriate one. Nevertheless, a feature that lets users switch among different kinds of display depending on their preferences and information-seeking needs is clearly desirable.

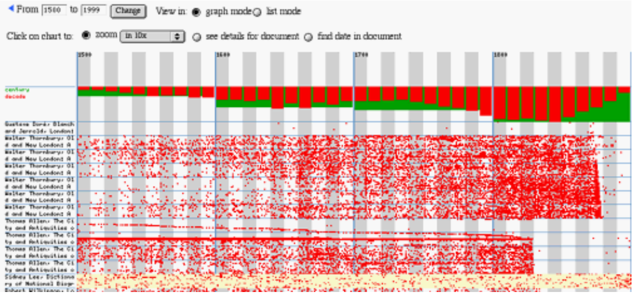
Documents that discuss many disparate places and historical personages that were famous in the nineteenth century obviously benefit most from this environment, but these links also help contextualize works that occupy a largely fictive London. We automatically identify, for example, more than one hundred and fifty London locations in Dickens' *Our Mutual Friend*. Likewise, the reader confronting the phrase "the Lord Chancellor sitting in Lincoln's Inn Hall" in the opening chapter of *Bleak House* will find links to a picture and a description of that building.

Links to Visualize Time and Space

Space and time are fundamental axes for most historical collections. We decided to extract as much temporal spatial information as possible, with the goal of generating useful maps and timelines automatically.

Given its chronological focus, it is not surprising that the London collection contains many dates. Early dates usually have labels such as *A. D.* and *B. C.* to disambiguate them from other small numbers. (The consistency of this practice varies, of course, from book to book.) Furthermore, in the samples that we have examined, more than 98% of the unlabelled numbers between 1000 and 2000 in running text are dates. Most of the falsely recognized dates in this corpus come from tables — a class of data structure on which we have not yet begun serious analytical work. Overall, we have automatically identified more than 69,000 dates in the London materials; by contrast, classical source texts contain few precise dates.

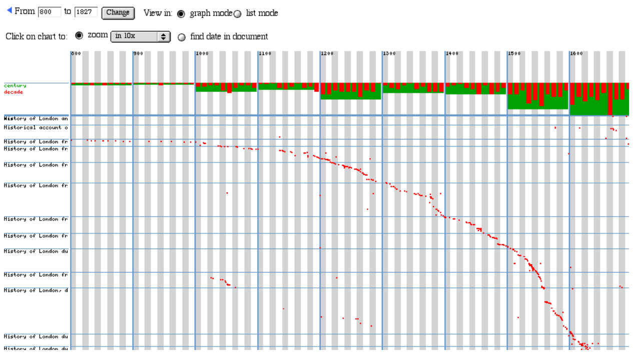
Electronic timelines are hardly new (e.g. [19], [20]). In our case, we generate them from automatically extracted data as a visualization tool for documents and collections of documents (see figures 1 and 2).


From 1500 to 1999 Change View in: ● graph mode ○ list mode

Click on chart to: ● zoom in 10x ○ see details for document ○ find date in document

1500
1600
1700
1800
century
decade
Gustave Doré; Blanch and Jerrold; London
Walter Thornbury; Old and New London: A
Walter Thornbury; Old and New London: A
Walter Thornbury; Old and New London: A
Walter Thornbury; Old and New London: A
Walter Thornbury; Old and New London: A
Walter Thornbury; Old and New London: A
Thomas Allen; The City and Antiquities o
Thomas Allen; The City and Antiquities o
Thomas Allen; The City and Antiquities o
Thomas Allen; The City and Antiquities o
Sidney Lee; Dictionary of National Biogr

Figure 1: Part of the timelines generated for the London collection. The x-axis tracks dates and the y-axis lists the titles of books within the collection. The top bar exhibits aggregate date counts by decade and century and shows that the collection as a whole increases its coverage over time, with richest coverage focused on the 19th century. The bottom section plots dates in separate books, including six-volume and four-volume descriptions of London and, at the bottom, the summary volume of the Dictionary of National Biography. Note that the slight rightward creep of the timelines above lets us see that the two multi-volume descriptions of London were produced in installments over time.


