The Perseus American Collection 1.0

Gregory Crane,Alison Jones

March 13, 2006

Contents

Status of Draft: This covers the basic topics of what we have done and provides initial evaluation of current results but only begins to document related work and the contributions that others have made on which we have drawn.

1 Introduction

The initial release of the Perseus Nineteenth Century American collection contains 55,000,000 words and is designed to demonstrate in a practical and public fashion some ways in which text mining and information extraction can be applied to historical documents. Information extraction involves locating specific types of data from either structured or unstructured documents, while text mining seeks to discover useful knowledge or previously undiscovered patterns from unstructured text [56]. Named entity recognition, a particular form of information extraction, seeks to identify references to particular kinds of entities such as people, places or organizations. The potential uses of these technologies in historical documents is a topic that has received surprisingly little attention, yet interest in this area is growing. Two recently announced projects of note include the University of Sheffield's plan to explore data mining technologies in a series of digitized eighteenth century materials[1] and the NORA project, which is examining how text mining technologies can be used in currently existing digital libraries of historical materials.[2] Recent work by Ian Witten, et. al has also considered the potential importance of text mining, particularly named entity extraction, for making digital libraries easier to browse. Of particular relevance, is their suggestion that extracting named entities such as person or place names and building them into separate browsing structures or searchable indexes could potentially be of great benefit to users[94].

[1] http://www.hrionline.ac.uk/armadillo/sources.html
[2] http://nora.lis.uiuc.edu/description.php

In this initial release, we particularly focus on the problem of automatically identifying people, dates, places, organizations and other named entities. Automated systems have added 12,000,000 tags to the collection, identifying 500,000 references to organizations, 600,000 dates, 1,000,000 place names, and 1,500,000 personal names. Users can search for mentions of, or documents relevant to, dates (e.g., "July 4, 1863"), places (e.g.,"Springfield, MA" vs. "Springfield, IL") or people (e.g., "Robert E. Lee" vs. "Fitzhugh Lee").

At present, the automated system is able to correctly identify approximately 75% of the personal names and 85% of the place names in the corpus as a whole, with performance rising to around 90% for both functions in Civil War publications which form the largest component of the collections. While much can be done to improve these results, this level of performance is sufficient for many real-world tasks.

This automated analysis offers three strategic advantages. First, it is scalable: the same system could process either small collections or those that exceeded a billion or more words. Indeed, the larger the collection, the more accurate results would become, as machine learning algorithms developed more precise models of topics within the collection. Second, these methods can be improved with the refined results then propagating out to improve tagging in the collection as a whole. Third, the results can themselves be selectively refined and used as training sets to improve results in other similar document collections.

In May 2005 we downloaded and analyzed the links within Wikipedia. Within the then 500,000 articles, there were approximately 15,000,000 "disambiguating links," i.e., links from a particular mention of an unspecified name (e.g., "Washington," "Springfield,") to a Wikipedia article for that particular entity (e.g., "George Washington," "Springfield, IL"). We found that these community generated links were not only extensive but highly accurate, and a manual survey of two hundred links found only two that were problematic. Similarly, in a survey of twenty five biographies in Wikipedia by Roy Rosenzweig, he found that only five of them had clear cut errors, most of which were "small and inconsequential"[78].

As the example of Wikipedia illustrates, immense amounts of productive human energy are available to reinforce what the machines can do, creating a dynamic interaction among machines, collections and their users. The challenges of integrating historical corpora, automated systems, and user corrections and annotations into a workable system that can learn and improve on its own results is a significant problem. A great deal of recent research has examined how to make information extraction systems portable between different domains, how to learn new extraction rules from annotated corpora, and how to embed these learning systems into end user tools [21].

Douglas Oard also recently argued that a key question facing digital libraries is how to utilize systems that learn from examples in order to build more robust and scalable collections.[60] In addition, despite rapid advances in automated systems that allow tagging performance at near human levels, these systems that learn still require materials to learn from or training data such as annotated corpora. This problem has often been labeled the "knowledge acquisition bottleneck." In fact, a recent overview of semantic annotation

platforms used to support information extraction found that no system is yet able to automatically identify and classify all entities in a given document with complete accuracy [72]. Consequently, human users must be integrated into the machine learning cycle. Luca Gilardoni has recently proposed that systems and interfaces must be developed that are able to turn "end users into seamless training corpora builders"[32]. He has proposed that the most significant challenge is how "to collect the information in the form we need to use to feed our learning algorithms, in such a way to keep the users happily and (possibly transparently) within the collect-learn-deploy-correct cycle." A great deal of recent research has thus focused on how to best solicit and harness user contributions into the knowledge capture process and then feed that information back into the underlying algorithms [75, 59, 3, 12, 11]. We hope that our work may make some contributions in this area.

This document briefly describes the collection and the motivation for this work, then goes on to describe the basic components of the initial system. It also outlines all the major components of this system that apply high precision rules to extract names (e.g., "George Washington," vs. "Washington, DC"), develop statistical models (e.g., Washington appears as a place name four times as often as it does a personal name in a given document), and apply these models to ambiguous data (e.g., in "son of Washington," the systems balances the phrase "son of" against the general priority of Washington as place name and concludes that this "Washington" designates a person). These components can be improved or replaced individually within the existing system architecture or the architecture as a whole can (and surely will) be revised to test new approaches. We hope that others will apply their own systems to some or all of these materials and compare the results. Ultimately we also hope that at least some components of this collection can serve as a reference set against which others can test new techniques.

1.1 Collection contents

The Perseus American collection covers a range of topics designed to provide insights into the needs of various types of use. The main body of the collection consists of a selection of Civil War memoirs, periodicals, and reference works. It is designed to support research within a heavily documented subject area with a relatively narrow chronological focus but wide ranging geographic scope, spanning a substantial portion of the United States. The Perseus American collection also includes a growing collection of local history materials, currently focused on the Boston area, that by contrast concentrates on a much narrower geographic space but covers hundreds of years. Nineteenth century print culture expanded the capacity of small groups to integrate their stories into national topics: histories of Civil War regiments and of individual towns embed detailed information about small groups within a broader historical framework.

By contrast, nineteenth century literary publications such as the works of Whittier draw upon an imaginative space that veers from contemporary events to the legendary and mythological. Individual memoirs and biographies of military figures, politicians, scientists,

and cultural leaders trace paths that move steadily forward in time but wind through space. Reference works such as *Knight's American Mechanical Dictionary*, Frederick Dyer's *Compendium of the War of the Rebellion*, and George P. Rowell and Co.'s *American Newspaper Directory* of 1869 provide not only background information for human readers but also data to help automated systems ferret out and determine references to technical terms, Civil War regiments, and nineteenth-century newspapers. More interpretive reference works such as the massive *Harper's Encyclopaedia of United States History from 458 A.D to 1902* and the *Cambridge History of American Literature* provide windows into earlier patterns of thought (consider the article on "Imperialism" in Harper's Encyclopedia of US History).

A fuller description of the content follows:

- thirty seven volumes of the Southern Historical Society[3], sample volumes from the massive Official Records of the Union and Confederate Armies (Shiloh[4]; Atlanta Campaign[5]), complete sets of the Confederate Military History (e.g., South Carolina[6]), Battles and Leaders of the Civil War[7] and Photographic History of the Civil War[8], memoirs of Civil War participants (e.g., Grant[9], Porter[10], Early[11], Longstreet[12], Alexander[13], Beatty[14]), and early histories of the conflict (e.g., the complete Rebellion Record including the diary[15] and volumes of source documents[16] and poetry[17], Swinton (1866)[18]; Pollard (1876)[19]; Greeley (1866)[20] Comte de Paris (1876)[21])

- local history, with an initial focus on the Boston area including town histories (Cambridge (Paige 1877)[22], Medford (Brooks 1855)[23], Arlington (Cutter 1880)[24], Waltham

[3] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0001
[4] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0057
[5] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0032
[6] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0247
[7] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0007
[8] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0106
[9] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0019
[10] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0034
[11] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0015
[12] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0027
[13] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0130
[14] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0005
[15] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0041
[16] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0135
[17] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0077
[18] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0227
[19] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0227
[20] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0066
[21] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0321
[22] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0228
[23] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0104
[24] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0320

(Nelson 1879)[25]) multi-volume publications of local historical societies (30 volumes of the Medford Historical Society Papers[26], and all 8 volumes of the Somerville Historic Leaves[27]), city directories (Cambridge 1857[28]), guide books (Mount Auburn Cemeter, 1839[29]), personal memoirs (Cambridge Sketches (Merrill, 1896)[30], Olde Cambridge (Higginson, 1900)[31]; Cambridge Sketches (Stearns 1905)[32]), accounts of local monuments (Cambridge Civil War memorial[33]), and miscellaneous publications (e.g., Boston events: a brief mention and the date of more than 5,000 events that transpired in Boston from 1630 to 1880 (Savage, 1884)[34]).

- the overlap of local and Civil War history with an emphasis on Massachusetts: histories of individual regiments (e.g., Mass 19th (Adams 1899)[35]; Mass 54th (Emilio, 1894)[36], 2nd Mass Battery of Light Artillery (Whitcomb 1912)[37]; Bennett (1886)[38]; NY 121st (Best 1921)[39]), histories of Massachusetts in the Civil War (Schouler (1868)[40]; Schouler (1871)[41]; Higginson (1895-1896) vol. 1 Mass regiments[42], officers and men who died[43]; preliminary narrative[44] and vol. 2[45]) and biographies of the fallen from particular communities (e.g., Harvard Memorial Biographies (Higginson 1866)[46]).

- multiple biographies of individual figures such as Charles Sumner (Lester 1874[47],

[25] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0325
[26] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2005.05.0001
[27] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0309
[28] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0210
[29] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0211
[30] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0224
[31] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0157
[32] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0155
[33] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0149
[34] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0235
[35] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0168
[36] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0151
[37] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0194
[38] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0193
[39] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0171
[40] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0152
[41] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0169
[42] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0222
[43] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0223
[44] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0219
[45] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0233
[46] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0229
[47] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0184

Nason 1874[48]; Pierce 1877 vols 1[49], 2[50], 3[51], 4[52]) and William Lloyd Garrison (Francis Jackson Garrion 1885-1889: vols 1[53], 2[54], 3[55], 4[56]; Crosby (1905)[57]; Chapman (1921)[58]) as well as Ulysses S. Grant (Crafts (1868)[59], Badeau (1885)[60]; Badeau (1887)[61]; Wister (1901)[62]; Lyman (1922)[63]). These are designed to illustrate problems of aligning disparate narrratives about the same individuals.

- the primary published works of Thomas Wentworth Higginson (e.g., *Army Life in Black Regiment*, 1870[64], *Women and Men*, 1888[65], *The New World and the New Book*, 1891[66]) not only bring to light the work of a remarkable anti-slavery activist and intellectual but allow us to study the problems of creating a virtual comprehensive edition of a well-published author.

- the works of John Greenleaf Whittier[67] illustrate not only the problems of representing an existing print edition as part of a larger digital collection but also of applying a named entity identification system, initially optimized for historical works, to a literary corpus.

- George Bancroft's 10 volume History of the United States (1859)[68], which reflects the nation's image of itself at the dawn of the Civil War and illustrates the challenges of managing a history that covers several centuries.

- surveys of American Literature representing the state of thought in late 19th/early 20th century: Cambridge History of American Literature (vol. 1[69], 2[70], 3[71]), Reader's

[48] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0150
[49] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0206
[50] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0207
[51] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0208
[52] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0217
[53] http://www.perseus.tufts.edu/hopper/collection.jsp?collection=Perseus:collection:cwar
[54] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0186
[55] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0201
[56] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0205
[57] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0159
[58] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0158
[59] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0063
[60] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0259
[61] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0262
[62] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0061
[63] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0069
[64] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0021
[65] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0192
[66] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0267
[67] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0312
[68] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0326
[69] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0068
[70] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0199
[71] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0212

History of American Literature (Higginson 1903)[72]; Short Studies of American Authors (Higginson 1880)[73]); Perry (1921)[74])

- reference works such as Dyer's Compendium of the Civil War (with coverage of Battles[75], Union regimental histories[76] commands, and a machine generated list of officers and their commands[77], Fox's Regimental Losses (1888)[78]; Knight's Mechanical Encyclopedia (1877)[79], Rowell's American Newspaper Directory (1869)[80], and Harper's Encyclopedia of United States History (1902)[81] provide contemporary information and document ideas of the period.

2 Background and motivation

This collection carries forward more than twenty years of work on digital libraries that exploits the automatic identification and analysis of full text to provide advanced searching, browsing and visualization. Our original focus was upon Greek and Latin language. In the late 1990s we also began to examine other domains, including early modern English, the history of science, and the history and topography of London [10, 19, 20]. David Smith developed a general system to extract dates and places from all texts in the Perseus Digital Library [85, 84]. That work demonstrated that American English posed greater challenges than our Greco-Roman or European collections: Americans were far more likely to reuse the same names such as Springfield and Washington over and over for different places and much less likely to mark place names with semantic classifiers. For example, we have far more cities with names like Jackson, MS, than like Jacksonville, FL.

Evolving work on various language technologies, such as machine translation, automatic summarization, cross language information retrieval, and question answering to name only a few, has immense potential for the humanities. Most research has, however, focused upon contemporary materials and especially news reports. These tend to be structurally homogeneous and to cover only the recent past. Assessments of this work are published in an alphabet soup of evaluation forums (TREC, ACE, CLEF, DUC), but these results have only begun to find their way into working systems accessible to wider audiences (Google News being probably the most important example). While these and similar technologies will, we believe, become fundamental to the work of humanists, we felt that our ability to

[72] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0198
[73] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0216
[74] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0200
[75] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0140
[76] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0146
[77] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0148
[78] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0068
[79] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0138
[80] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0124
[81] http://www.perseus.tufts.edu/hopper/text.jsp?doc=Perseus:text:2001.05.0132

recognize people, places, organizations and other named entities constituted a fundamental challenge that would enhance the performance of many other services.

We therefore decided to create a system that would tag historical materials that, while not so very far removed from contemporary materials, were sufficiently old that they would reveal the problems of using existing techniques and knowledge sources to tag historical documents. Thus we chose to develop a tagged corpus of nineteenth century U.S. English and to expand our tagging coverage from dates and place names to personal names, organizations, and other named entities. Our research has thus shifted emphasis, moving from depth (e.g., intensive analysis of Greek and Latin morphology) to breadth (extracting all place names from documents in a heterogeneous collection) and back again to depth (intensive focus on nineteenth century American English).

Other researchers have also pointed out the limitations of modern knowledge sources and language technologies for historical collections. Researchers at the University of Treste are currently examining how the creation of temporal language models might support the linking of modern search terms to historical equivalents in nineteenth century materials. They argue that "the development of statistical language models requires the availability or a huge digitised reference corpus" and that they are currently limited by not having access to enough digitized historical text corpora on which to train their models [22]. Our American Collection will hopefully come to serve as one potential testbed for related research.

