White Paper Report

Report ID: 98426

Application Number: HJ5001310

Project Director: Gregory Crane (gregory.crane@tufts.edu)

Institution: Tufts University

Reporting Period: 1/1/2010-3/31/2011

Report Due: 6/30/2011

Date Submitted: 8/8/2011

What did we do with a million books? Rediscovering the Greco-Ancient world and reinventing the Humanities

Bridget Almas, Alison Babeu, David Bamman, Federico Boschetti, Lisa Cerrato, Gregory Crane, Brian Fuchs, David Mimno, Bruce Robertson, David Smith

1.1 Introduction

In March 2006, we had the opportunity to publish an issue of *D-Lib Magazine*[1] on the general topic of digital library evolution where we posed the question: "What do You Do with a Million books?"(Crane 2006). The Digging into Data (DID) Program has allowed us to do just that, as we worked over the past eighteen months with more than a 1.2-million volume collection that our colleagues from at the Department of Computer Science at the University of Massachusetts at Amherst had downloaded from the Internet Archive.[2] This paper reports upon what we learned as we worked with this collection. Support from the DID Program was essential to us but we depended as well upon support, past and present, from other NEH programs as well as from the Institute of Museum and Library Services (IMLS), the National Science Foundation (NSF), the Mellon Foundation, the Cantus Foundation and Google.

Our practical goal is to explore the challenge of creating dynamic variorum editions[3], i.e., editions that allow readers to view various versions of, and major threads of discourse about, any passage in a given work. A variorum edition is a kind of super edition – an edition of editions. This is hardly a novel idea – the first volume of the *New Variorum Shakespeare* series was published one hundred and forty years ago in 1871 (Horace Howard Furness's *Romeo and Juliet*) and the Modern Language Association (MLA) has, rather heroically, continued the series into the twenty-first century.[4] The task deserves the label heroic because it is admirable and, for all practical purposes, impossible, at least with the hand-crafted methods currently in use. There is simply too much scholarship for human editors to maintain up-to-date variorum editions for the 37 plays commonly attributed to Shakespeare – as of June 2011, the MLA offers only nine volumes from the *New Variorum Shakespeare*.[5]

The tools exist with which to develop dynamic variorum editions – the same explosion of data that overwhelms traditional methods provides automated methods with the data that they need to detect patterns and structures. Technologies such as optical character recognition (OCR), named entity recognition (NER), topic detection and other forms of text

---

[1] http://www.dlib.org/dlib/march06/03contents.html
[2] http://www.archive.org
[3] The creation of sophisticated digital editions and dynamic variorum editions is a long standing research problem within the digital humanities, for some recent work in this area see (Price 2009), (O'Donnell 2009), and Schmidt (2010).
[4] http://www.mla.org/variorum_handbook
[5] http://www.mla.org/store/CID38

mining exist today, and, as more machine actionable data becomes available, these technologies become increasingly effective.

Our goal is to develop a model for such variorum editions that applies not only to primary and secondary sources produced in print culture and primarily focused upon a single language (the vast majority of scholarship reviewed in the *New Variorum Shakespeare* series is in English) but that can also track versions of, and discussions about, sources across time and space, language and culture.[6] In this case, classical authors such as Homer and Cicero are better test cases than Shakespeare because they imply coverage for thousands, rather than hundreds, of years and materials in a wide range of languages. A variorum edition for Aristotle would demand not only Greek but also Latin and Classical Arabic sources, as well all the usual languages of modern scholarship (English, French, German and Italian as a minimum).

At the same time, if a small number of canonical works have attracted an immense amount of attention, we also consider the challenge of contextualizing the rest of our sources about Greco-Roman culture or in Greek and Latin. The vast majority of surviving Greek and Latin sources fall well outside of the heavily studied canons and do not have the print infrastructure of editions, lexica, commentaries, and scholarship upon which we can build when we focus on more commonly read sources. It is at least as important to automate the production of lexical information and background information needed to support use of the vast majority of surviving sources.

This paper will report upon our findings as we addressed the practical issues of moving towards a system that could organize information about the heavily studied sources and bootstrap understanding of the rest. But before considering the practical results, we must frame this eighteen-month effort against the broader perspective.

1.2 Advancing the intellectual life of society

The ultimate goal for our work – and arguably for all work in the humanities – is to advance the intellectual life of humanity as broadly as possible. Just as primary sources and large data sets can surge beyond the handful of academic libraries and flow across the globe over the open net, we in the academy have an opportunity to realize this fundamental goal far more fully than with the system of specialist print journals and monographs that had largely reached its maturity in the nineteenth century.

Classicists are responsible for enabling the Greco-Roman world to play the most vigorous possible role within the general intellectual life of humanity. Ultimately this entails placing the complete record of antiquity in an accessible form where humanity can most fully realize it – the pioneering German Classicists of the 19th century called this the *Totalitätsideal*, the goal of getting all of it, all our data, archaeological as well as linguistic,

[6] Tracking the influence of specific works as well as the ideas they contain through automatic quotation detection, text reuse and other algorithmic measures has been explored quite prominently in terms of the Google Books corpus, see (Baptiste Michel 2011 et al.) as well as (Schilit and Kolak 2008).

about the full Greco-Roman world in an accessible format. The brothers Alexander and Wilhem von Humboldt, the one a biologist and explorer, the other a philosopher and linguist, lived from the eighteenth to the mid-nineteenth centuries. Wilhelm founded Humboldt University in Berlin, while his brother gave the Alexander von Humboldt Foundation its name. Both would have loved almost every aspect of the DID challenge (except for the absence of German participation).

Within this eighteen month project, we focused our work upon sources in Greek and Latin – itself a vast subject, because it entails not only working with primary sources in these languages but also with secondary sources in (at a minimum) English, French, German and Italian. Arabic language materials, in particular, are a critical, if largely underutilized, resource for anyone studying Aristotle in particular or ancient science in general. A student of Greek and Latin sources thus must work with scholarship produced in every period since antiquity – inscriptions and papyri from the ancient world, manuscripts and scholarly notes from the middle ages, Arabic editions of Aristotle, Galen, Euclid and other Greek scientists and philosophers produced between 800 and 1000 CE, translations of these Arabic editions into Latin from the twelfth century onward, a tradition of print scholarship that begins in the fifteenth century, and now an increasing body of born digital scholarly work.