From 800 to 1827 Change View in: ● graph mode ○ list mode

Click on chart to: ● zoom in 10x ○ find date in document

Figure 2: A Timeline for an Individual Document (in this case a narrative history of London). The y-axis lists chapters while the x-axis plots dates. A user can zoom into the timeline and/or use this as a front end to the text: clicking on a dot for 1666, will call retrieve the particular page and will highlight the date. The stretch of red dots curving downwards in the middle is the temporal signature of a narrative history moving through time: the dates move steadily forward in time (i.e., they move right on the timeline on top) as we move through the text (measured by the y-axis, with chapter breaks as blue horizontal lines and marked by labels in the left hand margin).

Integrating maps with each other and with texts

The London collection at Tufts contains approximately 50 historical maps ranging in date from 1790 through the end of the nineteenth century. The integration of geographic information systems (GIS) with a larger digital library has been a long-term interest for us [21] and the extensive and precise spatial data available for London opened up possibilities not feasible with our much sketchier knowledge of ancient Rome or Athens. We georeferenced each map, aligning the historical maps to a modern GIS. Each map varies somewhat from the others, but the overall alignment works well and we can now locate the same subset of London in any map within the collection, comparing historical maps to one another or to the modern GIS. At present, we have georeferenced two dozen maps. The time required for georeferencing is less than one hour per map.

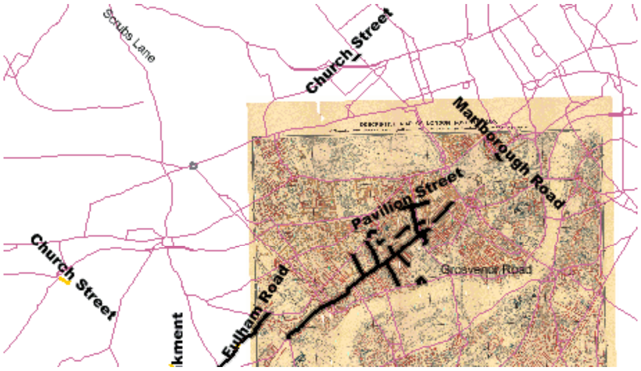


Figure 3: The above map plots vectors from a modern GIS for all the streets mentioned in a table from the 1902 edition of Charles Booth, *Life and Labour in London*. We have overlaid the modern GIS data on a georeferenced map from the period (in fact from Booth). Although some street names (such as Church Street) are ambiguous (thus limiting precision) and the modern GIS picks up no more 53% of the possible street names (thus limiting recall), the automatically generated map clearly reveals the geographic context. The user can now zoom into the modern GIS, historical map, or both.

Finally, we used the *Getty Thesaurus of Geographic Names* (TGN) to search for major geographic features. The TGN has proven to be the most difficult source to leverage. Not only is the TGN huge (more than 1,000,000 names for 886,000 locations) and ambiguous (92% of the place names that we actually encounter can refer to more than one place), but American practices of naming render semantic classification particularly challenging: Hot Coffee is the name of a town in Mississippi, for example, and there is a Monday in both Ohio and Missouri (as well as a Paraguay).

FUTURE WORK

We are clearly at an early stage of development and a great deal more could be done at every level. The scattered authority lists should be unified. We need to develop better tools to disambiguate and to filter what the users see when they pursue an automatic link. We need more content to create a richer environment for browsing and exploration. We need to develop evaluation measures that take into consideration the disparate materials within, and audiences for, this collection. Short term issues include the following:

- **Other sources of link data: Arguments and Conventional Indices:** There are other sources of information that we can use to generate useful links. Many 19th century books include brief, itemized “arguments” that summarize the content of a chapter. These break down into lists with items separated by dashes and can easily yield discrete phrases similar to the headers that we have already mined. Conventional indices likewise provide a wealth of information, including brief descriptions of people and places that do not appear in the larger reference works. Even brief hand-generated indices can disambiguate referents (e.g., the Smith on page 12 is "John" and that on page 32 is "Mary"; or the "All Saints' Church" on page 212 is in Blackheath while that on page 461 is on Margaret Street). Older books often have separate indices for people and places, thus helping bootstrap the problem of semantic classification (e.g., is Wellington a person or a place?). Some indices are as long and informative as entire books: the index to the six volume *Old and New London* contains half a megabyte of raw text and 15,000 page references to 5,600 disambiguated people and places. We need to develop strategies to mine such resources.

