Cyberinfrastructure and the Next Generation of Ancient Corpora

Gregory Crane and David Bamman

The Perseus Project, Tufts University

1. Introduction

The Perseus Digital Library (Crane 1987, Crane et al. 2001) has for 20 years created an open reading environment for the study of Classics, serving 3.4 million words of carefully curated Latin source texts and 4.9 million words of Greek along with attendant commentaries, translations and linguistic annotations. These texts are all public-domain materials that have been scanned, OCR'd and formatted into TEI-compliant XML, and are used on average by 400,000 distinct users every month.

The scope of this corpus, however, pales in comparison with the large million book collections that are now taking shape, as Google, the Open Content Alliance (OCA) and the European i2010 initiative are all laying the foundations for vast digital libraries. Rather than containing a single edition of a source text, these massive libraries will contain multiple editions (by different editors from different eras), translations into dozens of languages, and thousands of books that quote some passage in the original text. The availability of these collections has the potential to significantly transform our ability to analyze textual materials, and to create new reference works built from those texts.

The Perseus Digital Library has been developing technologies for creating the next generation of ancient corpora – a "cyberinfrastructure" that emerges from the interaction of a small, highly structured corpus with a much larger, unstructured one. In particular, our research has focused on the ways in which we can automatically mine high value data using the human curated reference works in our existing collections to automatically create new "cybereditions" of source texts for a demanding scholarly audience.[1]

In the following we will present the state of structured collections within the Perseus Digital Library – highlighting especially the extensibility of the infrastructure to other languages – and outline our current research into integrating those structured collections with the massive unstructured ones now coming into existence.

---

1 For more on the application of cyberinfrastructure to the humanities, see Gietz et al. 2006 and the final report of the ACLS Commision of Cyberinfrastructure (2006).

2. Structured collections in the Perseus Digital Library

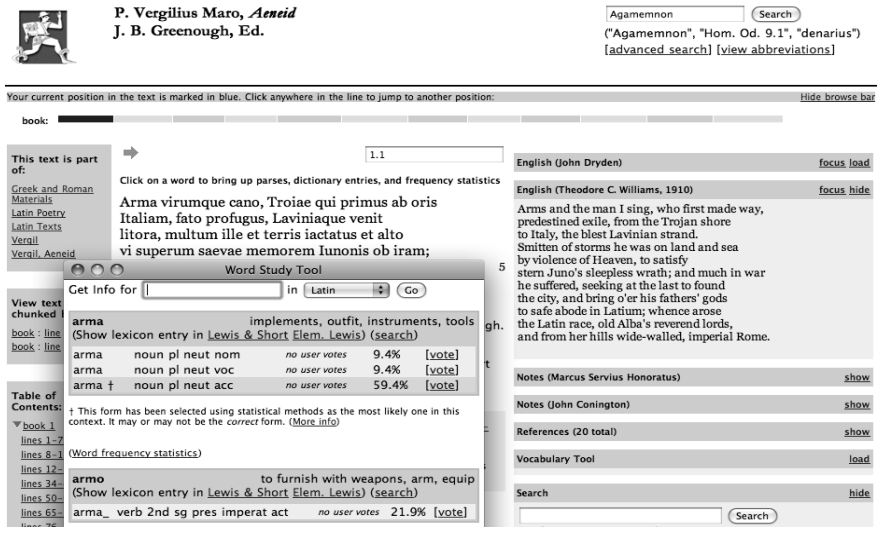

Figure 1: A screenshot of Vergil's *Aeneid* from the Perseus digital library.

Figure 1 shows a screenshot from the Perseus Digital Library. In this view, the reader is looking at the first seven lines of Vergil's *Aeneid*. The source text is provided in the middle, with contextualizing information filling the right column. This information includes:

- Translations. Here two English translations are provided, one by the 17th-century English poet John Dryden and a more modern one by Theodore Williams.
- Commentaries. Two commentaries are also provided, one in Latin by the Roman grammarian Servius, and one in English by the 19th-century scholar John Conington.
- Citations in reference works. Classical reference works such as grammars and lexica often cite particular passages in literary works as examples of use. Here, all of the citations in such reference works to any word or phrase in these seven lines are presented at the right.

Additionally, every word in the source text is linked to its morphological analysis, which lists every lemma and morphological feature associated with that particular word form. Here the reader has clicked on *arma* in the source text. This tool reveals that the word can be derived from two lemmas (the verb *armo* and the noun *arma*), and gives a full morphological analysis for each. A recommender system

automatically selects the most probable analysis given the context, and users can also vote for the form they think is correct.