We already face a catastrophe of riches. Even if we restrict ourselves to the 2.8 million digitized books available for public download from the Internet Archive, the general public now has physical access to more sources about human cultures in more languages, covering more chronological periods, and representing more cultures than all but a handful of university libraries could offer (simply adding Arabic, Sanskrit and Chinese establishes this, since many academic libraries without subject specialists have simply not collected materials in these languages).[7]

Physical access does not, however, entail intellectual access – we may be able to view beautifully scanned page images of every Greek source text from antiquity, but, unless we have had training in this language, each of us can only shrug, with Shakespeare's Julius Caesar, that it is "Greek to me." The relative handful of advanced researchers and library professionals cannot provide the introductions and explanatory notes, much less the translations, specialized glossaries, and descriptions of people, places, and organizations that appear in this vast space. Meaningful public access, combining both physical and intellectual access, entails public participation. Society cannot realize the value of the materials available already in digital form unless advanced researchers and library professionals engage student researchers and citizen scholars in a shared task of discovery and conversation.[8] Wikipedia is arguably the most important phenomenon of the early

[7] For further discussion of how mass digitization projects such as the Internet Archive and Google Books, as well as the digitization of cultural heritage materials in general, are challenging academic libraries to redefine both their core mission and what services/collections they need to offer in a digital world, see Shaw (2010) and CLIR (2010) among many others.

[8] Lisabet Rausing has also discussed the importance of opening up access to databases and digitized collections both created and licensed by research libraries in order to more fully engage the general public

twenty-first century because it established the viability of the new decentralized, community driven mode of intellectual production upon which we must depend if society is to realize the potential intellectual value of vast public collections.

There is a corollary to the proposition above. If a handful of professionals working with traditional methods cannot make intellectually accessible the contents of very large collections, then very large collections can only realize their full potential insofar as they are open, not only for reading but also for correction, annotation, and refinement. If we compare the text that the Internet Archive has produced with OCR software to that which we can find in commercial textual databases, the commercial databases are infinitely superior.

If, however, we consider collections to be living entities and their value to consist not only in what they are at any given point in time but in what they can become, then the situation becomes quite different. If communities can customize OCR software for their particular domains (e.g., Latin and Classical Greek, as well as Classical Arabic, German Fraktur, eighteenth century English with the long s), correct the results of that OCR, add structural markup and linguistic annotation, and produce new versions, then, in effect, the communities can develop the collections that they need.[9] By contrast, every researcher in the humanities is probably aware of digital collections, produced years ago, which have remained frozen in time, in formats that are now idiosyncratic at best, and hidden behind search interfaces and encryption software at worst. Commercial business models that depend upon monopoly access and that use licensing agreements and the threat of law suits may have served corporations and university-based commercial entities well, but even open source projects such as Perseus know all too well that small editorial teams, however well-meaning, constitute bottle necks when curating large and diverse corpora.[10] But of course to realize the value of digital collections, we need to develop and then republish our own versions of existing corpora – such repurposing is the logical extension, in a digital age, of what we have always done in building new editions, lexica, commentaries.

2. Current Status

Both the Internet Archive and ultimately Google made available to us digitized versions of Greek and Latin source texts. We had worked with the Internet Archive to develop a collection of Greek and Latin sources for several years. In 2010, Google made available to us approximately 1,000 digitized versions of Greek and Latin sources with which to work.[11] The Google sources were particularly valuable because Google made available to us the full

---

with existing academic scholarship as well as to create new scholarly opportunities both in the humanities and the sciences (Rausing 2010).

[9] Libraries and cultural heritage institutions are increasingly looking to use crowdsourcing as a means of improving and adding value to digitized collections, particularly in the areas of OCR correction and text (Chrons and Sundell 2011, Holley 2010), image tagging (Vaughn 2010), and the addition of structured metadata (e.g. subject headings or concept hierarchies) (Eckert et al. 2010).

[10] On the need for new collaborative and open digital editing models, see (Crane 2010)

[11] For the list of fully downloadable texts and page images, please see (http://www.google.com/googlebooks/ancient-greek-and-latin.html) and for a list of books where download is restricted to users in the United States, see http://www.google.com/googlebooks/ancient-greek-and-latin-limited-distribution.html

scans rather than simplified PDF versions of the page images. JSTOR had also provided us with approximately 100 years of journals, primarily in English but also including German, on Classical Antiquity. On top of this, we had access to the over 8 million words of carefully curated Greek and Latin sources in XML along with parallel English translations, as well as machine actionable reference works such as lexica and biographical and geographical encyclopedias that make up the Perseus Digital Library.

Our work focused on several major threads:

1) We worked with large amounts of often very noisy OCR-generated text and especially Latin to revisit our earlier question on an even larger scale: "What do you do with a billion words?"
2) We customized open source OCR to provide useful Greek text.
3) We developed an infrastructure whereby we could work with far more data than we could store, much less process, with desktop systems.
4) We coordinated our efforts with computer scientists working with the NSF-funded Mining a Million Books Project at UMass Amherst and with computer scientists working on topic modeling with support form a Google Digital Humanities Award at Princeton.

2.1. What do you do with a billion words?

In 2006 we had asked "What do you do with a million books?" In 2010, we found ourselves faced with the new challenge of "What do you do with a billion words of Latin? "Most of this work has either been published or is under editorial review, with drafts available. This section summarizes the major findings from this work and particularly results published in Bamman and Smith (2011), Bamman and Crane (2011), and Bamman, Babeu and Crane (2010). These papers address the basic questions: (1) How much Latin do we have in this collection and from what periods? (2) How can we view the changes in word senses over two thousand years and how can we read Latin source texts where the words may have very different meanings from those in classical Latin when we do not have a lexicon, wordnet or other lexicographic resource? (3) How do we organize many different versions of the same work, including not only different scholarly editions but also translations of that work into multiple languages?

First, we developed a corpus from the relatively unstructured collection available from the Internet Archive. David Smith had over several months in 2009 systematically downloaded 1.2 million books and made these available to the project. We wanted to extract the Latin sources from this collection – a task that did not prove to be straightforward (Bamman and Smith 2011).

The available metadata listed approximately 28,000 of the 1.2 million books as being primarily in Latin. When we ran language detection software on these 28,000 books, however, we discovered that about 4,000 of them were, in fact, not primarily in Latin – they were actually in some other language (such as Italian) or they were Greek editions with Latin title pages and introductions. When, in turn, we ran the same language detection software on the full 1.2 million books, we found that the available metadata had missed

about 4,000 books that actually were in Latin. Thus, in effect, the available language metadata did not prove to be very useful.

When we tried to study the 28,000 books that were primarily in Latin, the metadata involving dates also proved to be problematic. The available cataloguing data listed the publication dates for book publication, rather than the dates when the works found within the books were originally produced – thus, we found Cicero's *Orations* with dates ranging from the sixteenth through the twentieth centuries rather than in the first century BCE. A team of student researchers worked during the 2010/2011 academic year to provide the best dates that they could for the creation dates of the contents within each digitized book. Such dates are clearly rough – some books contain works that were produced over decades or even centuries, or works produced in a single period but for which we have no clear date. Nevertheless, we were able to create a dated corpus of 9,000 books primarily in Latin. These books contained 385 million words of Latin produced from approximately 200 BCE through 2000 CE, with the majority of Latin being produced after the Classical period. Extrapolating across the rest of the 28,000-book collection, we have more than 1 billion words of Latin.