- **Quotes and Citation Linking**: Designers of digital libraries now routinely scan their source documents for citations and where possible convert these into active links ([22]; [23]; [24]). Classicists have been careful to establish and then maintain standard reference schemes so that the citations in nineteenth century commentaries, grammars and lexica normally work with contemporary editions. We have thus been able to mine our on-line classical reference works for more than 900,000 links. Of these, 380,000 are "commentary" notes that cite not only "Vergil Aen. 1.1" but one or more words within that reference (e.g., *arma virumque*): since each commentary note is part of a defined chunk of text, the 380,000 commentary notes are "span-to-span" links. If we follow the Dexter Hypertext reference model [25], we can generate 900,000 "LinkTo" and 380,000 "LinkToAnchor" objects, thus converting each citation into a bi-directional link. The average page of Greek and Latin text in Perseus has nine links pointing into it. For highly canonical texts such as the *Iliad*, the number of links already exceeds 100 per page. For us such density is a feature as it allows us to study problems of filtering and visualizing dense, relatively stable collections of links.

  The London collection is highly intertextual. Many of the works cite earlier authorities extensively — in some cases, more than half of a text consists of quotes from earlier authorities. In fact, over twenty percent of the collection as a whole consists of quoted material. Many of these earlier authorities are, or will in the foreseeable future become, parts of the collection. We should be able to generate a rich web of links, allowing us to see links to and from individual passages and to visualize the relationships between documents (e.g., who cites which parts of which documents). For anyone studying the development of discourse about London such links are essential.

  Unfortunately, the London books rarely use conventional citations. They will often refer to "Stow" without providing any typographic or formatting clue that Stow is an author. Even when an author cites another by page number, the edition (and pagination) cited may be different from the one that we have online. And, indeed, the citing work often contains no a bibliography and fails to specify which edition it happens to be citing. We have carefully used the distinction marked by the <Q> and <QUOTE> tags in the TEI DTD [26] to distinguish between literary inventions (e.g., the dialogue of characters within a novel), and true quotations drawn from sources external to the text. A digital library system should be able to search its own and federated holdings to locate the source for any text enclosed in <QUOTE> tags. If the query string is extensive enough and the source text is on-line, the chances of retrieval are good (if one can choose ahead of time, five words are usually enough to define a document: [27]). The average <QUOTE> element contains more than fifty words and this should enough data to retrieve the source document if it is available and on-line.

- **Tabular Information**: The London collection contains at present 1,600 tables with 154,000 elements. These need to be mined for data. Several of the works that we include (Mayhew's *London Labour* and Booth's *Life and Labour*) contain important statistical information that would benefit from visualization within a GIS. Many of the books contain scattered tables with prices and wages illustrating social and economic history.

- M**onetary sums**: Monetary sums are another class of easily extracted and historically significant data — relative prices for commodities and labor are both important for scholarship and useful for giving students a sense of what people purchased and how expensive things were at a given time. The precision of monetary sum extraction is good because the texts contain various labels to indicate when number defines a currency. Where tables primarily affect the precision of our date tagging, they conversely reduce our recall of monetary sums. Our collection contains many historical lists of products and their prices: e.g. a table of prices for fowl in 1274 ("the best hen," for example, cost 3s. 2d.). As we do not yet interpret the forms of tables, we currently lose these values. Even parsing simple tables will be useful because such a process will not only yield more monetary sums but will firmly bind these monetary sums to their referents. Nevertheless, we have extracted more than 10,000 monetary sums. Simply allowing users to search for similar sums of money would be useful. Our goal is to associate those sums with their probable referents as well (e.g., "3s. 2d." refers to the cost of the "best hen").

- **Temporal Spatial Querying**: Given automatically generated timelines and maps, the next logical step will be to query the collection by time and space: e.g., search for documents relevant to the area around St. Paul's in the 1630s.

- **Providing Link Services to External Datasets**: However much work we do on London (or any other subject), no one collection will contain everything of value. We have worked to create an initial critical mass of information on London both because we felt that this would be useful in itself and because we hoped to build an extensible environment. We will continue to expand our internal collection, but we also plan to provide linking services for third party resources (e.g., "value added surrogates" [28]). Others (e.g. [29]) could filter their documents through our linking and visualization tools. We would thus offer linking services similar to those contemplated as part of the Open Citation Project ([30]) but covering other categories including people and place names, as well as specialized language tools (e.g., links from inflected foreign language terms to their dictionary entries). The rise of XML will immensely simplify such services, since well-formed XML fragments can readily contain detailed formatting information that could enhance the precision and recall of any third party linking service.