3 Significance: Creating corpus editions

Studies of information seeking behavior frequently emphasize the importance of being able to search for particular people, places, and other proper nouns. A recent examination of the information seeking habits of humanities scholars using digital libraries found that all those surveyed frequently used proper names in their searching, though this strategy often met with only limited success [8]. Several research studies that focused exclusively on historians also cited their preference to search by proper names, particularly the names of individuals and geographic locations [17, 26, 13]. Similarly, a study on the research habits of genealogists, another frequent user group of historical collections, illustrated that they wanted to be able to search by name, distinguish between different individuals with the same name, and be able to link current place names to historical ones in their searches [25].

Named entity analysis also serves as a fundamental technology for any developed cyberinfrastructure. In addition, named entity analysis, and many other language technologies, benefit from knowledge sources. Humanists cannot depend solely upon clever algorithms but must develop annotated corpora, gazetteers and machine readable reference materials to drive emerging systems. Since 2000, our own group has produced maps of place names and timelines of dates automatically extracted from collections, documents, and smaller

units of text [20]. This American collection improves on those earlier results and extends our services beyond dates and place names to include people, organizations and other named entities.

But if the data and possible services are potentially important, the methodology underlying this effort is particularly significant, since this collection constitutes a new kind of publication, which we have called a corpus edition [18, 79]. The corpus edition occupies a middle ground between carefully designed electronic archives which apply traditional manual methods to the digital environment and the vast bodies of material automatically analyzed by Google and other services. Like the traditional editor, the corpus editor aims at perfection. Like the computer scientist, the corpus editor looks for methods that are completely automatic. Corpus editors, however, strike a different balance than their colleagues.

On the one hand, the corpus editor adds value to collections whose size makes traditional methods impractical. The traditional editor will scrutinize every comma of the source text, revise every note, produce carefully corrected indices, and review every decision. The corpus editor must think in terms of error rates and precision, relying upon careful sampling to assess the quality of the whole. The traditional editor can, for example, clarify whether a given "Brown" refers to the anti-slavery activist "John Brown" or one of the many other figures who shared this common name. Even our limited 55 million word corpus contains 1,500,000 personal names, far more than any one editor could have checked by hand. The corpus editor must apply automated methods such as rules and statistical models to determine when a reference to "Brown" describes the famous abolitionist or someone else. While this corpus contains 55 million words, we could apply the same methods and generate similar results for a related corpus of any size. General performance should, if anything, improve as the larger collection would allow us to generate better statistical models for more genres.

At the same time, the corpus editor focuses on a particular topic and can make some assumptions about the materials under analysis. Search engines can make few such assumptions. Researchers have only begun to make progress at extracting useful knowledge from the vast corpora indexed by Google and other search engines [16]. Corpus editors, even when dealing with fairly substantial topics, choose particular knowledge sources to enhance their automated systems. Thus, the corpus editor trying to sort out the various "Washingtons" or "Springfields" in nineteenth century English materials might get better results using a nineteenth century encyclopedia of places with 70,000 entries than a modern gazetteer with millions of entries to disambiguate these common place names.

It may be that automated methods become so powerful that many traditional editorial tasks become as obsolete as a concordance based on index cards. For the immediate future, however, we find that human labor is still needed to refine the results of automated systems. For example, computer scientists have demonstrated a variety of methods to analyze complex word forms. In our classical work, we drew on this knowledge and created a morphological analyzer that was highly tuned to Greek and provided better results than

we could get from more general tools. Another scholar then manually edited the results of our automated analysis to produce a 200,000 word subset of our larger corpus and thus created a more refined corpus for the study of classical Greek epic [57]. We have consequently provided an interface whereby third parties can improve on automated analyses in our Greek and Latin collections. While the American Collection 1.0 does not currently feature correction tools, the next release will extend such tools to people, places, dates and other named entities. Readers will be able to correct those many instances where automated methods identify the wrong Springfield and the wrong John Brown, and these user responses will feed back and update the statistical models driving the automated methods.

The present collection represents a partial step toward a true corpus edition: we have collected and begun preparing more specialized gazetteers, encyclopedias, directories and other knowledge sources but have only begun the process of transforming them into historically relevant knowledge resources that can drive the automated analysis of nineteenth century English. At this stage, we have concentrated most of our effort on bringing standard named entity identification techniques to bear on this collection and beginning to tune them for this particular period. We aim at this stage to provide initial benchmarks for our techniques, coverage and performance, against which to measure subsequent development.

4 Semantic tagging workflow

Semantic markup classifies a given "Washington" as a person or a place, or "June" as a month or personal name. In our work, semantic markup overlaps with identification: because we have access to pre-existing gazetteers of place names, for example, we can not only tag "Boston, Ma.," as a place but can identify it as place number 7013445 in the Getty Thesaurus of Geographic Names (TGN)[37]. This semantic markup takes place in the following stages: the creation of initial XML files, gazetteer based lookup of entities (essentially matching text strings against a large list), rule based look-up of entities (e.g., looking for patterns such as "Lieutenant NAME"), and statistical analysis (is Washington a person or a place in a given context?).

We work with the Text Encoding Initiative (TEI) tagset (P 4.0), using 88 distinct tags in this collection. While we work with multiple DTDs and schemas, the TEI provides core tagsets for modern names such as separate tags for forenames, surnames, and generational markers such as "Jr.". Moving to the schema based TEI 5.0 would allow us to combine the TEI tagset we currently use with other markup schemas. This would become an important feature as we move forward to capturing more complex relationships than the TEI is likely to support (e.g., markup to capture commodities along with their prices and features).

*The stages of production, described above, reflect our current workflow and are approximate. Some rules are applied to ferret out organizations in the "gazetteer" lookup stage (e.g., military units whose names can appear in many permutations: "10th Mass.," "Mass. 10th," "Tenth Massachusetts Infantry," etc.) We also use a geographical gazetteer

Table 1: Named entity tags in the 57 million word American collection

Stage*	persons	places	dates	orgs	all tags**
Initial	19,764	14,076	57	8,509	3,143,438
Gazetteer Lookup	38,590	18,141	57	432,552	6,832,460
Rule based analysis	773,780	553,881	622,033	438,620	12,905,359
Statistical analysis	1,551,513	1,030,355	622,033	519,438	15,202,589

during rule based analysis to help avoid tagging patterns such as "a visit by Sumner, Mass." (where Sumner is the senator from Massachusetts) as place names. Nevertheless, the above figures provide a reasonable view on the contribution that each of the three basic processing stages makes.

** Beginning and end tags are counted once: "<placeName>Springfield</placeName>" counts as a single tag.

4.1 Initial XML file creation

In this first stage, we capture key structural elements: page breaks, text divisions (e.g., chapters, sections, encyclopedia articles), tables, embedded quotes and texts within texts, openers and closers, datelines, signatures and salutations, and notes. We opportunistically tag bibliographic references in text and notes but do not yet attempt to analyze their (often very loose) structure into author, title, and publisher information.

```
   <p>The battle of Beverly Ford, as we call it, or of Fleetwood,
as General Stuart styled it, is interesting in the first place,
because it was the first occasion when the cavalry
of the Army of the Potomac went into action as a
body. The cavalry had been organized by General Hooker
into a corps under Stoneman during the winter of
<pb id=''p.136'' n=''136''/> 1862-63,
```

Initial XML with basic structural markup.

A few documents of central importance and regular structure (e.g., the 1857 directory of Cambridge, MA., *Harper's Encyclopaedia of United States History* , the *American Newspaper Directory*, and Thomas Wentworth Higginson's, *Massachusetts in the Army and Navy During the War of 1861-5*), have been more fully tagged at an early stage of the process. The tagged personal, place, organization and other names of these documents provide a seed bed for subsequent analysis.

Because *Harper's* and the *Cambridge History of American Literature* cite a wide range of publications, and bibliographic references are relatively easy to identify, we have tagged

document titles in these works (7,600 in *Harper's*, 10,500 in the *Cambridge History*). The tagged titles are intended to provide the start of an authority list of documents considered important a century ago. Such a selective list will, we hope, provide a good foundation for identifying bibliographic citations in the rest of the American collection. The next step would be to link these bibliographic citations as well as possible to professional catalog records (thus augmenting what the system knows about each reference).

While a few hand-generated human authored tags for named entities remain in the texts, we have found such hand-tagging to be, on the whole, counterproductive. We can change the format of automatic tagging much more easily than retrofitting older tags, and have also found that machine tagging has proven much more consistent.

The detailed structural metadata generated at this stage has relatively little effect upon the named entity analysis, which works as effectively on full text without markup. Some markup (such as italics) may even by its inconsistent application degrade automated analysis.

4.1.1 Normalizing case

Case information in nineteenth century (and contemporary) English can often help identify proper nouns, but headers and other specialized contexts capitalize all significant words or even at times display all words as capitalized. The resulting loss of case information degrades performance of named entity identification and classification. In the case of our own collection, we regularize case within the source texts themselves: e.g., "I SAW BROWN" becomes "I saw Brown." More conservative editors could leave the source text intact but apply the same methods to surrogate copies of the source text, then transfer the tagging back to the source text.

We capitalize words that are displayed in all caps in the text that are found in the QUOTE, HEAD, TITLE, ITEM, or ARGUMENT tags. In general, words are capitalized (1) if they are at the start of a sentence (2) if they appear capitalized in the current document, and (3) if they appear capitalized more than 25% of the time in the overall document. A word counts as capitalized if it shows up capitalized after a lower case word in standard text.

4.2 List lookup

After the XML files are created, the next stage is gazetteer or list lookup of potential entities. The easiest method for identifying entities is to compare the source text against a large list of terms, such as a gazetteer. This can be extremely efficient and, in some cases, very effective. While some technical language applies precise definitions to common words (e.g., compare the meanings of "strain" in biology and structural engineering), many technical terms are either distinctive (e.g., encephalitis, gastropod) or combines common nouns in unambiguous phrases (e.g.,"singular value decomposition," "information retrieval"). Sys-

tems such as Noosphere (which underlies PlanetMath.org) automatically link technical terms to background information [40]. Such automatic keyword linking has been a part of the Perseus Digital Library since the mid 1990s: the classical collections were homogeneous and the list of keywords relatively small.

Gazetteer lookup is an essential step in many named entity recognition systems, particularly in the identification and disambiguation of place names. The open source system GATE (Generalized Architecture for Text Engineering) created by the University of Sheffield, includes an information extraction system that detects different types of named entities such as persons, organizations and places by employing customized gazetteer lists [5]. Due to the difficulty and time involved in either creating extensive gazetteer lists or finding appropriate already existing ones, the system developers are currently exploring enabling new functionality for GATE that will support the automatic induction of gazetteer lists from annotated corpora [52]. Similar research in the semi-automatic creation of gazetteer lists has been conducted by Olga Uryupina, who used data mining techniques to create gazetteers semi-automatically from the Internet [91].

Despite the time consuming and often expensive nature of either creating customized or finding appropriate gazetteers, they still form an essential component of many named entity recognition systems. A variety of research has been conducted on how to best construct appropriate gazetteer databases of place names [2, 23] as well as how to best implement gazetteers in broader named entity recognition systems. Amitay, et. al. constructed an extensive gazetteer from freely available websources as an essential part of their place name disambiguation system [1]. Pouliquen, et. al. have also reported on the creation of a multilingual gazetteer database that is used to support place name recognition in natural language texts and generate relevant maps [66]. Siefkes used a variety of gazetteers, including a names list from the U.S. census to assign semantic classes to extracted entities [83].

Despite the general usefulness of gazetteers, list lookup frequently breaks down when terms are ambiguous and collections heterogeneous. Two frequently cited issues are the inability of gazetteers to provide exhaustive lists of all potential named entities and their inability to resolve entity ambiguities, or whether an entity is a person or a place [72]. For example, in a collection of predominantly Civil War materials, determining whether an Athens in question is located in either Greece or Georgia becomes a difficult question. As collections become more diverse and the set of keywords grows larger, this list lookup approach becomes rapidly less effective. While we collaborated with colleagues at Johns Hopkins to rewrite our own list lookup system as a service to match technical terms to background data, we also began looking for more powerful solutions [64].

Nevertheless, some entities are easily found through one-word list look up even in fairly large collections. While individual words may be common and ambiguous, combinations of words or keyphrases are much less common and can be more powerful keys for describing objects. Thus, we begin our named entity analysis by scanning for multiword phrases. While technically we could scan for personal names in this way (e.g., "Harriet Lane"),

we found too many instances such as the "Harriet Lane" (a ship) or the "Harriet Lane Society"(an organization). At this stage we thus scan for multiword phrases whose final words probably mark the end of a phrase such as "Oxford University."

Organization names are an open class for which exhaustive gazetteers are rarely available. We experimented with open ended mining of phrases that looked like organization names by creating rules such as tag all instances of "UPPER CASE WORDS + (Association or Society or University)" as organizations. The results were good but we chose not to incorporate them automatically. Instead, we generated a list of candidate organization names which we then scanned for errors. For example, the expression "Martin's Massachusetts Public School System" was flagged as a possible organization because it ended with "system" but it was actually a book by an individual named Martin about the Massachusetts Public School System.

We currently scan for a wide range of organizations. Besides military units, newspapers, and railroads (to which we have paid particular attention), we tag administrative departments and districts, colleges, universities, schools, associations, and federations, to name only a few.

4.3 Rule-based analysis

The next stage of our system is rule based analysis, which attempts to identify patterns that clearly mark various categories of entity. Many information extraction systems, particularly those involved in named entity recognition, make at least partial use of rule based systems, where manually created rules or encoded patterns are used to identify different types of entities. Since manually developing pattern and rules is both difficult and expensive, many systems are now relying on a combination of supervised machine learning and human annotated corpora to optimize performance [56]. At this stage of our process, we try to maximize precision. The results of this phase also form the foundation for the statistical analysis which follows. The major contribution of this rule-based analysis will lie less in its overall technique than in the careful analysis of how effective various rules are in tagging entities in documents that are very different from the news feeds, e-mails and contemporary reports that dominate evaluations of most language technologies.

```
   <milestone unit=''sentence'' n=''1721''/>The
<rs n=''Battle of Beverly Ford'' type=''battle''>battle of Beverly Ford</rs>,
as we call it, or of Fleetwood, as
<persName><roleName n=''General''>General</roleName>
<surname>Stuart</surname></persName> styled it,
is interesting in the first place, because it was the
<num value=''1'' type=''ordinal''>first</num> occasion
when the cavalry of the
<orgName n=''Army of the Potomac'' type=''army''>Army of the Potomac</orgName>
```

```
went into action as a body.

<milestone unit=''sentence'' n=''1722''/>
The cavalry had been organized by
<persName><roleName n=''General''>General</roleName>
<surname>Hooker</surname></persName>
into a corps under Stoneman during the winter of
<pb id=''p.136'' n=''136''/>
<dateStruct value=''1862--''><year reg=''1862''>1862</year></dateStruct>-
<dateStruct value=''1863--''><year reg=''1863''>63</year></dateStruct>,
```

An example of text after list lookup and rule based analysis: list lookup tagged the terms "Battle of Beverly Ford" and "Army of the Potomac." Rule-based analysis tagged "General Stuart" and "General Hooker," recognized that 1862-63 constituted adjacent dates rather than separate numbers 1862 and 63, and labeled sentence breaks.