The implications of this figure are significant. First, the figure represents a lower bound – if the first million books that we analyzed contain more than 1 billion words of Latin, then that figure will increase as the 15+ million books digitized by Google are analyzed. If we consider Latin sources not available in printed books among the vast archival holdings of European culture, the surviving corpus of Latin will clearly encompass billions of words. This is hardly surprising – Latin remained a major language of formal publication long after antiquity until at least c. 1800 (with Latin remaining the official language of the Kingdom of Hungary for another fifty years).

The consequences of this exercise are significant. The public corpus of Latin – the amount of Latin to which members of the public could realistically expect to view has exploded. Only a handful of research libraries had editions for any but the most important Greek and Latin sources. Academic libraries with a full collection of the Loeb Classical Library offered perhaps 10 million words of Latin – a billion words thus constitutes an increase of 10,000% in the amount of Latin available to most students of Latin in colleges and universities. A generation ago, the public corpus of Latin – the amount of Latin that members of the public could expect to encounter in public libraries or bookstores – was negligible. The surviving corpus of Latin was, for all practical purposes, as visible to the public as the dark side of the moon. Now anyone around the globe with access to the net can download a reasonable edition of almost any classical source and a vast body of postclassical sources.

But physical access does not, as noted above, imply intellectual access. Members of the public may be able to download beautiful scans of works by Augustine, Boethius, Thomas Aquinas, Galileo, Newton and other authors, but few could do much with the unmediated sources. Even students proficient in Classical Latin will find works produced after the Classical period obscure, because these often use familiar words in very new ways or assume specialized knowledge of subjects such as astrology, early modern Law, theology, painting, and architecture.

Second, we want to be able to trace the development of words and their meanings, both so that we can study the evolution of ideas over time and so that we can better read particular texts with non-classical words and word senses. Producing the intellectual infrastructure with which to understand such sources is not an easy task. Editions of individual works can take years and large reference works can take decades. The leading lexicon for Classical Latin, the *Thesaurus Linguae Latinae* (TLL), focuses upon the approximately 10 million words from the first 800 years of Latin. The TLL began in 1894, currently employs a full-time staff of 20, has completed almost 2/3 of its work and expects to finish its task in around 50 years.[12] Clearly, traditional methods will not scale when we consider the challenge of documenting the widely variable sources produced in every part of Europe over more than two millennia.

The Dynamic Lexicon Project, initially sponsored by the NEH Preservation and Access Research and Development Program, and the Greek and Latin Treebanks[13], supported by NSF and the Cantus Foundation, provided us with the instruments whereby we could address the challenge of providing background information about what words mean and how they are used (Bamman and Crane 2008). We faced at least three major challenges. First, the 385 million-word corpus of dated Latin was approximately two orders of magnitude larger than the 5 million words of Latin in curated XML source texts with which we had worked. Second, this same corpus was also variable in quality, because the OCR-software used by the Internet Archive and Google was optimized neither for sources in the Latin language nor for the print in many earlier editions. Third, the machine actionable data that we had developed for Latin focused upon the first five hundred years of Classical Latin (c. 220 BCE -250CE) whereas the new corpus covered a period of more than two thousand years, four times as long, and composed in wide range of cultural contexts throughout Europe.

We set out to address three problems: (1) to identify word senses for Latin words over two thousand years (e.g., How often does Latin *oratio* correspond to "prayer" vs. "speech"? How often does Latin *scientia* correspond to "knowledge" vs. "science" as systematic inquiry?); (2) to use English keywords to detect semantic shifts within Latin (e.g., How often does the English word "knight" correspond to the Latin word *eques*, which could designate in Rome a horseman or a member of the equestrian class, or to the Latin word *miles*, literally "soldier," but more generally the term that corresponds in later Latin to the conventional English sense of "knight"?); (3) to determine the most likely sense for any particular word in any particular context (e.g. in one passage, *scientia* probably corresponds to "knowledge" but in another passage – perhaps in the same work – it probably corresponds to "science").

To address these challenges, we built upon two pre-existing domain specific resources that we had developed at Perseus: (1) a morphological analyzer for Latin and (2) a 2.9 million word parallel corpus of Classical Latin and English. To augment this, we manually identified

[12] For more on the methodology behind the creation of the TLL and their progress, see Hillen and Coleman (2007).
[13] http://nlp.perseus.tufts.edu/syntax/treebank/

129 pairs of Latin source texts and English translations, covering the full 2000-year range of Latin. The translation pairs consisted of raw OCR-generated text and Latin source texts corresponded precisely to the English translation only 32.6% of the time. Even in these cases, the raw OCR lacked any structural markup to distinguish headers and footnotes, introductions and appendices from the Latin source texts and English translations. In more than 2/3 of the cases, the English translation translated only a part of the Latin text or only subsets of the two books corresponded (e.g., an anthology of Latin works translated into English). We were nevertheless able to extract more than 40,000 parallel sentences in Latin and English and then to align more than 500,000 Latin and English words. This 500,000 word parallel corpus provided the training set with which to explore the relationship between Latin and English and between English and Latin and to rank the most probable sense for particular words in particular passages.

The outcome of this work is that we can in fact trace changes in the meanings of words over thousands of years. This ability supports the development of services that address the three problems outlined above. We can start with a Latin word and use its corresponding English senses to view its semantic development. Conversely, we can start with an English word and see how different Latin terms correspond to it over time. We can also provide reading support for those reading an unfamiliar text in postclassical vocabulary or with classical vocabulary that has new meanings.

Third, we need to be able to organize many different versions of the same work, not only aligning multiple editions of the same text in its source language but also translations of that work into many different languages, as well as identifying associated reference works such as commentaries, specialized lexica, and indices, and detecting quotations of passages from these works in their original language and in various translations. Users of the Perseus Digital Library are familiar with such integration applied to a curated collection: in calling up the text, the reader is presented with a list of available editions, translations, commentaries, and works that cite this particular passage. Perseus draws upon curated markup within its source texts to accomplish this goal. As we scale up to very large collections, creating such markup for every digitized book is simply not practical.

We can already with reasonably high accuracy align versions of the same text across languages. Thus, we have methods with which to line up many different Latin editions and modern language translations for a passage of Cicero. We need to address two additional tasks. First, while we may not be able to annotate every edition of Virgil or Cicero available in digital form, we do have XML editions for these and a growing body of other authors. Can we use the markup in one edition to organize many other editions and translations? Second, if we have different markup in different versions of the same text (e.g., one text has marked the speeches of Thucydides and the other has identified people and places), how well can we integrate the information in these disparate sources and then transfer that information across hundreds or thousands of versions?

Our goal is to support queries such as:

- Identify all versions of book 2, chapter 38, section 2 of Thucydides' *History of the Peloponnesian War*, identifying where Greek editions differ, and providing a visualization of where this passage has been quoted, in what languages, and in what contexts?
- Compare the language of two characters in a play as their words have been translated into multiple languages across time?
- Show for an influential eighteenth century edition of an author, all annotations about a given passage.