CONCLUSIONS

Generating metadata from diverse and opportunistically acquired sources has proven extremely useful. While a great deal of effort could profitably be spent merging and resolving inconsistencies between the various authority lists that we have collected, the quickly assembled materials at hand have proven surprisingly effective. While the approach that we are pursuing may not scale to collections that contain thousands of distinct authority lists, we tentatively believe that this relatively simple approach will work well with hundreds of documents and hundreds of millions of words. We suspect that scalability will not prove a major problem for the foreseeable future because crucial reference works are much scarcer than general documents. Thus if the collection increased by two orders of magnitude, the number of key reference works would increase much more slowly and the aggregate pool of potential automatic links slower still.

- We would urge anyone bootstrapping a digital collection on a coherent subject to begin, if at all possible, with creating well-structured on-line key reference works. Such reference works are often very expensive and difficult to manage, but they lay a foundation that may make more conventional materials easier to add and then make these materials more useful when integrated with the online reference environment. We found this to be the case when we started a Digital Library on Roman culture by entering a dictionary and only then adding texts [31]. The same principle seems to be holding true with London.

- Well-organized XML documents are enormously useful for any finely grained, hypertextual digital library, but the value of XML resides in its ability not only to describe overall document structure but to precisely associate unambiguous identifiers with references to people and places. While readily available XML editors are a desideratum, we also need connectivity between these editors and external databases. We can generate useful automatic links, but these automatic links are only a starting point. Subject experts should be able to go through and refine these links, adding some, removing others and disambiguating still others. A great deal of work needs to be done on user systems (e.g., click on a map to indicate which Springfield is meant in a particular text) and on back end data processing (e.g., systems that can intelligently compare local indices or particular reference works against more global resources like the authority lists from the US Library of Congress or the Getty *Thesaurus of Geographic Names*).

ACKNOWLEDGMENTS

Support for digitizing the London Collection has come from Tufts University and especially the *Berger Family Fund for Technology Transfer*. A grant from the *Digital Library Initiative Phase 2* (NSF IIS-9817484), with particular backing in our case from the National Endowment for the Humanities, supported the system development. We would particularly like to thank our colleagues in the Perseus Project, our collaborator, Gregory Colati, Tufts University Archivist and his colleagues for the contributions in every facet of this work.

REFERENCES

1. Colati, G., *Bolles Collection Overview*. 2000, Perseus Digital Library. http://www.perseus.tufts.edu/cgi-bin/ptext?doc=2000.01.0043.

2. Smith, D., J.A. Rydberg-Cox, and G. Crane, *The Perseus Project: A Digital Library for the Humanities.* Literary and Linguistic Computing, 2000. **15**(1): p. 15-25.

3. Rydberg-Cox, J., et al., *Knowledge Management in the Perseus Digital Library.* Ariadne, 2000. **25**.

4. Crane, G., et al., *The symbiosis between content and technology in the Perseus Digital Library.* Cultivate Interactive, 2000. **1**(2).

5. Crane, G., *Designing Documents to Enhance the Performance of Digital Libraries: Time, Space, People and a Digital Library on London.* D-Lib Magazine, 2000. **6**(7/8).

6. Crane, G., *Generating and Parsing Classical Greek.* Literary and Linguistic Computing, 1991. **6**: p. 243-245.

7. Corns, T.N., *The Early Modern Search Engine: Indices, Title Pages, Marginalia and Contents*, in *The Renaissance Computer: Knowledge Technology in the First Age of Print*, N. Rhodes and J. Sawday, Editors. 2000, Routledge: New York. p. 95-105.

8. Rhodes, N. and J. Sawday, eds. *The Renaissance Computer: Knowledge Technology in the First Age of Print*. 2000, Routledge: New York. 212.

9. Blustein, W.J., *Hypertext Versions of Journal Articles: Computer-aided linking and realist human-based evaluation*, in *Computer Science*. 1999, University of Western Ontario: London, Ontario, CA. p. 180.

10. Levy, D.M. *I read the news today, oh boy: reading and attention in digital libraries*. in *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space&mdash;structure in hypermedia systems*. 1997. Pittsburgh, PA USA: ACM Press.