While the above example demonstrates our coverage of people and places, we also identify dates, monetary sums and various other quantities. There is substantial coverage of dates expressed in various patterns (e.g., "June 12, 1864," "the 12th of June, 1864," "the twelfth of June, 1864," "12 June, 1864," etc.) All numbers are tagged (e.g., "<num value="12">12</num>," "<num value="12">twelve</num>"). Numbers associated with dollar or pound signs are captured as currency sums. Keywords for various measures (e.g., "acres," "kegs," "hogsheads") adjacent to numbers are also captured. In addition, the current version of the named entity system allows for searching of dates.

Scanning noun phrases that end (or begin) with keywords such as "system" or "association" is another example of rule based analysis. Military organizations follow naming conventions so complex that we chose to develop a rule based scanner to tag them more effectively. A simple grammar with rules such as NUMBER STATE SERVICE can recognize "3d Ala. Cav.," "Third Alabama Cavalry," and "2nd Mass. Arty." At present, we have collected 36,670 references to regiments with state, number and service. We have also performed initial work linking citations of Union regiments to the 3,385 units we have identified in Frederick Dyer's *Regimental Histories*. Interestingly, the Massachusetts 20th and Alabama 12th prove to be the most frequently cited military organizations, with 189 and 160 citations respectively. We have also tested other rule based systems to augment our existing lists of organizations (e.g., the PLACENAME NEWSPAPER-TITLE, PLACENAME RAILROAD).

The order of analysis in the rule based stage is quite important. The system generally looks for longer phrases first and then works down to individual words. Place name rules are a partial exception. Although place names may be a bit longer, we first scan for hierarchical statements such as "PLACE, REGION". This rule successfully tags expressions such as "Cambridge, Mass." and "Rome, Italy" but also tags "Centerville, Augusta, Virginia"

which might turn out to be "Centerville, Fairfax, Virginia". In this first pass, we only accept combinations that we can match against the Getty Thesaurus of Geographic Names (TGN) or that we have manually verified. We found too many patterns such as "Next came Smith, Ohio, who said" (a phrase describing a man named Smith who was a representative from Ohio) tagging as place names.

Since we are working with historical collections, many places in our texts have changed names, no longer exist or have shifted borders. Thus, we have rules such as "if you can't find the place in VIRGINIA, try WEST VIRGINIA" (because West Virginia was not yet a separate state in many of our collections). Access to historical gazetteers would allow us to avoid such ad hoc strategies, but the current system yields reasonable results. The lack of historical place names, consistent links between current and former place names, and historical contextual data about placenames in modern gazetteers and library name authority files have been cited as major issues facing the development of more robust geographical searching in digital libraries and traditional library catalogs[9]. One group that implemented a local gazetteer using the ADL standards found that they needed to create large numbers of entries for local regions and places that their users would expect to find [39]. While there are a number of projects that either make use of or have developed historical gazetteers, such as the Great Britain Historical GIS Project[87], the Electronic Cultural Atlas Initiative[45], and the China Historical GIS Project[4], the resulting gazetteers are limited in geographic scope and not suitable for our needs.

Nonetheless, we tentatively tag all states, countries and other geographical units. If we don't find an adjacent place name to confirm a "PLACE, REGION" pattern, but the context suggests a place name (e.g., "WORD, Mass.,"), then we retain the place name. Otherwise, we remove the tentative tag.

This stage of named entity analysis also attaches possible identifiers to matched place names. For example, "Boston, Mass." is linked to number 7013445. PLACE and STATE combinations can, however, be ambiguous – the TGN reports 9 places in Virginia alone named Centerville. In these cases, we include all possible TGN numbers up to a maximum (at present, 12).

Our rule based analyzers primarily focus upon people and places in this release. Before looking for these directly, however, we look for other patterns, besides organization names, that may incorporate personal and place names within them. In our collection, ship names are among the most difficult to recognize. We apply rules such as SHIPTYPE + UPPER CASE PHRASE (e.g., "the frigate President") or SHIPTYPE + POSSIBLE NAME (e.g., "the schooner F. W. Dana") and other contextual clues (e.g., "on board the Congress") to tag these entities.

The last step in the rule based stage is to analyze personal names. To assist in this process, we use a list of rolenames that commonly introduce a name. This list currently includes more than 275 separate prefixes, all accumulated during the course of analyzing our corpus. Some rolenames overlap with personal names such as duke, lord, or dean, while other role names can be combined (e.g., "chief justice," "master sergeant," "lord

admiral") but not all ("duke lord"). Some important role names produce noisy results. For example, the role name "general" leads to the tagging of various expressions such as "General Mills," "General Assembly," and "General Order." Consequently, we have accumulated a stoplist of 386 words that we will not accept as personal names. The system also looks for suspicious patterns, for example, we do not assign personal name tags after an article. While the system will accept "General Grant" it will not allow "the General Grant."

This personal name analyzer also scans for place names that tend to follow the pattern PREFIX PERSONAL NAME (e.g., "Fort Sumter," "Camp Beauregard"). Prefixes such as "fort" and "camp" are treated as if they were personal role names (thus taking advantage of the code that manages personal name extraction), then converted to place names.

We apply two other, more problematic, methods to mine personal names. First, we look for patterns such as "SURNAME, FORENAMES." These are common in indices, tables, and are often the only source that we have from which to mine names. Nevertheless, patterns such as "PARAGRAPH START: Doubtless, Mr. Lincoln," can easily be captured as names "Mr. Lincoln Doubtless." We have needed to restrict this pattern to parts of the document that seem to be indices. Since we are trying to automate as much as possible, we look for key words such as "Index," "Roster," "members," "wounded," and other clues that we are looking at a list. Better strategies may include analyzing the length and opening of paragraphs to identify alphabetized lists.

The final – and most problematic strategy – entails simply scanning the text for one to four initials or known forename abbreviations followed by an upper case word. Such a strategy picks up "N. England," "S. H. Soc." (Southern Historical Society) and similar bits of noise. Nevertheless, this pattern, loose as it seems, plays an important role in ferreting out personal names that show up later. It allows the system to recognize "U. S. Grant" as a personal name and helps it to identify a personal name in the expression "I visited Grant."

Finally, when we have collected all potential personal names, we look again for residual place name patterns. In this phase, we scan for "City of NAME," "NAME County," "NAME River," "Mount NAME," and other prefix/suffix clues that we are dealing with a place name. Ambiguous place suffixes such as Field, Ford, Hill, or Mills are accepted if the suffix is lower case (e.g., accept "James hill" but not "James Hill" as a place name). Instances such as "James Hill" are left for the statistical analyzer to resolve.

In cases where the place name is unique (e.g., "Mississippi River") we attach a TGN number to it. Many place names captured at this stage (e.g., "Smith's Farm") may have no record in any standard gazetteer, while others are so common (e.g., "Elk Creek" with 206 TGN entries) that they may be either impossible to locate or not exist within the larger gazetteer.

4.3.1 Lists of entities

Lists of multi-word entities are particularly challenging in the semantic classification of personal and place names. While we find patterns such as "Generals Smith, Jones, Hill and Jackson", we also find "Generals U. S. Grant, Sherman, Stonewall Jackson, and Robert E. Lee," which require more processing. In addition, we find expressions such as "counties of Middlesex and Suffolk" as well as "Middlesex and Suffolk counties."

4.4 Statistical Classification

In preparation for this stage, the system has identified entities with various semantic classifiers accompanying them (e.g. "Lieutenant Jackson," "the frigate Ohio," "Boston, Mass.") This problem of the semantic classification of named entities into specific categories such as personal, place or organization names has been examined in a variety of related research [29, 62, 81, 80, 55, 47]. The system most similar to ours is that discussed by Rosenfeld, et. al. that takes a hybrid approach to named entity extraction. In the TEG system, the rules for the extraction grammar are written manually and the probabilities for entity identification are trained on an annotated corpus of examples [77]. Our system as well combines both rule based and statistical analysis to first classify and then identify potential entities.

At this point, the system now moves on to classify floating individual terms such as Jackson and Boston. At present, we mainly focus on identifying personal and place names. Nevertheless, to accomplish this limited goal, we need to recognize as many other classes of noun as possible. Most of our errors occur when we misidentify words such as "French" or "Justice" as personal names. In these cases, we do find people with these last names, but we don't yet have a good automatic method in place to calculate the far more numerous sets of times when these words are not personal nouns but instead identify ethnics groups or abstract concepts.

```
   <p><milestone unit=''sentence'' n=''1721''/>The
<rs n=''Battle of Beverly Ford'' type=''battle''>battle of Beverly Ford</rs>,
as we call it, or of
<placeName reg=''Fleetwood, Berks, Pennsylvania''
key=''tgn,2088486''>Fleetwood</placeName>,
as <persName n=''Stuart,General,,,,
reg=''nearbymention:Stuart,J.,E.,B.,''><roleName n=''General''>General</roleName>
<surname>Stuart</surname></persName>
styled it, is interesting in the first place, because it was the
<num value=''1'' type=''ordinal''>first</num> occasion
when the cavalry of the
<orgName n=''Army of the Potomac'' type=''army''>Army of the Potomac</orgName>
went into action as a body.
```

```
<milestone unit=''sentence'' n=''1722''/>The cavalry had been organized by
<persName n=''Hooker,General,,,,''
reg=''mostcommon:Hooker,Joseph,,,:1''>
<roleName n=''General''>General</roleName>
<surname>Hooker</surname></persName>
into a corps under
<persName n=''Stoneman,,,,,''
reg=''mostcommon:Stoneman,nomatch:0''>
<surname>Stoneman</surname></persName>
during the winter of <pb id=''p.136'' n=''136''/>
<dateStruct value=''1862--''><year reg=''1862''>1862</year></dateStruct>-
<dateStruct value=''1863--''><year reg=''1863''>63</year></dateStruct>,
```

The above is an example of text after statistical classification. The floating names "Fleetwood" and "Stoneman" have been classified correctly as a place name and a personal name. Note that the system has posited an incorrect identification for "Fleetwood" – the correct Fleetwood for this reference is not in the currently used gazetteer and the system selects the most plausible match. General Stuart and General Hooker were already classified as personal names, but the system has matched Stuart with the initials "J. E. B." and Hooker with "Joseph." The rules for name matching are embedded in the attribute: there was a nearby mention of "J. E. B. Stuart" in the document while "Joseph Hooker," although not nearby in the text, is the most common Hooker in the text as a whole.

4.4.1 Document vs. collection models

At the moment, we apply two statistical models to each word. First, we assess the usage patterns of the word in the physical source document (usually a book or a pamphlet) itself. If the document refers often to General Early or General Battle, then we will be more likely to tag references to "Early" and "Battle" at the start of a sentence as personal names. Second, if the document based model of term usage does not provide enough guidance, we look at the statistics for usage of the term in the collection as a whole.

Neither of these models is satisfactory in scope and can lead to many false tags. The overall collection provides good statistics if its individual documents refer to the same topics in the same proportions, but even the Perseus American collection contains documents on many diverse topics. Individual documents may contain chapters on a wide range of topics and thus be themselves heterogeneous. Newspapers are perhaps the most challenging because they contain many small articles on unrelated topics, making the development of statistical models largely useless.

Statistical models to assess term usage should reflect usage in documents similar to that being examined. We need (and plan) to create "virtual collections" by clustering

chapter, reference work entries and other document chunks into topical groups. Using clustering algorithms to group or classify large document collections or document chunks by topic is a well established research field in computer science [97, 43, 82, 61] but has only recently been explored by humanities scholars. Some recent interesting work includes [41, 63, 30, 70]. The use of clustering algorithms to assist in named entity identification has also been explored by a number of researchers [89, 65, 42, 31, 51]. We are considering the differing algorithms discussed in this vast literature to create our "virtual collections".

4.4.2 Statistical models

Where the document structure allows us, we add tags to original source documents and check the results. Gazetteers, for example, may contain lists of the format "CITY, STATE." Biographical dictionaries may have entries that begin "SURNAME, FORENAME1, FORENAME2 etc." Many authors only specify uncommon entities so that we may find more references to "Rome, Ga." than to "Rome, Italy," even though most references to "Rome" designate the city in Italy. To address this imbalance, we might tag all references to "Rome" in a general reference work (e.g., *Harper's Encyclopaedia of United States History*) to provide a more satisfactory training set. At present, however, such semi-automatically generated examples constitute only 2.5% of the examples on which we base our statistical models (19,764 of 773,780 personal names, 14,076 of 553,881 place names). Thus the rule based system provides 97.5% of the data underlying our statistical models.

We analyze the results of the rule based system and create statistical models from the results. These models include the following:

1. Occurrence of words by entity class. Thus we currently identify 4,146 instances where we can classify "Washington" into various semantic classes. The most numerous are general place name (2,020), surname (676), city name (647), street name (243), fort name (Fort Washington: 186), and forename (102).

2. Occurrence of entity class. Thus, we can calculate that we have recognized 746,647 surnames vs. 228,942 forenames, etc.

3. Words preceding entity classes. We calculate how often n-grams of one, two and three words precede each class of entity: "on board the" appears 740 times, 662 of which are followed by a ship; "neighborhood of" appears 249 times, of which 235 instances are followed by a place name; "from" occurs 42,674 times, 15,042 of which before a place name, and 8,637 before a personal name. At present, we count only n-grams that appear ten or more times.

4. Words following entity classes. This is the same as the preceding except that it tracks patterns after an entity class. We have not yet implemented this model, but following n-grams promises to address the problem of classifying entities at the start of a sentence.

5. General context words. These seem more promising for problems such as determining which General Lee (of several) or which Cambridge (of many) is meant in a given passage. Since personal and place names often co-occur in the same sentences, general context seems unlikely to distinguish these classes. As we focus again, however, on distinguishing the "Constitution" as document from the "Constitution" as ship, general clues about the overall context are likely to become more useful, if not crucial. A variety of related work in this area may prove helpful in implementing this possible approach [65, 15].

6. Word classes and entity classes. We have not yet calculated models for VERB followed by personal name vs. place name, ADJ followed by personal name vs. place name etc. Such data is, however, worth exploring.

The rule based system from which we derive our statistical models is not perfect. Even if the rules produced results that were 100% accurate, any large corpus includes small errors that can confuse the system. Thus, in a list of generals killed in battle, we find "General John Sedgwick Spotsylvania" rather than "General John Sedgwick, Spotsylvania." Without the comma, the rule set assumes that Spotsylvania is a surname and adds an incorrect example to the model.

We address such noise for now in two ways. First, the system looks for terms that have appeared a minimum number of times in the corpus, since current rules that fail usually do not fail consistently. For example, Spotsylvania may be classed a surname once or twice in a very large corpus but not repeatedly. At present, ten is the minimum number for us to accept an entity without question. Second, many valid entities do appear less than ten times (e.g., "Fairview," "Littlestown," "Weakley"). We thus analyze the appearance of low frequency entities in the overall corpus. If the term looks like a common noun, then we assume an error and do not accept it.