Bamman, Babeu and Crane (2010) presented a method for automatically projecting structural information across translations, including canonical citation structure (such as chapters and sections), speaker information, quotations, markup for people and places, and any other element in TEI-compliant XML that delimits spans of text that are linguistically symmetrical in two languages. We evaluate this technique on two datasets, one containing perfectly transcribed texts and one containing errorful OCR, and achieve an accuracy rate of 88.2% projecting 13,023 XML tags from source documents to their transcribed translations, with an 83.6% accuracy rate when projecting to texts containing uncorrected OCR. This approach has the potential to allow a highly granular multilingual digital library to be bootstrapped by applying the knowledge contained in a small, heavily curated collection to a much larger but unstructured one. The paper was well received and was chosen from the 130 submissions for the best paper award.

2.2 Customizing OCR and Classical Greek

The work reported in Section 2.1 built upon the raw OCR-generated text available for download from the Internet Archive. This work thus establishes a baseline for what is possible, even with source texts that can contain extensive noise, because the OCR software used by the Internet Archive is not customized either for Latin or for the printed typefaces found in many digitized books.

For many purposes, small error rates do not have a major impact.[14] Consider, for example, the established task of searching Greek and Latin source texts. Traditionally we search a curated text collection and then hunt down the print originals so that we can check the variant readings or other ancillary materials. For this purpose, image-front searching, where we type a query and then see page images (as in JSTOR and Google Books) is often a preferable solution. Instead of viewing de-contextualized snippets, users can instead see the original page image, with its notes and associated materials. If the OCR is good enough, then users can still cut and paste the OCR-generated text into their notes, correcting it as needed. Earlier research determined that between 5% and 15% of the unique words in a given Greek or Latin edition appears only in the textual notes – so 95% accurate OCR of a Greek or Latin source text gets more of the words than a perfect transcription of the reconstructed text alone, because the reconstructed text only provides 85% to 95% of the textual data in the edition (Stewart, Crane and Babeu 2007). Such a system can sustain small error rates well by drawing upon standard techniques for fuzzy searching. And, if we

[14] For a thorough overview of how OCR errors can affect natural language processing tasks, see (Lopresti 20008)

are able to align multiple editions of the same work, then if the OCR fails on a word in one edition, it may well succeed on the same word in another.[15]

A great deal of work remains to be done to improve Latin OCR, but the available OCR-generated text was good enough for initial work.[16] OCR for Classical Greek, however, remains the major barrier for digital classics. Earlier publications (Crane, Stewart and Babeu 2007, Boschetti et al. 2009) established that OCR could generate Greek text at a level of accuracy that would easily support searching. In the best case, we compare the results from multiple OCR engines and use various methods (e.g., Greek spell checking, language models, voting) to determine which OCR engine has the most plausible analysis where they differ. Boschetti worked with the open source OCRopus[17] in bundle with Tesseract OCR engine as well as with the commercial Abbyy[18] and Anagnostis[19] systems.

Within our Digging into Data project, we needed to scale this work up so that we could run Greek OCR on millions of pages, not only analyzing known Greek editions but hunting for quotations in Classical Greek in sources that were primarily in English, French, German, Italian, Spanish or other languages with a Roman alphabet. In this context, two of the three OCR engines that we had used were not suitable. Anagnostis could not, as far as we could tell from the documentation, work with large bodies of text – each page needed manual attention. More seriously, while the Abbyy OCR package could run on large bodies of source texts, the Abbyy company offered a licensing model that charged by CPU – a model that does not make economic sense when we move, as move we must when working at scale, to systems with hundreds or thousands of CPU-s. To develop a reliable system, we needed to move to an open source OCR service that we could customize and deploy as we needed.

Federico Boschetti, who had returned to his teaching job in Italy after six months at Perseus, and Bruce Robertson, the Principal Investigator from Mount Allison University in Canada, worked on the challenge of customizing open source OCR engines to the wide range of Greek source texts available to us.

The open source document analysis framework 'Gamera'[20] already provided a preliminary Greek OCR application as well as GUI application to classify characters. To make the Greek OCR code more effective we added to it the character splitting and component joining

---

[15] Feng and Manmatha (2006) reported great success in using this methodology in a digital library of historical books.

[16] Research into the development of OCR systems as well as advanced document recognition technology for historical languages such as Latin and Greek has grown extensively in the last few years through projects such as the IMPACT (Improving Access to Text) project (Ploeger et al. 2009) and for some other recent work see (Leydier et al. 2009) and (Vamvakas et al. 2008).

[17] http://code.google.com/p/ocropus/. Other research has also examined the potential use of OCRopus with historical languages such as Sanskrit (Breuel 2009) and its potential use in a large-scale historical document recognition workflow (Bryant et al. 2010).

[18] http://www.abbyy.com

[19] http://www.ideatech-online.com

[20] http://gamera.informatik.hsnr.de/addons/ocr4gamera/index.html, and for an earlier examination of the potential uses of Gamera in digital libraries see (Choudhury et al. 2006).

features from the base Gamera library. We also fine-tuned the OCR's word splitting algorithm.

Teams of undergraduate students worked on producing classifiers suitable for ancient Greek OCR. Rather than trying to provide a single classifier that necessarily would work poorly on many texts, they identified eighteen distinct font families among the 18th- to 20th-century Greek texts, and produced classifiers tuned to each of these families.

Moreover, near the end of the grant period, it became clear that more sophisticated methods of automatically analyzing page layout would most dramatically improve the OCR of pre-twentieth-century ancient Greek texts. Columned pages – for instance, those containing side-by-side columns of Greek and English – ruin OCR output because the height of the Greek lines are different from those in other languages. By analyzing the page in two steps, effectively forcing it to treat each of these columns as a new context, we have improved the OCR engine's accuracy greatly, and we expect this technique will be widely applicable.

The Mount Allison group also explored the effectiveness of various GUIs in searching their OCR'd texts. They integrated ancient Greek morphology data from Perseus with the OCR generated data to create a morphologically aware, image-fronted OCR search tool that made use Ruby on Rails.[21] Figure 1 below shows the results of a search on one Greek dictionary word as it retrieves thumbnail images of all the pages that contain forms of that word. Figure 2 then shows the result of clicking on one of the thumbnail images. The page image is displayed full size, and the word(s) that are instances of the dictionary word are highlighted in yellow. We imagine adding OCR correction and other services to a GUI such as this.