11. Aarseth, E., *Cybertext: Perspectives on ergodic literature*. 1997: Johns Hopkins University Press. 216.

12. Douglas, Y. and A. Hardagon. *The pleasure principle: immersion, engagement, flow*. in *Proceedings of the eleventh ACM conference on Hypertext and hypermedia*. 2000. San Antonio, TX USA: ACM Press.

13. Allan, J. *Automatic Hypertext Link Typing*. in *Proceedings of the seventh ACM conference on Hypertext*. 1996. Bethesda, MD USA: ACM Press.

14. Green, S.J. *Building hypertext links in newspaper articles using semantic similarity*. in *Third Workshop on Applications of Natural Language to Information Systems (NLDB '97)*. 1997. Vancouver, CA.

15. Green, S.J. *Automated link generation: Can we do better than term repetition?* in *Proceedings of the 7th International World-Wide Web Conference*. 1998. Brisbane, Australia.

16. Shin, D., S. Nam, and M. Kim. *Hypertext construction using statistical and semantic similarity*. in *Proceedings of the 2nd ACM International conference on Digital Libraries*. 1997. Philadelphia PA USA: ACM Press.

17. Rydberg-Cox, J., *Announcing a Greek and Latin Synonym Tool*. 1999, Tufts Universiity: Medford, MA. http://www.perseus.tufts.edu/PR/syn.ann.html.

18. Rydberg-Cox, J., *Word Co-Occurrence and Lexical Acquisition in Ancient Greek Texts.* Literary and Linguistic Computing, 2000. **15**(2): p. 121-129.

19. Kumar, V., R. Furuta, and R.B. Allen. *Interactive Timeline Editing and Review*. in *Digital Libraries '98*. 1998. Pittsburg PA USA: ACM.

20. Kumar, V. and R. Furuta. *Visualization of Relationships*. in *Proceedings of hypertext '99 on Hypertext and hypermedia*. 1999. Darmstadt, Germany: ACM Press.

21. Chavez, R.F. *Using GIS in an integrated Digital Library*. in *Proceedings of the fifth annual ACM Digital Library Conference*. 2000. San Antonio, TX USA: ACM Press.

22. Hitchcock, S., et al. *Citation linking: improving access to online journals*. in *Proceedings of the 2nd ACM International conference on Digital Libraries*. 1997. Philadelphia PA USA: ACM Press.

23. Baldonado, M.Q.W. and T. Winograd. *Hi-cites: dynamically created citations with active highlighting*. in *Conference proceedings on Human factors in computing systems*. 1998: ACM Press.

24. Lawrence, S., C.L. Giles, and K. Bollacker, *Digital Libraries and Autonomous Citation Indexing.* IEEE Computer, 1999. **32**(6): p. 67-71.

25. Halasz, F. and M. Schwartz, *The Dexter Hypertext reference model.* Communications of the ACM, 1994. **37**(2): p. 30-39.

26. Sperberg-McQueen, C.M. and L. Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange*. 1994, Text Encoding Initiative: Chicago and Oxford.

27. Phelps, T.A. and R. Wilensky. *Robust Intra-document Locations*. in *9th World Wide Web Conference*. 2000.

28. Payette, S. and C. Lagoze, *Value-Added Surrogates for Distributed Content.* D-Lib Magazine, 2000. **6**(6).

29. Levenson, M., D. Trotter, and A. Wohl, *Monuments and Dust: The Culture of Victorian London.*, Institute for Advanced Technology in the Humanities. http://www.iath.virginia.edu/london/.

30. Harnad, S. and L. Carr, *Eprint archives through open citation linking (the OpCit project).* Current Science, 2000. **79**(5): p. 629-638.

31. Crane, G., *Extending a Digital Library: Beginning a Roman Perseus.* New England Classical Journal, 2000. **27**(3): p. 140-160.

32. Landow, G.P., *Hypertext 2.0*. 1997, Baltimore: Johns Hopkins University Press.

33. Joyce, M., *Of two minds : hypertext pedagogy and poetics*. Studies in literature and science. 1995, Ann Arbor: University of Michigan Press. viii, 277.

34. Murray, J.H., *Hamlet on the holodeck : the future of narrative in cyberspace*. 1997, New York: Free Press. xii, 324.

35. Baker, N., *Double fold : libraries and the assault on paper.* 1st ed. 2001, New York: Random House. xii, 370 , [4] of plates.