4.4.3 Reference strings

We have currently delayed examining the problem of resolving "reference strings", typically indicated by strings such as "THE + NAME." For example, "The Congress," "the Constitution," and "the President" can designate political entities or ships. "The Ohio" can be a river or a ship, and if a ship, then one of several ships that have borne that name over time. "The Times" and "The Herald" are probably newspapers, but many newspapers use these names at any given time and the specific references vary by context (e.g., London vs. New York). In initial work, we found too many false positives for ships: naval accounts may, for example, contain a burst of references to the ship Ohio, and these seem to skew our statistical model more sharply than similar clusters of references to people and places.

For now, we have set aside such reference strings, mainly analyzing them so that they do not generate false positives among pure named entities. We assume any contiguous set of upper case words following "the" is a reference string. If only a single upper case word

follows, we check to see if we have previously identified any candidate nouns to which this could refer. If we see "U. S. S. Ohio" or "Ohio river," then we accept "the Ohio" as a possible reference string without yet trying to decide whether it points to a ship, a river or some other entity.

4.4.4 Identifying possible names

Strings of upper case words in the middle of a sentence are usually proper names in our collection. We need to collect patterns such as "Thomas Jas. Smith" but avoid tagging expressions such as "...said Thomas Smith. Understanding..." Phrases appearing in sentence initial positions are particularly challenging for we need to distinguish phrases such as "Considering Smith" from "John Smith" and "Early in the morning" from "Early commanded his troops." To solve this problem, we calculate how often the sentence initial word appears capitalized elsewhere in the current document and overall corpus. If we find that the word is capitalized in the middle of the sentence (in practice, if it is capitalized after a lower case word) above a given threshold (currently 10%), then we assume it may be a proper name. If the frequency is too low, then we prune the candidate name (e.g., "Considering Smith" becomes just "Smith"). Besides the maddeningly named "Jubal Early," we find generals named "Battle," "Field," and other semantically ambiguous names.

4.4.5 Classes of proper names

We currently classify floating names into the following categories:

1. Multiword personal names: e.g., John Smith, Julia Ward Howe, Martin van Buren. In this case, we calculate how often we have seen constituent elements as forenames and surnames. Thus we would rule out "Accomplished Smith" but accept "John Smith."

2. Surnames and forenames. Single word personal names are harder to capture than multi-word personal names. These are related categories but distinct and very important: if we can pick out last names, then we have a much better chance at identifying additional information about a given individual. We assume that any surname can appear as a forename, but names do tend to cluster under one group or the other. We find 10,969 passages with Smith as a surname and 111 with it as a forename. (This ten to one ratio surely under-represents the forenames, since our texts provide three times as many surnames as forenames.) By contrast, the name "John" shows up 23,964 times as a surname, 665 as a forename. In this case, the disparity is probably even greater, given the general overrepresentation of surnames. Some names are also very hard to distinguish as either type of name, in our collection the name Thomas appears 6,842 times as a surname and 3,902 as a forename.

Genre heavily influences the probability that a floating name is a forename or a surname: surnames predominate in formal prose, forenames in fiction. Our current collection consists mainly of formal prose, and we can thus make do with a global model of forename/surname frequency. Isolated forenames are particularly rare, for historical accounts typically don't refer to Ulysses S. Grant as Ulysses or Robert E. Lee as Robert. If Thomas is equally common as a forename and a surname, we would currently estimate the probability that Thomas in isolation is a surname at five to one. This figure provides good results in practice, but this figure in particular and our approach as a whole warrant further study. We also need better methods to automatically calculate the probable distribution of forenames and surnames in a given document.

3. Place names. Some of these are multi-word (e.g., Florida Keys) but most are single terms (e.g., Springfield, Washington).

4. Fort names. American forts are named after, and often cited as if they were, people. Thus, we find "Sumter" as a reference to "Fort Sumter." We need to create a special classification to keep such references from erroneously being classed as people.

5. Streets. Street names constitute an important class which supports different services than other place names (e.g., they appear on local maps and can have addresses associated with them). They also often appear without semantic markers, we often we find expressions such as "I was on Elm" as well as "I was on Elm Street".

6. Organizations. Harvard, Yale, Cornell, Brown and other schools are named after people and often cited (as they are in this sentence) without an accompanying "College" or "University." We thus use phrases such as "Harvard University" or "Williams College" to build statistical models to find places where Harvard and Williams designate these institutions.

7. Month names. We capture MONTH DATE patterns (e.g., "May 15," "the fifteenth of May") in the rule based analysis. "March," "April," "May," "June" and "August" are, however, well attested personal names and almost any proper noun shows up as a personal name if the corpus gets big enough. We consider months also as a possible category at this stage when we analyze isolated floating names.

4.4.6 Classification methods

Sentence initial words pose a special case since we cannot use capitalization as a method to locate proper names. Thus we cannot tell if "Early," "Battle," etc. are proper nouns or simply words capitalized at the start of a sentence. At the same time, even if we did know that we had a proper noun, we cannot use the preceding words to classify a sentence initial term.

We have collected n-grams of subsequent words to help improve classification (e.g., in "Early went," we probably have a name; with "Early in the day" we probably don't). At present, we have not yet implemented this method. Instead, we conduct two tests. First, we check to see if the word in question shows up as a proper noun (i.e., capitalized after a lower case word) more than 10% of the time. If not, then we assume this is not a proper noun. If it does appear more than 10% of the time, then we simply pick the most likely class for this particular term: thus, sentence initial "Washingtons" would always be places, since most Washingtons in our corpus are place names. PRECISION/RECALL?

We apply two basic methods when classifying any names still remaining. First, we scan the preceding context (for the moment, the preceding ten pages in the print original) for labeled examples. If we see "Jackson" on page 48, we scan backwards until page 39 looking for a disambiguating example (e.g., "General Jackson" or "Jackson, Miss."). If we find such an example, then we classify the current "Jackson" as another instance of the same class. If the preceding example has a unique identifier attached to it, then the new instance inherits that as well. Thus, we map "Jackson" in Mississippi to the city and attach the TGN number 7016129. (Note that we default to the city rather than Jackson County. We would require "county of Jackson" or "Jackson county" in a previous disambiguating example to tag otherwise)

The choice of ten pages reflects short term experimentation. The question of how to measure the scope of a labeled citation in a given document (e.g., how long after we see a reference to "Philadelphia, Miss." should we assume that the next Philadelphia is also in Mississippi and not the major city in Pennsylvania) requires systematic study. A ten page window works well enough in extended narrative, but the reference scope in entry based reference works and especially in newspaper articles is very short, with abrupt changes of topic from one brief article to the next. Reference scope reinforces the need to create virtual document clusters that might, for example, connect scattered newspaper articles on Garibaldi in Italy or on a particular local election.

Second, we combine entity class frequency with n-grams. Consider the phrase "he went to Washington." In our statistical model, "Washington" designates a place name 68% and personal name 10% of the time. The three-gram "he went to" precedes place names 78% of the time and personal names 5% of the time. When we combine these two sets of figures, however, we will conclude that Washington is probably a place name in this context.

Similarly, consider the phrase "he spoke to Washington." In this case, 33% of the three-grams "he spoke to" precede personal names but we have no instances of these three words preceding a place name. We assign a small portion of the probability space to unattested patterns and thus do not assign 0 to the probability of "Washington" as place name, but "Washington" as a personal name is clearly more probable than as a place name in this phrase.

Students of historical and journalistic writing will know that places – especially capital cities – often act like people: "Washington responded to London," "Athens confronted Sparta." Conversely, people can act as markers for their location. When we considered

using the pattern "between A and B" to mine place names, we found enough instances of "Lee organized the defenses between McClellan and Richmond" to set this rule aside.

4.5 Identification of entities

Once we have completed the semantic classification of all entities and decided that an entity describes a person, place or organization, we still need to identify individual members of that class. The identification problem affects any named entity that has been cited allusively. For example, "the 5th of March" challenges us to identify the year, or, alternatively, a recurring event on a given date (e.g., "the fourth of July") does as well. "The Herald" is probably a newspaper, but a search of Rowell's *American Newspaper Directory* locates almost two hundred papers with the name "Herald". We may decide that "fifth division" designates a military formation, but we still need to determine to which corps or army this division belongs. Human readers are extremely good at determining the answers to these questions on the basis of general context. For now, we have set these larger identification problems aside and concentrated specifically on identifying places and people.

4.5.1 Identifying places

At this stage of our work, we concentrate on identifying places on the scale of a town, mountain, state or other location with an entry in a major gazetteer such as the TGN. The Alexandria Digital Library (ADL) Gazetteer contains much richer coverage of smaller scale locations such as housing developments, public parks and other geographic features, but these reflect contemporary development and were not judged useful for our analysis of historical documents [67]. We classify phrases such as "the Square" and "the Wilderness" as reference strings rather than place names (although the latter example illustrates the problems inherent in this approach). We have set aside this major class of problems for later analysis and currently look for places that either occur in the TGN or have some recognizable geographic head (e.g., "Smuggler's Gulch").

Setting aside reference strings has a small but significant impact upon our results. In one Civil War book with many references to rivers and other geographic features, we find 50 unique terms that could be classified as place names. These account for 328 total place names versus 6284 strings actually classified as place names: we thus focus upon 95% of potential place names. Of the 328 places designated by reference strings, however, the most frequent three ("the Mississippi" 46 instances, "the Tennessee" 44, and "the Potomac" 43) account for 40.5% (133/328) of the total. The remaining names include generic phrases such as "the Navy Yard" and "the National Cemetery" with little or no useful associated information.

Our work on this problem goes back to the 1990s, when David Smith developed a general tool to extract place names from all English texts in our collections [86]. Since our classical texts have English translations, this provided a general solution. Analysis of

automated place name extraction revealed that place names in some cultures were easier to locate. Greek and Roman names are easy to extract. First, cities named after Alexander are Alexandrias – not Alexanders. This makes the task of semantic classification much easier. Second, while there are a number of Alexandrias, Greco-Roman names are much less ambiguous: more Greco-Roman place names are unique and those that are ambiguous have fewer possibilities. Further research also demonstrated that European place names were a bit more ambiguous and harder to find than Greco-Roman.

American place names are, however, the most challenging. In addition to the difficult challenge of semantic classification, American settlers used the same names over and over again. A search for Centerville, Virginia, in the TGN, for example, lists nine places; a search for Lebanon finds 53 in the United States and four in Tennessee alone. Place names also change over time: the 1855 edition of *Harper's Statistical Gazetteer of the World* lists 163 Washingtons but that number declines to 92 in the TGN. Thus, more than 70 Washingtons of the mid nineteenth century – 43% of the total – have either vanished or changed their names. The challenge of place name identification was a major reason why we chose to study nineteenth century American texts. This issue has been explored by a variety of researchers [34, 35, 47, 69, 1, 71, 50, 87, 48] and we continue to learn from advances in this field.

We extract as many place names as possible early on in the analysis. In a few cases, we directly attach place ids to unqualified strings. While some odd phrases such as "Island number 10" can be extracted by rules, we have chosen to deal with the combination of name and number in "Bermuda Hundred," for example, by putting this phrase in a lookup table.

The system also spends substantial effort scanning for combinations such as "Cambridge, Mass.," and "Rome, Italy." While not all the results of such combinations are unique, this process nevertheless attaches unique TGN identifiers for 349,175 references to 30,456 unique places in the corpus. Over 9,336 (2.6%) of the "PLACE,PLACE" combinations can have two or more TGN identifiers.

Our gazetteer at the moment is two staged. While we default to the TGN, we also have a growing authority list of 5,250 "PLACE, PLACE" combinations that are listed in mid-nineteenth century sources. We began this list by analyzing 3,200 places from the index of the *Official Records of the Civil War Atlas* (OR). We identified unambiguous place names, looked up their coordinates in the TGN and then induced the minimum coordinates for the 800 maps. This process revealed a number of problems. First, this index does not distinguish between ambiguous place names. Thus, while the OR maps may plots several different Centerville, Virginias, the index lumps all Centerville, Virginias into a single entry. We also discovered unnerving examples of historical change. The TGN lists only a single Berlin, MD, (in Worcester county) as does the OR Atlas. When we compared the TGN coordinates to the coordinates of the OR map on which Berlin, MD occurred, we found that the TGN coordinates were 50 miles outside the scope of the map. It turns out that the Berlin, MD of 1860 changed its name and by 1870 another Maryland town had adopted

the name Berlin.

To supplement the OR Atlas and other manual ad hoc additions to our authority list, we systematically analyzed the 1,320 counties and 1,971 cities and towns in the US and Canada listed in Rowell's *American Newspaper Directory* as supporting newspapers. Whereas the OR Atlas documented engagements that usually occurred in the countryside, this list of cities with newspapers gave us a solid survey of economically significant towns in the nineteenth century. This directory also includes valuable demographic information such as population and relative location. In 1870, Cambridge, AL, had a population of 5,000 and was about 25 miles from Selma, while Cambridge, MA, had 36,000 inhabitants and was near Boston. Such data lay the foundation for more effective methods of distinguishing one Cambridge from another.

We thus initially check name combinations against the 5,250 entries in our smaller authority list. The combination "Centerville, Va.," will therefore return only the four Centerville, Virginias, that we identified on mid-nineteenth century maps, rather than all nine returned by the TGN. A reference to "Berlin, MD" will return not only the city in Worcester County but the current, Brunswick, MD, which previously bore the name Berlin. We only turn to the TGN if we fail to find an attested PLACE, PLACE combination.

PLACE, PLACE combinations will clearly generate errors in cases where names have shifted or the text refers to a place name found in the authority list but the actual instance of that place does not exist in the list (e.g., a fifth Virginian Centerville other than the four Virginian Centervilles in the list). Such errors are difficult to find and in some cases impractical to find. If we had not possessed a very precise geographic context, we would not easily have realized that our maps designated a Berlin, MD that differed from the Berlin, MD, of the TGN. Nevertheless, these processes seem to provide generally reliable data.

More problematic than the PLACE, PLACE combinations are place names of the form NAME, PLACETYPE (e.g., Mississippi + River, Blue + Hills). If we find a unique match for a name in the TGN, we currently assume that the TGN contains the correct location for this word. Normally, the problem with the TGN is that it contains too many low frequency names that are irrelevant. Nevertheless, we find instances where the unique key in the TGN is, in fact, not the location meant in the text. Although "Milford Station" occurs more than thirty times in our collection, the TGN knows of only a single Milford Station, which happens to be in Nova Scotia. Our texts describe a Milford Station in Virginia. Additionally, some rules allow us to collect state and country names without picking up false positives such as "Virginia Ham" or "Virginia Smith."

Once we have assembled our training set of 350,000 places (e.g., "Washington, DC") and categorized names as probable place names (e.g., "Washington" in a given context is a place), we try to associate unmarked place names with unique identifiers. At the moment, we combine two features in determining which "Springfield" or which "Washington" is meant: relative frequency of identified locations in the corpus and analysis of other recently identified places in the individual document.