[21] http://rubyonrails.org/

Figure 1: Searching for a Greek word

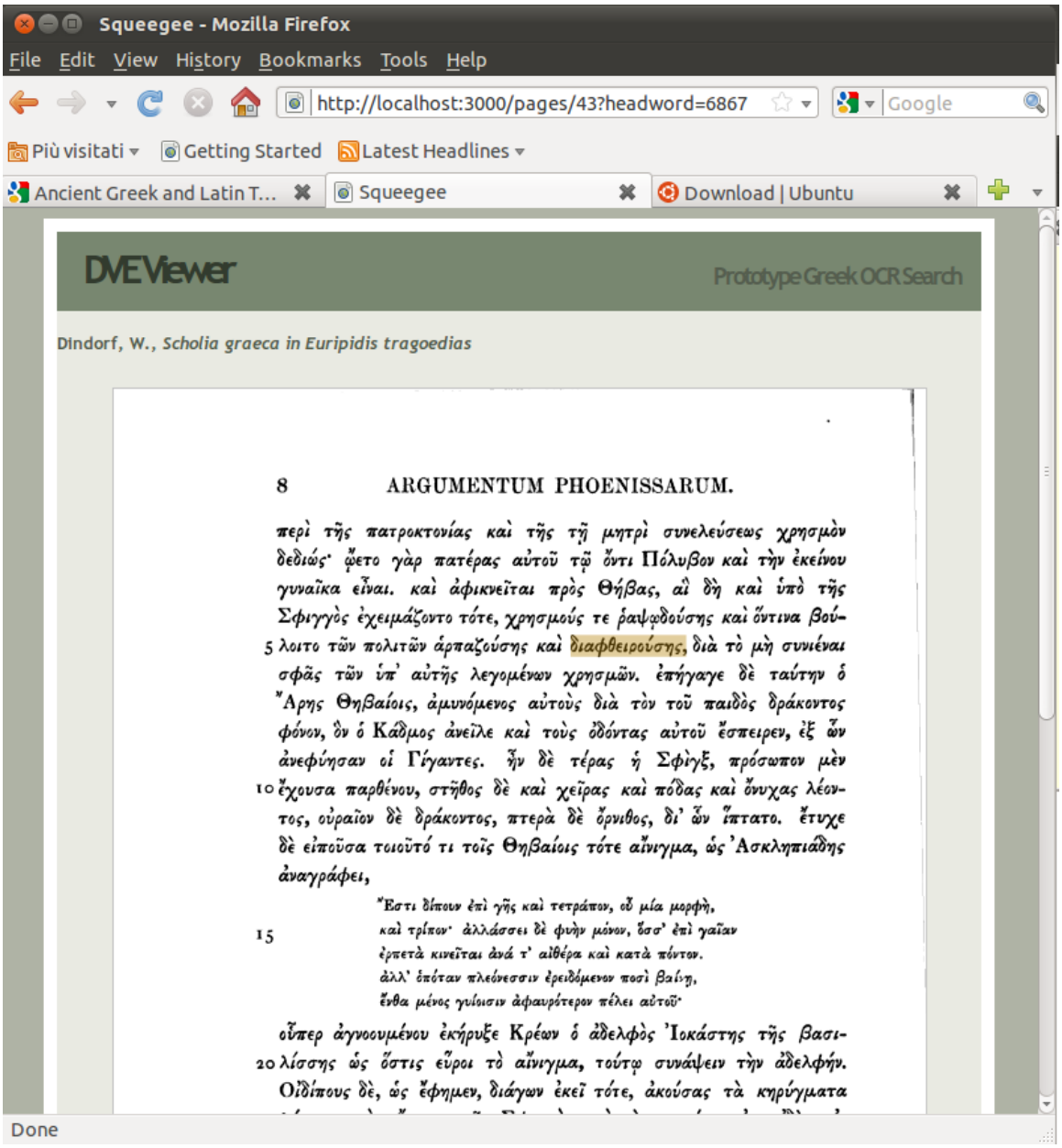

Figure 2: A fuller view of one of the thumbnail images.

In the end, the largest amount of development work was conducted with Tesseract. While the Mount Allison Group made a lot of progress with the open source OCR engine GAMERA, they were unable to get the new OCR engine in OCRopus to work. Ultimately, we do expect that OCRopus will provide a third open source OCR engine.

2.3 Scaling up to work with millions of pages.

Imperial College London's contribution to the DVE was to design and implement a scalable infrastructure for processing large datasets characteristic of humanities data, in this case, page images from books printed before 1900.

The motivation for building such an infrastructure comes from the Internet—specifically from the influx of enormous quantities of scanned page images as the world's libraries are digitized. Traditionally, humanities computing has been a 'one-man one-machine' business, with digital work focusing mainly on the creation of high-quality datasets with intensive input from a single researcher, and large-scale computing resources required only to support publication of these datasets. A good example of this in Classics would be the Thesaurus Linguae Graecae (TLG)[22] or the Perseus Project in its early years.

The enormous growth in the availability online of scanned material from the world's libraries and archives has radically altered this picture in several ways. First, there is now far more material available online than any team of researchers can mark up, let alone read. Secondly, a great deal of that material lies outside of the traditional canon of Classical literature and scholarship, with the result that researchers are in many cases simply unaware of its existence. Neo-Latin,[23] where many texts were published only in small quantities, and survive in only a few exemplars, is a good example of this. Lastly, the number of online page images containing references to[24] or uses of classical literature dwarfs classical literature itself in size and transience—while classical literature consists of an almost fixed body of texts which are available in well edited versions, the enormous body of literature that refers to classical culture and literature is constantly changing—in the quality of the scans that are available, in our knowledge about the quality and nature of the resources (duplicates? translations? Are the editions related?), and in our ability to process them with the help of metadata about their language and content. In each of these cases, the ability to process large data at scale becomes central. This shifts both the kind of computational support needed for research in the humanities and, more significantly, the kind of research that Humanities scholars are likely to do.[25]

Our approach to scale was to look at parallelizing core OCR tasks and, once we had parallelized them, at ways of scaling the infrastructure on demand. To support our parallel infrastructure, we chose Hadoop, a parallel processing infrastructure from the Apache Foundation[26] and for a scalable infrastructure, we chose Eucalyptus, an open-source Cloud platform, which enables us to work with virtual computing resources.

Out first task was to devise approaches to OCR that took advantage of parallel methods. This required some rethinking of standard OCR approaches. Typically, OCR is concerned with achieving maximum accuracy for the outcome for each page. However, with large

---

[22] http://www.tlg.uci.edu/

[23] There are a large number of neo-Latin texts available online, as documented by the "Philological Museum: An Analytic Bibliography of On-Line Neo Latin Texts", http://www.philological.bham.ac.uk/bibliography/

[24] The need to provide sophisticated linking services between digitized classical texts and the scholarly literature that cites them (e.g. in academic journals or monographs) as well as the challenges of automatically extracting primary text citations has been discussed extensively in (Romanello 2008, Romanello et al. 2009a)

[25] A number of large ongoing research projects are also exploring what type of infrastructure will be required to support both the scalable creation and storage of large-scale humanities data as well as researching the development of algorithms to process and make this data reusable, including Project Bamboo (http://www.projectbamboo.org/), DARIAH (http://www.dariah.eu/), CLARIN (http://www.clarin.eu/) and TextGrid (http://www.textgrid.de/en.html)

[26] http://hadoop.apache.org/

datasets, researchers may have preliminary questions that are less concerned with accuracy and more concerned with retrieval. For instance, a researcher may want to know which pages in a large dataset of books contain polytonic Greek, and be content with a simple list of pages without full-text OCR.