The user interface of our library is designed to be modular, since different texts have different contextual resources associated with them (while some have translations, others may have commentaries). This modularity allows us to easily introduce new features, since the underlying architecture of the page doesn't change – a new feature can simply be added.

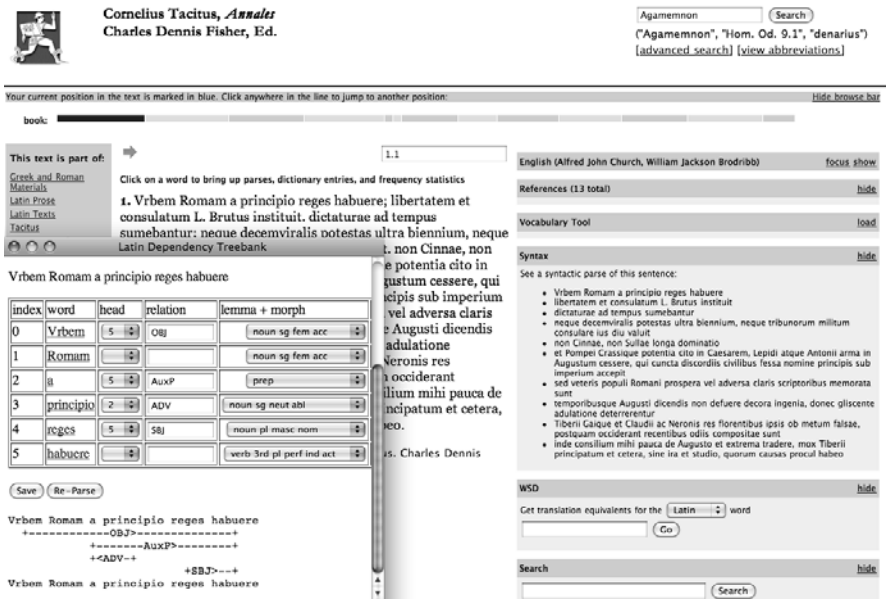

Figure 2: A screenshot of Tacitus' *Annales* from the Perseus digital library.

Figure 2 presents a screenshot of the digital library with a syntactic annotation tool built into the interface. In the widget on the right, the source text in view (the first chunk of Tacitus' *Annales*) has been automatically segmented into sentences; an annotator can click on any sentence to assign it a syntactic annotation. Here the user has clicked on the first sentence *(Vrbem Romam a principio reges habuere);* this action brings up an annotation screen in which a partial automatic parse is provided, along with the most likely morphological analysis for each word. The annotator can then correct this automatic output and move on to the next segmented sentence, with all of the contextual resources still in view.

Perhaps the most significant feature of this underlying modularity is the ability to extend existing services to new languages. The Perseus Digital Library has long provided these features for Greek and Latin, but we have recently extended them to Classical Arabic as well, as shown in Figure 3.

Figure 3: Arabic infrastructure in the Perseus Digital Library.

The services provided for Arabic are the same as those provided for Greek and Latin: a source text (here, the *Quran*) is shown in the middle of the page, with contextualizing information (three English translations) on the right. Each word in the source text is linked to its morphological analyses. As with our Greek collection, we provide several options for viewing the non-Latin character set, either as Unicode (shown here), or in the form of a Latin transliteration (using the Buckwalter transliteration scheme).

We can in principle extend these services to any language provided we have the following:

TEI-compliant XML texts governed by a common citation scheme. The document structure provided by TEI-compliant XML allows us to serve discrete chunks of a text to the user.[2] While we can break any XML document into smaller chunks, we are able to provide meaningful contextual information when multiple documents are organized according to the same scheme. In Classics, this citation scheme allows us to use a single identifier (such as Cic. Cat. 1.1) to refer to the

2 For more on the Text Encoding Initiative (TEI) standard, see `http://www.tei-c.org`

same section of text (the first section of the first speech of Cicero's *In Catilinam*) – whether a source text, a translation, or a commentary.[3] Like the chapter/verse citation scheme of the Bible, the sura/verse scheme of the Quran provides this same universal identifier; this allows us to automatically link chunks of a translation with its source text and present them together.

Morphological analyzer. Like Latin and Greek, Classical Arabic is a highly inflected language with an intricate morphological structure. In order to serve morphological analyses for all words in the source text, we first need an engine capable of generating those analyses. For Greek and Latin we use the Morpheus analyzer (Crane 1991, Crane 1998); for Arabic we use the Buckwalter Arabic Morphological Analyzer (Buckwalter 2002), translating its free text output into a format our infrastructure can utilize.