The relative frequency is fairly straightforward: "Washington, Beaufort County, North Carolina" shows up 34 times in our model versus 19,470 recognizable references to "Washington, DC." If these figures reflect the relative frequency of these two locations among unmarked references to "Washington", then "Washington, DC", is more than 570 times as likely to be the correct reference as "Washington, Beaufort, NC". In fact, preliminary statistics may underrepresent dominant names, since writers seem more likely to fully specify less common places (e.g., "Washington, NC") than dominant ones ("Washington, DC").

4.5.2 Identification of particular people

In our collection, we have found there are far more people than there are places. Thus, while a name such as "Jackson" might apply to 80 locations in the TGN, the upper bound for the number of people named Jackson is hard to quantify. The problem of personal name disambiguation has been extensively studied. Of particular importance for our work is the issue of coreference resolution, or determining whether different expressions in a single text or across a document collection refer to the same entity. A variety of approaches to solve this problem have been explored [53, 95, 28, 49, 96, 33]. The following figures reflect some of our experience with the variables that need to be considered in identifying entities in our historical collections.

4.5.3 Ambiguity of surnames

The more material a document covers, the greater the chance of ambiguous names. If we analyze the personal names in the print index to the four volume work *Battles and Leaders of the Civil War* (BLCW), we find that 74% of the unique surnames within the four volume series describe only one person. The ambiguous surnames are, not surprisingly, disproportionately common and only 43% of the 14,195 surname mentions are unambiguous: i.e., 57% of the time, if we just looked at the surname, we would not be able to uniquely identify the individual.

The bigger the document, the more likely we are to find the same name applied to more than one person. In the above table, we present the results if we look at the four individual volumes of the *BLCW* and only consider the names in each separate volume. Thus if "Albert Sidney Johnston" were only in volume 1 and "Joseph Johnston" were only in volume 2, the name Johnston would be unambiguous in each volume. By this measure, approximately 83.4% of the unique surnames and 62.6% of the name mentions are unambiguous.

If we further divide individual volumes and look at chunks of text, ambiguity naturally declines. The average 10 page chunk, for example, contains 32 unique surnames and 53 surname mentions which are 97.4% and 94.2% unambiguous. The above automated metric does not tell us how often in a ten page chunk we can identify a floating surname (e.g.,

Table 2: Ambiguity of surnames as document size decreases

work	total surnames	unambiguous	total surname mentions	unambiguous
blcw1-4	2,846	74.2%	14,195	43.5%
blcw1	573	82.5%	2,149	59.4%
blcw2	638	86.0%	3,210	69.7%
blcw3	787	83.4%	4,348	61.9%
blcw4	848	82%	4,488	59.4%
avg	711.5	83.4%	3549	62.6%
100 page chunks	173	92.6%	507	82.5%
10 page chunks	32	97.4%	53	94.2%
1 page chunks	5.7	99.1%	6.2	98.1%

"I saw Johnson") with the last specified surname in that ten page chunk (e.g., if we saw "Albert Sidney Johnson," we could assume that "I saw Johnson" refers to Albert Sidney), but we assume at this stage that such nearby mentions are helpful.

4.5.4 Frequency of forenames/initial and of recent mentions

We depend upon two primary methods to identify particular names. In the best case, the author provides us with forenames or, at least, initials. Such a specified full name allows us to identify nearby names in the text without explicit forenames or initials with reasonable accuracy. The more names either directly specified or near to directly specified names, the fewer names that need to be identified with other (and generally less reliable) methods.

When we studied our corpus, we were shocked by how widely the frequency of specified surnames and nearby mentions varied. We were not surprised to find that some literary works may provide no references to formal names, for example, they only rarely spell out full names such as "John Smith" or "Jane E. Woods". At the other extreme, in reference works that list people (e.g., regimental rosters, indices) more than 98% of all listed names may contain initials. *Harper's Encyclopaedia of United States History* published in 1902 reflects the careful editing of an established publishing house: specific names and nearby mentions account for 83.6% of all surnames. The median figure in our entire collection is an even 50% (explicit names + nearby mentions). What surprised us was the variance of name mentions within books of the same general type. In Ernest Crosby's 1905 biography of William Lloyd Garrison entitled *Garrison the non-resistant*, the figure is only 13%, but that figure rises to 77.6% in the third volume of *William Lloyd Garrison, 1805-1879; the story of his life told by his children.* William Boynton's *Sherman's Historical Raid* comes in at 67%, but the figure in Edward Porter Alexander's *Military Memoirs of a Confederate*

is only 17.5%.

A related issue is that the same name can also describe many individual people. There have been many John Browns, and even relatively uncommon names can often refer to different people. In many cases we can unambiguously connect initials to a fully expressed forename, such as linking "J. Johnston" to "Joseph Johnston." In a small percentage of instances, however, we do find ambiguities where a mention of "J. Johnston" is found in a document that includes references to both a "Joseph Johnston" and a "James Johnston." On the average, approximately 1.5% of the initials are ambiguous within any given work. For now, we choose the nearest match, if "J." is closer to a reference to a "James" than a "Joseph," then we resolve "J." to "James."

4.5.5 Identifying names

The current system considers the following features in its identification of potential names:

1. Analyze abbreviations. Convert standard abbreviations to full names: e.g., Jos. becomes Joseph, Jas. becomes James.

2. Compare rolenames. Within a given document, we assume that Mr. Smith and General Smith are not the same person. Likewise, we assume that Lieutenant Smith and General Smith are different because of the difference in rank. In some contexts, this causes problems, such as when a biography of Grant describes him as both Lieut. and General U. S. Grant at various points in his life. Colonel Smith and General Smith are assumed to be the same person because many colonels do reappear as generals. These rolename matching rules need to be studied more carefully but provide a useful heuristic.

3. Match floating surnames to nearby specified surnames. If a surname has no initials or forenames, we match it to the nearest comparable name within a given range. We have experimented with set windows (e.g. 10 pages) and structural windows (e.g., within the same chapter). Thus, Johnston probably becomes "Albert Sidney" in a chapter on Shiloh, but "Joseph" in a chapter on the Battle of Atlanta. This algorithm has weaknesses: e.g., "John C. Smith, who was survived by his son Peter T. Smith, was ... Smith was a famous." In this case, all subsequent Smiths become "Peter T." rather than "John C."

4. Match unmatched floating surnames to the most common surname. Many texts allude to figures that are, within the context, thought to be obvious. That context could be an individual document (e.g., a biography of William Lloyd Garrison where unmarked mentions of "Garrison" designate William Lloyd) or a corpus (e.g., a Civil War memoir where unmarked mentions of "Grant" designate Ulysses S.) Very common names present a particular challenge, since many texts may spell out dominant

names less often than uncommon ones. Thus while Robert E. Lee may be the default "Lee" mentioned most often in a text, some Civil War narratives may for that reason mention Fitzhugh Lee by name more often than Robert E. Lee. Automatically eliciting the most important default names can therefore be quite difficult, with one single error resulting in dozens or hundreds of mismatches within a text. Because of the difficulty of automatically identifying the most important names in a text, we did not implement this rule in the 1.0 release.

5. Unify initials and forenames. With the current system, for example, "U. S. Grant" is matched against "Ulysses Simpson Grant." When such matches exist in a given document, they are unambiguous 98.5% of the time. There will be instances, of course, where "J. Smith" refers to "John Smith" but the document only refers to a "James Smith." Within most single documents, such hidden collisions are unlikely to be common. The problem grows greater, however, as we unify hundreds of thousands and millions of names across documents.

4.6 Validating XML against TEI P.5 DTD

Information extraction scans data and outputs results. The Perseus Digital Library system has exploited this technique since the late 1990s by creating separate index files of dates and places automatically identified in the original text. In our current system, we add annotations to the TEI source files. When markup becomes this dense, the XML can become hard to read and the tagging confusing. Sophisticated schemas have emerged to manage annotations that become too numerous and too heterogeneous to maintain in a single file, with "standoff markup" stored in separate files. We currently treat markup for people, places, dates and other named entities as core data that belongs in the base XML source file. While such files can become complex, they are self-contained. Individuals can then correct and extend the automated tagging with standard XML editors.

Consequently we need to be careful not to overload the source files with too many tags. We also want to produce files that validate when checked against the TEI DTD. At present, we do find problems with the TEI DATELINE, which we use as often as possible to capture the place from which a letter or other document originates. The DATELINE element has very strict limitations on the types of entities it can contain. If the automated tagging system places personal names, raw numbers or reference strings in a DATELINE element, the documents do not validate. For now, we correct these errors by hand, using the DATELINE tags as a check to identify problems that may arise when we tweak the named entity identification routines. The current collection contains more than 18,000 DATELINE elements, enough to provide a useful measure to catch anomalies such as"Boston" suddenly being classified as a person rather than a place.

If we were to expand the corpus substantially, we would automatically convert problematic tags into unoffending tags with labels to invite subsequent correction if and when

human editors wished to improve on the automated markup.

5 Evaluation

5.1 Variability of results

Evaluation results measure the performance of a particular system on a particular set of documents. While the results can be useful, it is important to understand their limitations. To begin with, performance of named entity identification depends upon a range of variables including the types of document genres found within the collection being analyzed. Some types of documents follow structural patterns that may aid automatic analysis. Much, if not most, current research on language technologies, for example, concentrates on recent news articles from major publications such as the *Wall Street Journal*, *der Spiegel* or the Chinese news agency Xin Hua. Professionally authored and edited news articles, however, have well-defined formulas to set the context. In a national publication, the dateline "Philadelphia," if unspecified, will always describe Philadelphia, PA. In most cases, even common names are glossed when first cited ("President George W. Bush"). News publications are designed to provide the context which automated systems need even more than many human readers. The same tool that performs well at extracting dates from homogeneously structured news articles may perform much more poorly when applied to longer documents that are less careful to contextualize the people and places to which they refer.

The content of documents also affects system performance. A Civil War memoir may not carefully mark which "Shiloh" or "Grant" is meant, but we may be able to infer the right reference from other more meticulous documents. On the other hand, a scrupulously edited document that introduces many people and places that the system has never seen may be much harder to process.

The combination of new structural conventions and new content can seriously degrade system performance. The works of Whittier included in this collection present an interesting challenge. His poetry often capitalizes many common nouns in the middle of a sentence. The system therefore interprets "Frost" in "Of Frost, the early comer" as a surname. At the same time, the field of reference is very different in Whittier than in other documents within the collection. References to "Eden," "Ida," or "Plato" are cultural referents that point to a general cultural background that plays a much smaller (though noticeable) role in many of the documents in our collection.

5.2 Precision, recall and the combined score (F-Measure)

Computational linguistics spends substantial amounts of time developing methods to measure how well a particular technique works. Identifying personal and place names lends itself, however, to some of the most well-established evaluation methods typically used in information retrieval: precision, recall and the F-measure [46, 7].

1. Precision. This metric measures how many returned values in a set of results are correct. It also measures how many were false positives. If we search for "Washington" as a place and the system returns 8 instances, of which 6 are places and 2 are people, then 6 of the 8 returned instances would be correct and our precision would be 75% (6/8).

2. Recall. This metric measures how many instances of a selected entity were missed or not found in an entire corpus. Thus, if the system located 6 references to places named Washington but the collection actually contains 10 references to Washington as place, then we have missed 4. The recall would, in this case, be 60% (6/10).

3. F-measure. It is often useful to provide a single measure for the success of the system as a whole. In terms of precision and recall, this combined score is historically called an F-measure. There are various formulas used to obtain this measure [74]. One popular method combines precision and recall in the following formula: (2 * recall * precision) / (recall + precision). In the example above, the F-measure would be 67% [(2*.60*.75)/(.60+.75)].

Systems can emphasize precision or recall, often minimizing one to emphasize the other. Any system can achieve perfect recall by returning everything it sees, for example by assuming every "Washington" refers to a place. Conversely, we can usually minimize false positives by only returning those instances in which the system has confidence: e.g., return phrases such as "Washington, D.C.," or "city of Washington," but not "he heard from Washington." We may find that we wish to maximize one or the other depending upon our changing needs.

5.3 Collecting evaluation data

It is not feasible to examine all 1.5 million personal names and 1 million place names automatically tagged in the test corpus. We therefore depend upon evaluating subsets of the data to generate reasonable approximations of performance as a whole. To this end, we performed two types of evaluation, one manual and one automated. Time constrained the amount of manual evaluation we could perform. Automated evaluation was based on comparing pre-existing book indices. Both methods have their limits, and our results, given the size and heterogeneity of the collection, are preliminary. They nevertheless seem to provide a reasonable model for how the system performs and what progress can be made.

In the manual evaluation, we corrected subsets drawn from across the collection. We chose one page from each of the three hundred plus documents in the collection. This approach is not wholly unbiased since a book with 200 pages is better represented than a book with 500 pages. Nevertheless, we felt that this gave us a reasonable model of the collection as a whole. The larger books also tended to be reference works that are, on the whole, more regular in form and easier to analyze, so the easier cases were slightly underrepresented.

5.4 Semantic classification

To evaluate the results of our semantic classification, we randomly selected one page from each of the three hundred books in the collection and analyzed the results. Although we are not formally presenting our work with extraction of organizations, preliminary analysis was encouraging: of 1112 strings classified as organizations, only 11 were erroneous, thus yielding a precision of 99%. This was not surprising, since organizations tend to have multi-word names that are more distinctive. The results for the tagging of person and place names are summarized in the following table.

Table 3: Evaluation of results for personal and place names

Category	total identified	errors	precision	total missed	recall	F-measure
surnames	3202	150	95.3%	65	98.0%	96.6%
place names	2627	159	93.9%	43	98.4%	96.1%

In a second smaller scale examination we found 283 correctly labeled and 21 incorrectly labeled surnames (precision of 93%) as well as 288 correctly labeled and 8 incorrectly labeled place names (precision of 97%).

All systems make strategic decisions as to how their default behavior balances precision and recall. We can maximize recall if we only return hits for which we are confident – but we may miss many important pieces of evidence in the process. We can always achieve a recall of 100% – if we return every record that we look at and filter nothing out. Our system clearly emphasizes recall over precision.

A closer look at the results suggests that the major challenge is to identify a fuller set of competing classifications. For example, references to "Easter" are typically classified as a surname, and properly so, given what the system knows. Easter does sometimes appear as a surname and the system is not (yet) considering holidays as a potential class. We could and will add this as we already use the TEI OCCASION tag to mark hundreds of "Christmas," "New Year's," and "Easter" references. The system has not, however, utilized this knowledge in classifying personal and place names.

If we consider our success in distinguishing known classes, the results are encouraging. The use of "Washington" rather than "Washingtonia," "Washingtonville," etc. makes distinguishing personal and place names difficult. Of the 150 erroneously retrieved surnames, only 19 were really place names: thus, our surname precision with respect to place names was 99.38%. Conversely, only 30 of the incorrectly classified place names were in fact personal names. The place name precision with respect to surnames was 98.7%.