Another consideration with large datasets is utility, or the ability to answer the question: "Which portion of this dataset is it most efficient to process first?" OCR is notoriously bad with polytonic Greek. However, once a training set for a particular font and series is produced, accuracy increases dramatically. One of the most useful services a large parallel processing infrastructure can provide is a vetting process that assesses the likelihood of a good OCR outcome for a particular book. This can be achieved by running a few pages of a book against all available training sets and scoring these results.

A third consideration was the elimination of bottlenecks in the OCR workflow that prevent parallelism. Of these, by far the most significant is ground-truth, i.e. the need to consult the original page when first running a text through an OCR engine or what is called "training". Each training set requires a pair of eyeballs to discard incorrect identifications and add correct ones. In this case, the solution was the Boschetti Aligner, which uses innovative techniques such as likelihood of syllabic sequences, dictionary lookups, and competitive workflow scoring, to achieve an accuracy that begins to rival that of a human corrector. Initial training sets are still needed, but far less of them, and therefore far fewer eyeballs.

We boiled these approaches down to three basic "recipes":

1) Multiple engines – run each page on multiple OCR engines, score results using the Boschetti Aligner. The goal of this process is to free OCR training from reliance on eyeballs.
2) Filters: run a single page image through multiple filters.
3) Multiple training sets: run selected pages against multiple training sets; compare scores to find the OCR output with the highest predicted accuracy.

To implement and test these strategies, we first created a small test Hadoop cluster on fixed resources and implemented the OCR stack on this cluster. The aim was to establish that the parallel approaches we had devised, which required several different kinds of software, could actually be implemented in the Hadoop framework.

Once we were sure that a Hadoop implementation was feasible, we began the process of creating a custom virtual machine (VM) image that would contain the entire OCR software stack required for parallel processing. The motivation for creating a VM image was that it enables us to make use of "infrastructure as a service." This is a powerful new computing paradigm that allows users to customize operating systems on a per use basis and deploy compute nodes "on demand." It is an essential component in a fully scalable system, as it frees us from the need to deploy new hardware every time we need to increase the size of a cluster.

At the same time as we were creating the VM image, we were setting up a series of Eucalyptus VM clusters, for DVE and other research clusters. The final version of the Eucalyptus cluster consists of 20 dual quad core Viglen machines with a 16TB fibre-connected storage device on loan from the Large Hadron Collider project. The cluster runs Ubuntu and Eucalyptus 2. On this cluster it is possible to deploy up to 120 compute nodes. The diagram below (Figure 3) illustrates the framework from hardware (bottom) to software (top).

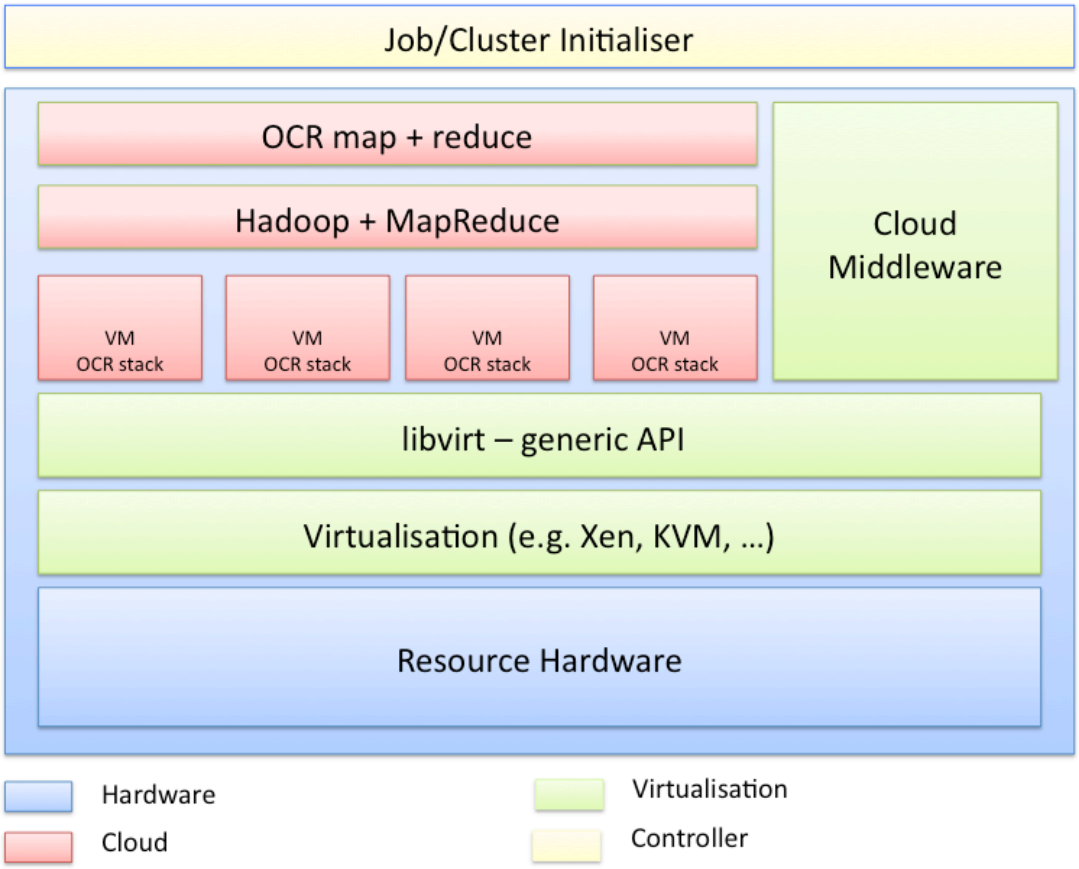

Figure 3: DVE Architecture

We have now reached the point where we are ready to process some large datasets. Initial tests have shown that, with an average processing time of around 2 minutes per page, an entire book of 120 pages can be processed in under 10 minutes. As we add capacity to the system, we expect this time to decrease and throughput to increase.

The parallel system we have built is not only a practical OCR tool which vastly accelerates the processing of polytonic Greek; it is also a proof-of-concept for a generic customizable parallel architecture for at-scale processing of humanities data. This makes it likely that a much larger set of humanities feature extraction processing, such as part of speech (POS)

tagging and named entity extraction, can be successfully refactored to take advantage of massive parallelism.

Many challenges, however, remain. To give just one example, our system is most efficient when we can make a single pass on data. The distributed architecture of Hadoop storage makes any iterative process that requires storing and then retrieving data extremely costly in computational terms. Many feature extraction codes—especially those that make use of statistical methods—rely on frequent iterations to achieve good results. Adapting these methods to parallel frameworks that rely on commodity hardware remains a significant research challenge.

2.4 Other collaborations

These collaborations illustrate what we can quickly do in the future by building on related work. The key points include:

1) The UMASS Mining a Million Books project has created a scalable research system that can apply named entity recognition, morpho-syntactic analysis and other key services to billions of words. This feeds into our need to scale up.
2) UMASS also has published work on automatically detecting translations within a large corpus – this also will help us build our parallel corpora (Mimno et al. 2009).
3) UMASS has done important work detecting multiple versions of the same work.
4) UMASS, under the direction of David A. Smith has used an NEH startup grant to improve named entity recognition for noisy OCR.[27]
5) eAqua[28] and now eTraces in Leipzig have developed systems whereby we can detect small quotations of Greek and Latin sources. They have demonstrated this by detecting testimonia of Plato and citations or text reuse of this author by later authors (Büchler and Gesner 2009).
6) David Mimno, now a postdoc at Princeton, received a Digital Humanities Grant from Google.[29] This has allowed him to continue applying topic detection to classical studies (focusing now on finding topics about archaeological sites from Google Books).

3. Conclusion

A great deal more needs to be done and an eighteen-month effort can only constitute one step in a larger process. Nevertheless, there are at least three major implications from the work done so far.

[27] http://www.neh.gov/ODH/ResourceLibrary/LibraryofFundedProjects/tabid/111/Default.aspx
[28] http://www.eaqua.net/en/index.php
[29] The project is entitled, "The Open Encyclopedia of Classical Sites" and for the grant announcement see http://googleblog.blogspot.com/2010/07/our-commitment-to-digital-humanities.html

First, the study of the Greco-Roman world in many ways enables – and entails – the study of all subsequent European culture. Some classicists from the 18th century German Art Historian Winkelmann on have stressed the separate and exceptional nature of Greco-Roman culture and seen subsequent sources as a means to the end of understanding the Classics. The rise of vast and comprehensive collections accessible to anyone on the net greatly strengthens the position of those who wish to emphasize the full classical tradition. This tradition points in two different directions. On the one hand, we depend upon scholarship from Ancient Alexandria, ninth century Baghdad, 15th century Venice, 19th century Germany, and 20th century North America to reconstruct that classical world. At the same time, every reader of Steven Saylor's detective novels set in Ancient Rome and every viewer of films such as *Alexander* or *Gladiator*, the HBO television series *Rome*, or the various documentaries on the Greco-Roman world from venues such as the History Channel experiences the influence of Classical Antiquity.

Professional classicists have turned with renewed vigor to reception studies – the study of how different periods read and understood classical sources. The rise of huge collections and new analytical methods is certainly important because it can provide new tools with which scholars can see what later ages had to say about Greek and Latin sources.

Very large open source collections are, however, even more important because they open up the classical tradition to society as a whole. We are no longer simply writing chapters for $150 monographs that only specialists will see where we talk about early modern or eighteenth century sources that appear in no public library or otherwise generally accessible source. We are able to develop intellectual conversations that can, if we so choose, serve to advance our understanding and to reach a wider audience – and to both at once without compromise.

Such a goal leads, however, to the next conclusions.

Second, the study of the Greco-Roman world involves more than Europe, North, Central and South America, Australia, and New Zealand. The Greco-Roman world extended from Rabat to Kandahar. Many of us know in general that Arabic scholarship played a critical role in making the modern West possible. Euclid re-entered Western Europe via translation from Arabic into Latin. Arabic translations of Aristotle, made between 800 and 1000 CE, reflect readings that are no longer preserved in Greek manuscripts and thus are essential to any comprehensive edition of this author. The University of Cairo hosts one of the most vigorous departments of Greek and Latin studies in the world. Almost a hundred years ago, the Egyptian nationalist author and intellectual Taha Hussein founded this department to express Egypt's participation in a broader European culture. When the Ayatollah Khomeini participated in developing the Islamic Republic of Iran, he went back to his training in Greek philosophy (he had followed ancient Islamic tradition by studying Aristotle's *Organon* in graduate school) and created an Islamic version of Plato's *Republic*, with philosopher king and council of guardians.

We may know something about Islamic contributions to, and engagement with, Greco-Roman culture but that knowledge remains vague and distant because the Arabic and later

Latin sources have traditionally been physically and intellectually inaccessible. We can now see this beginning to change, but physical access to Classical Arabic does not, of course, provide intellectual access. And Classical Arabic is very different from the Modern Standard Arabic that students labor to learn in American universities. We cannot conduct research in, or disseminate sources about, the great translation movements from Greek into Arabic (c. 800-1000 CE and starting in Baghdad) and then from Arabic into Latin (c. 1200 and starting in Spain) unless we develop new intellectual partnerships with our colleagues in Cairo and throughout the Arabic speaking world.

Third, we now have far more work to do than the relative handful of advanced researchers and library professionals can ever accomplish. We need a new, decentralized culture of intellectual activity, one where experts work hand in hand with student researchers and with citizen scholars. We must view community driven efforts such as Wikipedia as an inspiration and as a challenge to create shared conversations about the past that engage society as broadly as possible and that are at the same time more rigorous because the sources of those conversations are physically and intellectually accessible – projects such as the revolutionary Homer Multitext[30] have begun to make this statement tangible. Within the classical canon, even if we can convert our print information into machine actionable data, that conversion is only the first, incunabular step towards creating a true digital infrastructure. We now can support – indeed, we must support – a laboratory culture, where almost anyone can contribute at a very early stage and where each contributor can assume increasingly complex tasks as their abilities develop.

No field is better poised to realize the benefits of this new intellectual culture than the study of Greek and Latin. We have an opportunity to develop an education within K-12 and higher education, to foster life long learning and to ask bigger questions than we could in the past and to bring new data as well as new perspectives to old questions as well. In this, Classics has an opportunity not only to advance its own position but also to play a leading role in transforming the relationship between what we in the humanities do within the academy and our obligation to advance the intellectual life of society as a whole.

4. References

Bamman, David, Alison Babeu, and Gregory Crane (2010). "Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection. In JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries, New York, NY, USA, pp. 11-20. ACM. Preprint available at: http://hdl.handle.net/10427/70398

Bamman, David and Gregory Crane. (2008). "Building a dynamic lexicon from a digital library." *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, 2008, 11-20. Preprint available at: http://hdl.handle.net/10427/42686

[30] http://www.homermultitext.org/

Bamman, David and Gregory Crane. (2011). "Measuring Historical Word Sense Variation." In Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital libraries (JCDL 2011), ACM Digital Library. Preprint available at: http://www.perseus.tufts.edu/publications/bamman-11.pdf

Bamman, David and David Smith. (2011-Forthcoming). "Extracting Two Thousand Years of Latin from a Million Book Library." *Journal of Computing and Cultural Heritage,* Forthcoming, http://www.perseus.tufts.edu/publications/01-jocch-bamman.pdf

Boschetti, Federico, Matteo Romanello, Alison Babeu, David Bamman, Gregory Crane. (2009). "Improving OCR Accuracy for Classical Critical Editions." In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 156-167, Corfu Greece: Springer Verlag, 2009-09 Preprint available at: http://hdl.handle.net/10427/70402

Breuel, Thomas. (2009). "Applying the OCRopus OCR System to Scholarly Sanskrit Literature." *Sanskrit Computational Linguistics* (2009): 391-402.