5.5 Identification of particular places

One major question that we examined was: of those names that we correctly classify as place names, how often can we identify the particular place? This process yields two distinct types of failure. First, this question depends upon the comprehensiveness of our gazetteers. The TGN and the ADL Gazetteer contain millions of names, but historical documents often use local designations that occur many times even within the same state. "Bull Run" may seem distinctive, but a query to our copy of the TGN reveals 86 places under the name "Bull Run," including 22 in Virginia and West Virginia. Such heavily ambiguous names may defy automated resolution – and, indeed, we often find "Bull Runs" or other names that are not even in the already crowded gazetteer lists. Second, names such as "Smith's cross-roads," "Brown's tavern," "Castle Thunder," "Kettle Bottom Shoals" and similarly semi-formal place designations are common in our collection but either no longer exist or have in general been replaced with more modern place designations (e.g., street name and number). Overall, within our corpus we have been able to assign TGN identifiers to 88.6% of automatically tagged place names in the corpus. Many of the names misidentified as place names fall into the remaining 11.4% – many, but not all. The system occasionally has identified "Public" as "Public, Kansas" (tgn,2040930) or "Christmas" as "Christmas, Florida" (tgn,2018890).

Of those names that we have identified as place names and which we have identified as specific places, we need to estimate our success rate. We have conducted two evaluations, one manual, the other automated and drawing upon the machine readable form of a print index.

In the manual survey, we chose starting points at random in 30 collection documents and analyzed the next 10 place names. We found 237 properly classified place names with identifiers. Of these, 41 identified the wrong location, thus yielding a success rate of 85%.

To complement this manual survey of the collection as a whole, we used the substantial print index to the four volume *Battles and Leaders of the Civil War*(BLCW) as the basis for the automated evaluation. We chose this set to study performance on documents that matched a core focus of the overall collection and that would reflect the upper bound of current performance.

It was soon clear that, extensive as this index may be, it was much more selective than the named entity system. Where the print index listed 3935 "placename, page" matches where we had a TGN identifier for the place, the named entity system identified 24,370 places. If we filter out those high precision places (e.g., states and countries with distinctive names such as "Massachusetts" or "France"), we still have 21,118 places – more than five times as many as in the print index. Since the print index covers a relatively small subset of places mentioned, it cannot provide immediate evidence about precision. The majority of correctly identified place names in the text would appear as false positives if compared against the print index. The print index does, however, provide enough data that we can estimate the recall of our system. For example, if we find X% of the places in the index,

then that X% gives us some sense of what percentage of place names we are correctly identifying out of the whole. The print index is biased towards place names that human authors judged significant, so the resulting measures are interesting in themselves.

Our basic principle was straightforward. If the index stated that *BLCW* vol. 4 page 764 mentioned Abbeville, SC, then we could see if our system also found Abbeville, SC, on that page. Even this question, however, divided into three stages.

First, we checked each page in the text to see whether the index entry explicitly appeared. An index might cite a range of pages as relevant to a particular place but that place name might not explicitly appear in the text on all those pages. Since we are evaluating named entity identification rather than information retrieval, we set those pages aside. At the same time, minor inconsistencies appear (e.g., "Taylor's mill" in the index vs. "Taylors mill" in the text). The remaining placename/page pairs numbered 2,798. If the index stated that "Abbeville" appeared as a place on *BLCW* vol. 4, p. 764 and we found the string "Abbeville" on that page, we used this place/page pair.

Second, we checked classification accuracy, to see how often we found the appropriate string marked as a place name. If the index stated that Abbeville, SC, was on *BLCW* vol. 4 page 764, we checked to see if Abbeville appeared on that page tagged as a place name. In at least one case, this test will provide false data. If a page discusses "Washington" as both a person and a place (e.g., it covers how the capital was named after the first president), this test will return a successful result regardless of how well we separated out "Washingtons" as a place from the "Washingtons" as a person so long as it identified at least one "Washington" as a place. Such circumstances are theoretically possible and do occur in our collection (which contains encyclopedia entries on Washington, DC), but we saw no such problems in the *BLCW* and assumed that the effects of such noise would be marginal at worst.

Third, we examined how often we were able to link the correctly classified place names to the correct location. If we found that we had indeed identified an "Abbeville" on page 764 of vol. 4, we checked to see if we had assigned to it the same TGN identification number (tgn,2095219) as the "Abbeville" in the index.

Table 4: Comparison with place names in the print index

volume	all	found	recall	context		freq.		both		overall
blcw01	460	435	94.57%	397	91.26%	404	92.87%	416	95.63%	90.43%
blcw02	614	584	95.11%	524	89.73%	567	97.09%	542	92.81%	88.27%
blcw03	968	926	95.66%	842	90.93%	876	94.60%	860	92.87%	88.84%
blcw04	891	862	96.75%	799	92.69%	831	96.40%	821	95.24%	92.14%
totals	2933	2807	95.70%	2562	91.27%	2678	95.40%	2639	94.01%	89.98%

Table 4 shows all placenames listed in the index, the number actually found and the recall. It then shows the classification accuracy achieved under three conditions: (1) simply using the context (e.g., if you just saw Boston, then assume Cambridge is in Massachusetts); (2) simply using frequency (if Cambridge, Ma., is the most common Cambridge, always assume that any Cambridge is Cambridge, MA.); (3) a combination of context and frequency. The final figure provides an overall performance measure: of any 100 place names in the text, we correctly identify 90.

Notice that the recall for the *BLCW* collection listed in Table 4 is lower than that calculated for the collection as a whole. This turns out to reflect errors in the tagging of the index. "Stone Bridge," for example, shows up as an entry associated with the battle of Bull Run, but this location does not have an entry in the TGN. The TGN does, however, list three other places named Stone Bridge, and the automated system assumes that one of these must the intended place. The named entity system scanning the text notes the article in "the Stone Bridge" and classifies this as a reference string rather than as a formal place name.

The place name identification for volumes blcw02 and blcw03 is substantially lower than for the other two volumes. Several repeated errors draw down the overall score. The largest problem was, in fact, an error in the tagging of the human index: "Cemetery Ridge" – the location of terrible fighting at Gettysburg — does not appear in the TGN but the TGN lists five other places (all typed as "ridge"). The system assumed that one of these had to be the valid identifier and chose the Cemetery Ridge in Mississippi as the most likely. When scanning the "Cemetery Ridges" in the text itself, the contextual clues were different and the system generated different hypotheses, resulting in 18 errors. If we correct this error and remove "Cemetery Ridge", the identification accuracy becomes 92.8%. More reasonably, there were 11 mismatches where the system found the wrong Bridgeport (58 entries in the TGN). Similarly there were six times when the system identified Richmond, KY, as Richmond, VA. While correcting Bridgeport would bring the accuracy up to 94%, correcting Richmond would bring its accuracy up to 94.8%. Thus, a small number of repeated error types account for much of the error space. Such a pattern is typical and suggests that, if we can assess the most problematic automated decisions, relatively modest human efforts could lead to substantial improvements.

The most interesting and unexpected result in the above table is the fact that accuracy is slightly worse when we use both context and frequency than when we just use frequency. This probably reflects the fact that the *BLCW* volumes reflect major trends in the collection as a whole. Nevertheless, improved performance may, in fact, be counterproductive. We may generate better numbers overall if we always assign an ambiguous place to the common entry, but we would then never be able to find any other entry: e.g., if Cambridge, Ma., accounted for 75% of the Cambridges in our corpus, we would identify *any* ambiguous instance as Cambridge, England.

5.6 Identification of particular people

Unlike the TGN for place names, we currently do not have any knowledge sources that serve as anything remotely like a comprehensive gazetteer of people cited in our collection. While we have associated one thousand of the most common names in our collection with the Library of Congress Named Authorities List (LCNAF), the vast majority of names mentioned in our collection do not appear in the LCNAF or any other source currently available to us. Name authority files are of particular importance for personal name disambiguation but existing authority files such as the LCNAF are limited in terms of their coverage of the lesser known historical figures who populate our collection. The creators of one digital historical collection have suggested that one of the most significant challenges faced in the transcription and digitization of historical materials is the high degree of personal name variation [27].

In our current system, the resolution of personal names involves two steps. Once we determine that "Jackson," for example, is a personal name, we need to determine which Jackson is meant. Even if we are correctly able to determine that a particular "Jackson" describes "Andrew Jackson," we cannot be sure that the "Andrew Jackson" in question is the nineteenth-century president or some other individual of the same name. For now, we set aside this problem and concentrate on the task of determining the full name in a given context for floating surnames. In most individual documents, full names will be ambiguous, but the problem of identical names for different people grows as we work with larger and more comprehensive corpora.

Name identification thus becomes the problem of connecting partial names such as "Brown" or "J. Brown" to fuller names such as "John Brown" when they exist in the corpus as a whole. We do not at this point distinguish John Brown the Abolitionist (Brown, John, 1800-1859), John Brown the eighteenth-century minister in Massachusetts (Brown, John, 1696-1742), John Brown the author of books on the puritans (Brown, John, 1830-1922) or any of the other 54 entries that a search for "John Brown" elicits from the Tufts University Library Catalog.

We performed two evaluations of this function: a manual survey of the collection as a whole and a manual survey of a particular text. In the first manual evaluation we selected pages at random in the collection. Of 288 terms correctly classified as surnames, 213 names were correctly and 65 incorrectly identified: a success rate of 76.6%.

The second manual evaluation concentrated on names from the *BLCW* and was designed to measure success in a document that we felt would let us see optimal performance. First, *BLCW* contains thousands of names and an above average percentage (71.1%) of specific names ("A. S. Johnston") and floating names near specific names (e.g., "Johnston" near "A. S. Johnston"). Second, we have developed heuristics to factor rolenames into the process, and this index was particularly well suited to a collection with many military titles. For example, Mr. Sumner and General Sumner, Lieut. Smith and General Smith are probably not the same person in the same document. We analyzed 235 correctly classified

surnames and found that 212 of these names (90.2%) were correctly identified.

We have experimented with using the print-derived indices of the *BLCW* to augment our evaluation but found that this seemed to provide less precise results than with place names. We discovered that in those cases where the text specifies a name and that name also appears in the index, the two sources agree on the forenames only 95% of the time. In some cases, the naming convention varies: we match "R. E. Lee," "Robert E. Lee," "R. Edward Lee" and "Robert Edward Lee" as equivalents but not "R. Lee." In other cases, there seem to be errors in the index. While we can address the 5% noise ratio, this will require considerable work with the index.

6 Future work

We divide our on-going work into five broad categories. "Additional content" lists areas where we hope to expand our collections. "Features under consideration" describes particular digital library services that we may implement in the future. "Work in named entity identification" discusses research topics that may not directly affect the structure of the digital library but will improve the quality of the results. "Data sources to support named entity identification" examines how our current data sources could be better mined or structured. Finally, "Larger research issues" builds on more basic results to explore larger issues.

6.1 Additional content

Besides the American collection, we are applying the named entity system to other materials as well. We are also looking for groups who have content that they wish to analyze and with whom we could collaborate. At present, we are looking at the following collections:

1. Civil War issues of the *Richmond Times Dispatch* and the *Liberator*. We are collaborating with the University of Richmond on a project funded by the Institute of Museum and Library Services (IMLS) to digitize two Civil War era newspapers. The *Richmond Times Dispatch* reflects the high end of data entry and markup, with double keying maximizing transcription accuracy and careful markup of individual articles. The *Liberator* has been entered with minimal markup, and its text is based on OCR output, partially corrected by a data entry contractor. These documents allow us to approach a variety of problems, the most significant of which is how to automatically analyze the content of articles that are very short and very diverse in content. In these newspapers, an article on a local election may follow an allusive description of Garibaldi's activities in Italy. Since the source images are often hard to read, even the carefully entered texts contain more errors than in materials entered from clearer print. The newspapers contain thousands of personal, place and organi-

zational names, including detailed addresses (in a range of abbreviated formats), as well as advertisements, prices, and other commercial data.

2. Greco-Roman collections. We have long extracted place names, dates, citations, and morphological data from Greco-Roman materials. These raise interesting challenges both related to language and to cultural habits of naming (on which, see below).

3. The History and Topography of London. In 1999-2000, we helped begin a digital archive, based on one of the Tufts special collections, on the history and topography of London. Subsequent support from the IMLS allowed the Tufts library to extend this collection, which now contains 10 million words of text. This collection is of interest because it combines great similarities of language and structure with differing cultural emphases (e.g., words such as "Duke" and "Lord" are much more likely to be titles than surnames) and frames of reference.

4. American Memory Collections. The Library of Congress has created a number of public domain, highly structured collections covering US history and culture. We have already integrated several of these in the earlier Perseus Digital Library. The American Memory collections are in a modified form of the TEI Guidelines and cannot, as currently structured, easily accept the encoding that we can produce. These collections thus not only provide an opportunity to work with different chunks of content illustrating nineteenth-century American English but also raise interesting problems of portability, as we consider whether to transform them into a more standard form or extend their customized DTDs.

5. Early Modern Documents. We have developed a sample collection of early modern documents in English, Italian and Latin. These allow us to study issues raised by non-standard spelling and pre-modern systems of reference as well as multi-linguality.

6. Additional internally produced content. We have assembled several hundred other books on topics covered in the American collection. These will be entered both as image books (to study the problems of analyzing lightly structured documents typical of large collections) and, where the contents are of particularly strategic value, in more carefully structured formats. Thus, we have a number of nineteenth-century reference works including geographical gazetteers, newspaper directories, city directories and other semi-structured data sources, all of which we can use to provide background for human readers and to improve the performance of automated text analysis. The importance of this issue of providing contextualized information to support the users of digital libraries has recently become an issue of increasing importance as collections of digital materials are vastly proliferating and users need more advanced tools and technologies to help them wade through vast and disparate collections [93, 44].

6.2 Features under consideration

This list includes a number of potential services we hope to explore or extend.

1. Indications of which automated identifications are more and less certain.

2. Tools whereby users can improve data entry and markup.

3. Storing scans of pages and illustrations in the Tufts University long-term institutional repository (with better tools for viewing large images).

4. Ability to search for other classes of entity such as ships, bibliographic references, and organizations such as military units, or newspapers.

5. Better integration of the historical reference works, such as supporting a search that will automatically locate a photograph and an encyclopedia article about the same person. This work might also involve automatic linking, such as linking references to organizations such as military units and newspapers to contemporary reference works such as Rowell's *American Newspaper Directory* and Dyer's *Compendium of the War of the Rebellion*. The need for better automatic linking of related digital historical materials is increasingly receiving attention [90, 76] and we hope to learn from these research efforts.

6. Map locater that finds maps relevant to a given chunk of text. Since we extract place names, we can generate a geographic footprint for each chunk of text and search for geo-referenced maps that best illustrate the geographic context of the collection. As an initial dataset of historical maps, we have geo-referenced maps from the *Official Records of the Civil War Atlas* and from the 1910 *Century Cyclopedia Atlas of the World*.

7. Quote matcher. All quotes – more than 85,000 of thirty characters or more – in this collection are surrounded by TEI QUOTE tags. We can thus extract quoted text and search for its source elsewhere online. Many of these quotes also cite documents that are in our internal collection. While most of these cited documents are probably not in the Perseus American Collection, they may be available in other accessible online collections.