Bryant, Michael, Tobias Blanke, Mark Hedges, and Richard Palmer. (2010). "Open Source Historical OCR: The OCRopodium Project." *Research and Advanced Technology for Digital Libraries*, pp. 522-52

Büchler, Marco and Annette Geßner. (2009). "Citation Detection and Textual Reuse on Ancient Greek texts." *DHCS 2009-Chicago Colloquium on Digital Humanities and Computer Science*. http://lingcog.iit.edu/%7Eargamon/DHCS09-Abstracts/Buechler-Gessner.pdf

Choudhury, Sayeed G., T. Dilauro, R. Ferguson, M. Droettboom, and I. Fuginaga. "Document Recognition for a Million Books." *D-Lib Magazine* 12 (2006). http://www.dlib.org/dlib/march06/choudhury/03choudhury.html

Chrons, Otto and Sami Sundell. (2011). "Digitalkoot: Making Old Archives Accessible Using Crowdsourcing." *HCOMP 2011: 3rd Human Computation Workshop*. http://cdn.microtask.com/research/Digitalkoot-HCOMP2011-Chrons-Sundell.pdf

Council on Library and Information Resources. (2010). *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*. CLIR Publication, No 147, June 2010, http://www.clir.org/pubs/reports/pub147/pub147.pdf

Crane, Gregory. (2006). "What Do You Do with A Million Books?" *D-Lib Magazine*, 12 (3), http://www.dlib.org/dlib/march06/crane/03crane.html

Crane, Gregory. (2010). "Give us Editors! Re-Inventing the Edition and Re-Thinking the Humanities." In *Online Humanities Scholarship: The Shape of Things to Come*, University of Virginia: Mellon Foundation, 2010-03, http://cnx.org/content/m34316/latest/

Eckert, Kai, Mathias Niepert, Christof Niemann, Cameron Buckner, Colin Allen, and Heiner Stuckenschmidt. "Crowdsourcing the Assembly of Concept Hierarchies." *JCDL '10: Proceedings of the 10th Annual Joint Conference on Digital Libraries*. New York, NY, USA: ACM, 2010, 139-148.

Feng, Shaolei and R. Manmatha. "A Hierarchical, HMM-Based Automatic Evaluation of OCR Accuracy for a Digital Library of Books." *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, 2006, 109-118.

Hillen, Michael, (translated by Kathleen M. Coleman). (2007). "Finishing the TLL in the Digital Age: Opportunities, Challenges, Risks." *Transactions of the American Philological Association* 137 (2007): 491-495

Holley, Rose. (2010). "Crowdsourcing: How and Why Should Libraries Do It?" D-*Lib Magazine*, 16, 3/4, http://www.dlib.org/dlib/march10/holley/03holley.html

Leydier, Yann, Asma Ouji, Frank LeBourgeois, and Hubert Emptoz. "Towards an Omnilingual Word Retrieval System for Ancient Manuscripts." *Pattern Recognition* 42 (September 2009): 2089-2105.

Lopresti, Daniel. (2008). "Optical Character Recognition Errors and their Effects on Natural Language Processing." *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*. New York, NY, USA: ACM, 2008, 9-16.

Michel, Jean-Baptiste, et al. (2011). "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, 331 (January 2011): 176-182.

Mimno, David, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. (2009). "Polylingual Topic Models." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, 880-889.

O'Donnell, Daniel P. (2009). "Back to the Future: What Digital Editors can Learn from Print Editorial Practice." *Literary and Linguistic Computing,* 24 (April 2009): 113-125.

Ploeger, Lieke, Yola Park, Jeanna N. Gaviria, Clemens Neudecker, Fedor Bochow, and Michael Day. "'IMPACT Conference: Optical Character Recognition in Mass Digitisation'." *Ariadne* (June 2009). http://www.ariadne.ac.uk/issue59/impact-2009-rpt/

Price, Kenneth M. (2009). "Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?" *Digital Humanities Quarterly*, 3 (3). http://digitalhumanities.org/dhq/vol/3/3/000053/000053.html

Rausing, Lisbet. (2010). "Toward a New Alexandria: Imagining the Future of Libraries." *The New Republic*, (March 2010). http://www.tnr.com/print/article/books-and-arts/toward-new-alexandria

Romanello, Matteo. (2008). "A Semantic Linking Framework to Provide Critical Value-Added Services for E-Journals on Classics." *ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing*, (June 2008): 401-414.
http://elpub.scix.net/cgi-bin/works/Show?401_elpub2008

Romanello, Matteo, Federico Boschetti, and Gregory Crane. (2009a) "Citations in the Digital Library of Classics: Extracting Canonical References By Using Conditional Random Fields." *NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Morristown, NJ, USA: Association for Computational Linguistics, (2009): 80-87. http://aye.comp.nus.edu.sg/nlpir4dl/NLPIR4DL10.pdf

Schilit, Bill N. and Okan Kolak. (2008). "Exploring a Digital Library through Key Ideas." *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. New York, NY, USA: ACM, 2008, 177-186.

Schmidt, Desmond. (2010). "The Inadequacy of Embedded Markup for Cultural Heritage Texts." *Literary and Linguistic Computing*, 25 (3), 337-356.

Shaw, Jonathan. (2010)"Gutenberg 2.0: Harvard's Libraries Deal with Disruptive Change." *Harvard Magazine*, May, http://harvardmagazine.com/2010/05/gutenberg-2-0

Stewart, Gordon, Gregory Crane, Alison Babeu. (2007). "A New Generation of Textual Corpora: Mining Corpora from Very Large Collections." In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL 2007)*, pages 356-365, Vancouver, British Columbia: ACM Digital Library. Preprint available at:
http://hdl.handle.net/10427/14853

Vamvakas, G., B. Gatos, N. Stamatopoulos, and S. J. Perantonis. (2008)."A Complete Optical Character Recognition Methodology for Historical Documents." *Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop,* 2008, 525-532

Vaughan, Jason. (2010). "Insights Into The Commons on Flickr." *Portal: Libraries and the Academy*, 10 (2010): 185-214.
http://www.press.jhu.edu/journals/portal_libraries_and_the_academy/portal_pre_print/current/articles/10.2.Vaughan.pdf