8. Bibliographic matcher. References to books and documents can be very hard to extract, but there are more than 50,000 instances of the TEI BIBL tag in this collection. This service raises two issues. First, the tagged bibliographic references are usually unstructured strings (e.g., <bibl>Grant's Memoirs</bibl>). Second, the key bibliographic databases such as the Library of Congress Catalog are available as web services, thus making automated lookups with fuzzy strings an exercise in interoperability. Our goal would be to associate bibliographic citations with full catalog data

to support a range of queries (e.g., what are the average publication dates of works cited in document set N? How many documents cited are available elsewhere online or in the local library?)

9. Clustering of internal documents by various features. Many users are familiar with commands that "Find documents similar to this." We can (1) cluster chunks of documents as well as overall documents and (2) cluster according not only to overall content but according to various named entities that have been tagged (e.g., find documents with similar place names and dates).

10. Generating query terms for external searching. Identify the key words and phrases that are most closely associated with Daniel Webster vs. Noah Webster, Springfield, MA, vs. Springfield, MO, etc.

6.3 Work in named entity identification

Besides the above features, we will work on improving the results of the named entity analysis. Some strategies that we will evaluate include:

1. Preliminary text clustering. When deciding whether a given "Springfield" is in MA, MO or some other location, we analyze passages that specify one or the other (e.g., "... in Springfield, Mass., we find ..."). At the moment, we use two statistical models, one based on the collection as a whole, the other on the document being analyzed. If the corpus varies, then aggregate statistics are less effective. Thus, the Civil War corpus makes frequent reference to Rome, GA. When we try to identify references to Rome, Italy, in other documents, the system will tend to identify it as Rome, GA. Therefore, we check the statistics for the individual document first: if the individual document explicitly refers to Rome, Italy, then we give preference to Rome, Italy. But if the document does not specifically mention Rome, Italy, then Rome, GA, takes precedence. Furthermore, many single volume works – including such crucial resources as encyclopedias and gazetteers –contain chunks of text on a range of heterogeneous topics (and thus provide uneven statistical models). One possible new approach we are considering is to develop statistical models at the document level, such as for each encyclopedia entry, book chapter or other default chunk.

2. Applying human authored indices. This paper describes initial work with one particular document index, but a great deal more can be done, not only with that particular resource but with indices in general. While some previous work has explored how human authored indices can be better exploited in information retrieval and extraction, there is a relatively unexplored areas [88, 54]. We could use a survey of index types, citation/entity densities, problems, strengths, etc.

3. Personal name modeling. At present, we try to match isolated last names against last names with forenames. If we see "Lincoln" on page three and "Abraham Lincoln" appeared on page two, then we assume that both references point to the same person. This approach works well for news articles but is less successful with larger documents. Biographies mention many family members with the same name, and relatives often write memoirs of people with the same name. We need to make better use of other contextual clues (e.g., inferring that nearby references to "Clay" imply that a given Webster is Daniel rather than Noah) and representation of name frequency (e.g., the probability that two Smiths vs. two Rumpelstiltzkins, ten pages apart, describe the same person). We are exploring a variety of research in this area [6].

   At the same time, some names are so dominant in a corpus that they are not fully named as often. Thus in our Civil War collection, Robert E. Lee is by far the most frequently mentioned General Lee, but "Fitzhugh Lee" shows up 65% more often than Robert E. (or Edward) Lee (1486 vs. 896). Dominant names thus raise challenges for automated data mining.

4. More naming conventions. Our current work reflects the naming conventions implicit in modern English and especially American English. Thus, we correctly analyze patterns such as "William Lloyd Garrison" but not Roman names such as "Marcus Tullius Cicero" or "Gaius Julius Caesar." Even within the same culture, individual documents have very different conventions. While "Thomas" is more likely to be a surname in narrative prose, in fiction it typically appears as a forename. Cultural referents can also be difficult to identify. Civil War documents might discuss Vienna, Va., but then suddenly compare the defense of Richmond to the defense of Vienna against the Turks. Biblical names (e.g., Isaac, Ezekiel) may describe contemporary figures or refer directly to the Bible.

5. Multi-lingual naming. So far, we have analyzed documents either composed originally in or translated into English. Language and culture are separate, if interrelated variables. European names in seventeenth-century Latin texts may be easier to track (e.g., Isaacus Newtonus vs. Marcus Tullius Cicero). While we have extensive data to analyze standard terms in Greek and Latin, our databases are drawn from lexica that provide only cursory coverage of proper nouns. Over the past several years, we have converted encyclopedias of people, places and material culture for the ancient world into knowledge sources and extracted a database of stems with which to locate proper nouns in Greek and Latin.

6.4 Data sources to support named entity identification

Access to vast data sets (e.g., "web corpora" via the Google API) allows us to begin extracting general information and patterns automatically, thus enhancing the power of general purpose and sometimes scalable algorithms [16, 14]. Nevertheless, successful language

technologies still depend upon access to high quality knowledge sources. Data mining and analysis will augment and help generate well-structured knowledge sources but will not in the short term replace them. While humanists can build software tools and even contribute to computer science research, our most effective strategy may be to learn enough about language technologies to build electronic knowledge bases to support analysis of historical languages and cultures. The following list while not exhaustive offers a number of potential sources.

1. Historical gazetteers. Massive gazetteers listing geographic names are available listing millions of locations, thus allowing us to identify possible place names. The TGN, which we have used for years, contains more than 1,000,000 names. The ADL, with more than four million names, is even more comprehensive. While many places in our historical collections do not appear in these modern gazetteers, these gazetteers are, if anything, too comprehensive. They contain many names that were not in use in the nineteenth-century – the TGN lists two counties and two cities in the United States named "Jeff Davis," with two more counties named "Jefferson Davis." Only one city appears to be named "Abraham Lincoln." The ADL Gazetteer lists 31 locations named "Jefferson Davis," 14 named "Abraham Lincoln." Equally important, the online gazetteers do not record when any of these locations received the names "Jefferson Davis" and "Abraham Lincoln." Nor do they record contemporary data about population or other factors that could help an automated system identify locations in historical documents (e.g., what was the relative population of Wilmington, DE, vs. Wilmington, NC, in 1860?)

   Aside from easily quantifiable features, we can perform a more general analysis of article contents. An article on Cambridge, MA, will probably have more references to Boston, whereas an article on Cambridge, UK, will probably contain more references to England and London. An automated classifier that sees Cambridge and Boston in context can use this collocation to identify a likely reference to Cambridge, MA. The same techniques apply to any commonly associated terms: "Bull Run" is a very common place name (our version of the TGN lists 86), but only one of these streams was the site of Civil War battles. Thus, terms such as "battle" in combination with Bull Run help us identify the right Bull Run from a very long list.

   We have collected historical gazetteers that provide contemporary information for various points of the nineteenth-century. While the ADL Gazetteer may not yet contain extensive historical information, the ADL Gazetteer Content Standard schema provides a powerful language with which to encode the information within these contemporary resources [68]. We are studying the challenges of converting full text gazetteers into such complex data structures and the impact of using such contemporary gazetteer data on named entity recognition.

2. Biographical Dictionaries, City Directories, Census Records, and other "databases"

of people. In comparison to geographic names, human names constitute an almost unbounded set. While place names do change over time, the relative pace of change is radically different. Automatically identifying a particular "John Brown" in full text is thus fundamentally more challenging than identifying the right "Springfield."

Nevertheless, better data sources can help. Traditional publications tend to cite a small number of culturally significant figures. Biographical dictionaries document these cultural preferences and provide important clues for automated analysis. The same general techniques applied to gazetteers are also applicable to articles about people. The simple length of two entries provides evidence to the relative importance and probable frequency of reference. (Defining that relationship is an interesting problem: if an article on person A is three times as long as an article on person B, are references to person A three times as common as to person B? nine times as common? 33% more common? some other predictable relationship? or is there no solid correlation?) Unfortunately, we have discovered that we can't simply use human authored indices as a point of comparison, since these indices tend to be more exhaustive in citing less common figures and less exhaustive in citing very common figures. Likewise, we can apply general textual analysis to look for disambiguating terms: e.g., "Webster" and "dictionary" suggests Noah, "Webster" and "Calhoun" suggests Daniel.

Some biographical dictionaries follow fairly consistent formats so that we can automatically extract most birth and death dates. Documents published before Abraham Lincoln was born cannot cite Abraham Lincoln. Even if we can't ferret out the precise birth and death dates, we can analyze the dates cited and use these to create a general temporal footprint for a particular person. We can thus cut down on the potential search space as we try to match eighteenth century references to Lincoln. On the other hand, many primary sources appear as parts of editions, and these editions cite a number of both works and people who lived after the source was produced. Thus, good structural markup, which allows us to separate editorial contributions from the original source, is important.

3. Homogeneous Reference Works. Newspaper directories contain structured textual data with fields customized for newspapers (e.g., circulation, publication schedule, party affiliation). Railroad directories may describe incorporation, mileage, major stations, and number of passengers. Atlases contain indices that provide evidence for place names in use at a given time, locate those places within well defined geographic spaces (e.g., see box E7 on map 128) and often contain brief chunks of information (e.g., populations, status of a place as part of a province or perhaps a state that does not exist in the twenty first century). Literary biographical works can provide well-structured information not only about the authors themselves but also about the dates and titles of their publications.

As an example, consider Dyer's *Compendium of War of the Rebellion*. It essentially provides three views of a common database on Union military units and battles in the Civil War. Its regimental histories section contains brief histories for more than 2,000 Union Civil War regiments, including incorporation dates, affiliation in larger units, participation in particular battles, and casualties. Its section on battles enumerates thousands of conflicts, locating them within places, listing casualties (which help represent the probable significance of the event), and classifying what happened (e.g., skirmish, raid, etc.). Its section on Union commands details what larger command structures were established (e.g., Army of the Potomac), their start and end dates, the regiments of which they were composed and their various commanders.

One important insight from our work on named entity recognition is that specialized classes tend to have their own naming conventions. Thus, the Tenth Massachusetts Infantry Regiment can appear as the "10th Mass. Infantry," "Mass. 10th Regiment," or the "10th Mass," etc. The "Baltimore and Maine Railroad" is also the "B. and M. R. R." or just the "B and M" at times. We find the *New York Times*, *N. Y. Times*, or just the *Times* at various instances. These naming schemes often appear in distinct contexts and offer a range of challenges. Reference works on homogeneous topics generally provide full names from which abbreviations can be derived and often provide lists of abbreviations as well (e.g., if we know N. Y . is an abbreviation of New York, then we can infer "N. Y. Times" as a possible match for "New York Times").

Specialized reference works on homogeneous topics often thus have consistent formats that can support specialized services. For example, in terms of newspapers, once we have analyzed circulation, newspaper size and party affiliation, we can trace the claimed readership of different political parties. Similarly, once we have analyzed dates and places within a work on military commands, we can, for example, create temporal-spatial visualizations, plotting location of units over time. Bibliographic citations in a literary biography can also be linked to online library or book seller catalogues.

4. Specialized dictionaries. The reference works above document particular places, people, or organizations. Specialized dictionaries describe more general classes of objects, identifying not the "10th Massachusetts Infantry Regiment" but rather defining the terms "regiment" or "battalion" as military units. Specialized reference works can describe the usage of particular authors or corpora: thus, we have lexicons on Shakespeare and the Greek poet Pindar, which document how particular words from these authors are used in specified passages. Even when these works don't specify all potential usages of every word in each passage, they document which senses are most significant for a given corpus. Conversely, topical lexica can range from brief glosses (e.g., short definitions of rhetorical terms) to elaborate multi-volume reference works (e.g, the three volume *Knight's American Mechanical Dictionary* with thousands of detailed illustrations).

Specialized dictionaries have some potential for identifying named entities. Ship names can, for example, be very tricky: is "the Ohio," for example, a river or a ship?

5. Heterogeneous reference works. Heterogeneous reference works contain articles about various classes of topics. This category thus includes general encyclopedias such as the *Encyclopedia Britannica* and even relatively specialized works such as *Harper's Encyclopaedia of United States History.* Articles in reference works of this sort may yield structural data of high quality but they need first to be classified (e.g., is this article about a person? place? organization?) It is much easier to extract information if we know that every article is a biography or description of a place. Nevertheless, such general encyclopedias can not only provide information about particular topics but can help systems compare cross-class ambiguities: e.g., comparing the articles on people named Washington vs. places named Washington that appear in the same work.

6. Individual documents contain important categories of discrete information: timelines, portraits of people and places, maps, membership/office holder lists, or specialized glossaries to name only a few. *Harper's Encyclopaedia of United States History* contains timelines for the United States and individual states that include more than 10,000 entries. Individual entries in *Knight's American Mechanical Dictionary* contain dozens of glossaries for contemporary materials, products, processes and other topics. These scattered resources need to be identified and collected. Aggregation can be tricky. Suppose, for example, we want to collect entries from multiple timelines relevant to a particular date. Since print timelines are designed to be read in sequence and they often do not repeat information from entry to entry. We will thus find cryptic entries such as "elected president" or "captured" where the person or place was named in a preceding entry. We need to normalize entries if we are to reuse them as smaller units of information.

6.5 Larger research issues

1. Structured collections as training sets for larger, less structured collections. Most large existing collections of carefully entered and tagged documents contain hundreds or thousands of thousands of digital books. The Harvard Library – one of five libraries that Google is currently scanning in – contains roughly 10 million books in 500 languages. Are small and carefully curated collections obsolete? Or do these carefully structured primary sources and reference works provide training sets with which data mining and information retrieval systems can make more effective use of collections that are four orders of magnitude larger but much less structured?

2. How much information can we extract from documents? Ontologies rapidly become complex, culturally contingent, and idiosyncratic. Nevertheless, some important re-

lations can be mined from text corpora and be productively shared [36]. Places have locations on the earth. Human beings have birth and death dates, as well as locations as they move through space. The amount of information we can extract and share is a technical challenge. Developing ontologies to share such information is a social process. Both extraction and ontology development are, however, linked, since better extraction can push us to develop more powerful shared ontologies while well defined ontologies provide structures to guide development of extraction [92]. This is a rapidly developing research area and a number of different cultural heritage ontologies such as the CIDOC CRM have been developed that we could potentially explore [58, 24].

3. How can we personalize information for particular users? The need for personalized support of digital library users has become an important research topic [73, 38]. We can already extract far more information about many documents than any one person needs at any given time. How do we represent that information to make it more useful? Strategies include visualizations (e.g., plotting a map of all places mentioned in a document) and filtering (e.g., identifying important terms that may be unfamiliar to a reader).

References

[1] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *SIGIR 04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, New York, NY, USA, 2004. ACM Press.

[2] Amittai E. Axelrod. On building a high performance gazetteer database. In *NAACL 2003 Workshop on the Analysis of Geographic References*, 2003.

[3] David Bearman and Jennifer Trent. Social terminology enhancement through vernacular engagement exploring collaborative annotation to encourage interaction with museum collections. *D-Lib Magazine*, 11(9), September 2005.

[4] Merrick Lex Berman. Gazetteer development for the china historical gis project. In *Paper presented at the Social Science History Association Annual Conference*, 2001.

[5] Kalina Bontcheva, Diana Maynard, Hamish Cunningham, and Horacio Saggion. Using human language technology for automatic annotation and indexing of digital library content. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 613–625, London, UK, 2002. Springer-Verlag.

[6] Béatrice Bouchou, Mickael Tran, and Denis Maurel. Towards an xml representation of proper names and their relationships. In *NLDB*, pages 44–55, 2005.

[7] Silveira Chaves Bruno Martins, Mario J. Silva. Challenges and resources for evaluating geographical ir. In *Proceedings of the Workshop on Geographic Information Retrieval at CIKM 2005*, 2005.

[8] George Buchanan, Sally Jo Cunningham, Ann Blandford, Jon Rimmer, and Claire Warwick. Information seeking by humanities scholars. In *ECDL*, pages 218–229, 2005.

[9] Michael K. Buckland, Fredric C. Gey, and Ray R. Larson. Going places in the catalog: Improved geographic access. final report. IMLS grant report published on the web, September 2002. Available online at http://www.ecai.org/imls2002/imls2002-final_report.pdf.

[10] Robert F. Chavez and Thomas L. Milbank. London calling: Gis, vr, and the victorian period. In *VSMM '01: Proceedings of the Seventh International Conference on Virtual Systems and Multimedia (VSMM'01)*, page 335, Washington, DC, USA, 2001. IEEE Computer Society.

[11] Timothy Chklovski and Yolanda Gil. Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 35–42, New York, NY, USA, 2005. ACM Press.

[12] Timothy Chklovski and Yolanda Gil. Towards managing knowledge collection from volunteer contributors. In *Proceedings of 2005 AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC05)*, 2005.

[13] Youngok Choi and Edie M Rasmussen. Searching for images: The analysis of users queries for image retrieval in american history. *Journal of the American Society for Information Science and Technology*, 54(6):498–511, 2003.

[14] Rudi Cilibrasi and Paul M. B. Vitanyi. Automatic meaning discovery using google, December 2004.

[15] Philipp Cimiano, Günter Ladwig, and Steffen Staab. Gimme' the context: context-driven automatic semantic annotation with c-pankow. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 332–341, New York, NY, USA, 2005. ACM Press.

[16] Philipp Cimiano and Steffen Staab. Learning by googling. *SIGKDD Explor. Newsl.*, 6(2):24–33, 2004.

[17] Charles Cole. Inducing expertise in history doctoral students via information retrieval design. *Library Quarterly*, 70:86–109, January 2000.

[18] Gregory Crane and Jeffrey A. Rydberg-Cox. New technology and new roles: the need for "corpus editors". In *Proceedings of the 5th ACM Conference on Digital Libraries*, pages 252–253, San Antonio, TX, June 2000.

[19] Gregory Crane, David A. Smith, and Clifford E. Wulfman. Building a hypertextual digital library in the humanities: A case study on London. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 426–434, Roanoke, VA, 24-28 June 2001.

[20] Gregory Crane, Clifford E. Wulfman, Lisa M. Cerrato, Anne Mahoney, Thomas L. Milbank, David Mimno, Jeffrey A. Rydberg-Cox, David A. Smith, and Christopher York. Towards a cultural heritage digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2003*, pages 75–86, Houston, TX, June 2003.

[21] Hamish Cunningham. Encyclopedia of language and linguistics, 2005.

[22] F.M.G. de Jong, H. Rode, and D. Hiemstra. Temporal language models for the disclosure of historical text. In *XVIth International Conference of the Association for History and Computing*, pages 161–168, Amsterdam, 2005. KNAW. ISBN=90-6984-456-7.

[23] Ian Densham and James Reid. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. 2003.

[24] Martin Doerr, Jane Hunter, and Carl Lagoze. Towards a core ontology for information integration. *J. Digit. Inf.*, 4(1), 2003.

[25] Wendy Duff and Catherine M. Johnson. Where is the list with all the names: Information seeking behavior of genealogists. *American Archivist*, Spring/Summer, 2003.

[26] Wendy M. Duff and Catherine A. Johnson. Accidentally found on purpose: Information-seeking behavior of historians in archives. *Library Quarterly*, 72, October 2002.

[27] Nadine P. Ellero. Panning for gold: Utility of the world wide web for metadata and authority control in special collections. *Library Resources and Technical Services*, 46(3), 2002.

[28] Michael Fleischman and Eduard Hovy. Multi-document person name resolution. In *Proceedings of the Workshop on Reference Resolution and its Applications: ACL 2004*, 2004.

[29] Michael Fleischman and Eduard H. Hovy. Fine grained classification of named entities. In *COLING*, 2002.

[30] Paolo Frasconi, Giovanni Soda, and Alessandro Vullo. Hidden markov models for text categorization in multi-page documents. *J. Intell. Inf. Syst.*, 18(2-3):195–217, 2002.

[31] Dayne Freitag. Trained named entity recognition using distributional clusters. *Proceedings of the EMNLP 2004*, 2004.

[32] Luca Gilardoni. Machine learning for the semantic web: Putting the user into the cycle. In *Machine Learning for the Semantic Web Dagstuhl Seminar*, 2005.

[33] Chung Heong Gooi and James Allan. Cross-document coreference on a large scale corpus. In *HLT-NAACL*, pages 9–16, 2004.

[34] Linda L. Hill Greg Janee, James Frew. Issues in georeferenced digital libraries. *D-Lib Magazine*, 10(5), May 2004.

[35] L. Hill and Q. Zheng. Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with developing and implementing gazetteers, 1999.

[36] Andreas Hotho. Using ontologies to improve the text clustering and classification task. In Editors: Nicholas Kushmerick, Fabio Ciravegna, and Steffen Staab AnHai Doan, Craig Knoblock, editors, *Proceedings of Dagstuhl Seminar on Machine Learning for the Semantic Web*, 2005.

[37] J. Paul Getty Research Institute. The getty thesaurus of geographic names.

[38] et. al. Jamie Callan, Alan Smeaton. *Personalisation and Recommender Systems in Digital Libraries: Joint NSF-EU Delos Working Group Report.* May 2003.

[39] and John P. Wilson, Christine S. Lam. A new method for the specification of digital gazetteers. *Cartography and Geographic Information Science*, 31(4):195–207, 2004.

[40] Aaron Krowne. Building a digital library the commons-based peer production way. *D-Lib Magazine*, 9(10), October 2003.

[41] Aaron Krowne and Martin Halbert. An initial evaluation of automated organization for digital library browsing. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 246–255, New York, NY, USA, 2005. ACM Press.

[42] Anagha Kulkarni. Unsupervised discrimination and labeling of ambiguous names. In *Proceedings of the ACL Student Research Workshop*, pages 145–150, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[43] Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 658–665, New York, NY, USA, 2004. ACM Press.

[44] Carl Lagoze, Dean B. Krafft, and Sandy Payette andSusan Jesuroga. What is a digital library anymore, anyway? beyond search and access in the nsdl. *D-Lib Magazine*, 11(11), November 2005.

[45] Lewis Lancaster. A multilingual gazetteer system for integrating spatial and cultural resources. Grant report for NSF, August 2002. Available online at http://www.ecai.org/projects/gazetteer/nsfgaz.pdf.

[46] A. Lavelli, M.E. Califf, F. Ciravegna, D. Freitag, C. Guiliano, N. Kushmerick, and L. Romano. Ie evaluation: Criticisms and recommendations.

[47] Seungwoo Lee and Gary Geunbae Lee. A bootstrapping approach for geographic named entity annotation. In *AIRS*, pages 178–189, 2004.

[48] Jochen L. Leidner. Topoynm resolution in text: Which sheffield is it? In *Proceedings of the workshop on Geographic Information Retrieval, SIGIR 2004*, 2004.

[49] Xin Li, Paul Morie, and Dan Roth. Robust reading: Identification and tracing of ambiguous names. In *HLT-NAACL*, pages 17–24, 2004.

[50] James Reid Malvina Nissim, Colin Matheson. Recognising geographical entities in scottish historical documents. In *Proceedings of the workshop on Geographic Information Retrieval, SIGIR 2004*, 2004.

[51] Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *CoNLL*, 2003.

[52] D. Maynard, K. Bontcheva, and H. Cunningham. Automatic Language-Independent Induction of Gazetteer Lists. In *Proceedings of 4th Language Resources and Evaluation Conference (LREC'04)*, 2004.

[53] Andrew McCallum and Ben Wellner. Conditional models of identity uncertainty with application to noun coreference. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 905–912. MIT Press, Cambridge, MA, 2005.

[54] Massimo Melucci. Making digital libraries effective: automatic generation of links for similarity search across hyper-textbooks. *J. Am. Soc. Inf. Sci. Technol.*, 55(5):414–430, 2004.

[55] Andrés Montoyo, Rafael Muñoz, and Elisabeth Métais, editors. *Combining Data Driven Systems for Improving Named Entity Recognition*, volume 3513 of *Lecture Notes in Computer Science*. Springer, 2005.

[56] Raymond J. Mooney and Razvan Bunescu. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10, 2005.

[57] Martin Mueller. Electronic homer. *Ariadne*, 25, 200. Available at http://www.ariadne.ac.uk/issue25/mueller/.

[58] Gabor Nagypal. Creating an application-level ontology for the complex domain of history: Mission impossible? 2004.

[59] Roberto Navigli. Supporting large-scale knowledge acquisition with structural semantic interconnections. In *Proceedings of 2005 AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC05)*, 2005.

[60] Douglas W. Oard. Language technologies for scalable digital libraries. In *Presented at the International Conference on Digital*, New Delhi, India, 2004.

[61] Stanislaw Osinski and Dawid Weiss. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54, 2005.

[62] Marius Pasca. Acquisition of categorized named entities for web search. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 137–145, New York, NY, USA, 2004. ACM Press.

[63] JF De PasQuale and JG Meunier. Categorisation techniques in computer-assisted reading and analysis of texts (carat) in the humanities. *Computers and the Humanities*, 37(1):111–118, February 2003.

[64] M. S. Patton and D. M. Services for a customizable authority linking environment. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 420–420, New York, NY, USA, 2004. ACM Press.

[65] Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. Name discrimination by clustering similar contexts. In *CICLing*, pages 226–237, 2005.

[66] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. Geographical information recognition and visualization in texts written in various languages. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 1051–1058, New York, NY, USA, 2004. ACM Press.

[67] Alexandria Digital Library Project. Alexandria digital library gazetteer.

[68] Alexandria Digital Library Project. Guide to the adl gazetteer content standard, February, 26 2004. Available online at http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm.

[69] Ross Purves and Chris Jones. Adding geographic scope to web resources. *SIGIR Forum*, 38(2):53–56, 2004.

[70] Andreas Rauber and Dieter Merkl. Text mining in the somlib digital library system: The representation of topics and genres. *Appl. Intell.*, 18(3):271–293, 2003.

[71] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 50–4, 2003.

[72] Lawrence Reeve and Hyoil Han. Survey of semantic annotation platforms. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1634–1638, New York, NY, USA, 2005. ACM Press.

[73] M. Elena Renda and Umberto Straccia. A personalized collaborative digital library environment: a model and an application. *Inf. Process. Manage.*, 41(1):5–21, 2005.

[74] Jason D. Rennie. Derivation of the f-measure. Online, February 2004. Available online at http://people.csail.mit.edu/jrennie/writing/fmeasure.pdf.

[75] Matthew Richardson and Pedro Domingos. Building large knowledge bases by mass collaboration. In *K-CAP '03: Proceedings of the international conference on Knowledge capture*, pages 129–137, New York, NY, USA, 2003. ACM Press.

[76] Massimo Riva and Vika Zafrin. Extending the text: digital editions and the hypertextual paradigm. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 205–207, New York, NY, USA, 2005. ACM Press.

[77] Benjamin Rosenfeld, Ronen Feldman, Moshe Fresko, Jonathan Schler, and Yonatan Aumann. Teg: a hybrid approach to information extraction. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 589–596, New York, NY, USA, 2004. ACM Press.

[78] Roy Rosenzweig. Can history be open source: Wikipedia and the future of the past?, 2005. unpublished, copy in author's possession.

[79] Jeffrey A. Rydberg-Cox, Anne Mahoney, and Gregory Crane. Document quality indicators and corpus editions. In *Proceedings of the First ACM + IEEE Joint Conference on Digital Libraries*, pages 435–436, Roanoke, VA, 24-28 June 2001.

[80] Satoshi Sekine. Named entity: History and future. Available online at http://cs.nyu.edu/ sekine/papers/NEsurvey200402.

[81] Satoshi Sekine and Chikashi Nobata. Definition, dictionaries and tagger for extended named entity hierarchy.

[82] Arijit Sengupta, Mehmet Dalkilic, and James Costello. Semantic thumbnails: a novel method for summarizing document collections. In *SIGDOC '04: Proceedings of the 22nd annual international conference on Design of communication*, pages 45–51, New York, NY, USA, 2004. ACM Press.

[83] Christian Siefkes. Incremental information extraction using tree-based context representations. In *CICLing*, pages 510–521, 2005.

[84] David A. Smith. Detecting and browsing events in unstructured text. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 73–80, New York, NY, USA, 2002. ACM Press.

[85] David A. Smith. Detecting events with date and place information in unstructured text. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 191–196, New York, NY, USA, 2002. ACM Press.

[86] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 127–136, London, UK, 2001. Springer-Verlag.

[87] Humphrey Southall. Defining and identifying the roles of geographic references within text: Examples from the great britain historical gis project. In *NAACL '03 Workshop on the Analysis of Geographic References*, 2003.

[88] Lyne Da Sylva and Frederic Doll. A document browsing tool: using lexical classes to convey information. In Guy Lapalme Balzs Kgl, editor, *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005, Victoria, Canada, May 9-11, 2005. Proceedings*, pages 307–318, 2005.

[89] Hiroyuki Toda and Ryoji Kataoka. A clustering method for news articles retrieval system. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 988–989, New York, NY, USA, 2005. ACM Press.

[90] Gregory Toner. Enhancing scholarship through linking electronic resources. In *Digital Resources for the Humanities 2005*, 2005.

[91] Olga Uryupina. Semi-supervised learning of geographical gazetteers from the internet. In *Proceedings of the HLT-NAACL Workshop on the Analysis of Geographic References*, Edmonton, Kanada, 2003.

[92] Michael Uschold and Michael Gruninger. Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, 33(4):58–64, 2004.

[93] Rene Witte. An integration architecture for user-centric document creation, retrieval, and analysis. 2004.

[94] Ian H. Witten, Katherine J. Don, Michael Dewsnip, and Valentin Tablan. Text mining in a digital library. *Int. J. on Digital Libraries*, 4(1):56–59, 2004.

[95] Xiaofeng Yang, Jian Su, and Lingpeng Yang. Entity-based noun phrase coreference resolution. In *CICLing*, pages 218–221, 2005.

[96] Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. coreference resolution for information extraction. 2004.

[97] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524, New York, NY, USA, 2002. ACM Press.