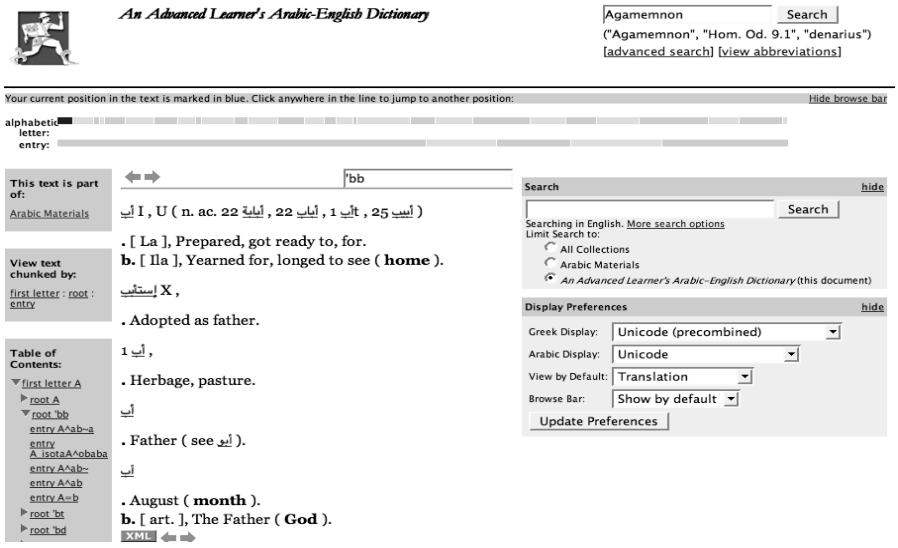

Figure 4: Arabic Lexicon in the Perseus Digital Library.

Lexicon. A prerequisite for morphological analysis is a list of lemmas from which each word can be inflected. Buckwalter's Arabic analyzer is based on its own lemma list (in which each lemma is tied to a short definition), but we have also digitized a full lexicon (Salmoné 1889) to provide more thorough contextual information for each word.

3 The Canonical Text Services (CTS) protocol also supports more sophisticated references, including individual spans of text within canonical chunks; see Blackwell and Smith (2005) and Porter et al. (2006).

While the core collection of the Perseus Digital Library has long been its Greek and Latin texts, the underlying architecture that enables this reading environment is not language-specific: any language with structured texts, a morphological analyzer and a lexicon can be brought into it to produce the same results.

The Perseus Digital Library is an example of the current generation of ancient corpora. The "structure" of its resources is found in the fact that they are all carefully curated by hand. The process that produces morphological analyses for all of the words in a source text may be automatic, as is the process of syntactic analysis, but they are both driven by human-created resources: a rule-based morphological analyzer (whose rules have been delineated by a human) and a syntactic parser trained on human-annotated data. The clean XML texts in our collection are also the products of careful curation – of human-corrected OCR and manual tagging. This structure allows us to create highly precise services for a small corpus of texts. Million-book collections, however, are orders of magnitude larger, and we cannot provide the same level of attention to all of the texts in those collections as we do to the several dozens in our own now. We can, however, leverage the services we have built on structured texts to exploit the unstructured ones now emerging.

3. Creating Cybereditions from Unstructured Collections

Million book libraries contain multiple editions of the same source work, often in the form of an OCR'd image book. By comparing these image books and their corresponding noisy OCR output to a clean, XML-structured edition within our digital library, we can mine new information: we can establish first if a work is an edition of a source text in a our digital library; we can project the XML markup of one corrected edition against the unstructured OCR output of an image book; and we can correct and collate multiple editions against each other to produce an automated textual history, which identifies the variants among different editions and the extent to which those editions relate to each other. By using technologies trained on tagged texts in our collection, we can annotate the new edition – identifying the named entities that it mentions and the quotations it references.

We need to extract machine actionable data from digitized print books to support new analytical and visualization services – simple information retrieval is the first and easiest step. Our current work concentrates particularly on the challenge of automatically collating and annotating multiple editions of the same author to create a "cyberedition." The result will include automatically generated histories of the textual traditions of canonical editions, but also other categories of scholarly information, such as lists of testimonia and specialized glossaries for individual texts. While some of the services continue traditional scholarly products, other outcomes have few predecessors. Scholars have, for example, created inventories of important readings and conjectures for particularly important texts, but we have never had

extensive databases that can compare multiple texts and arbitrary passages within texts against one another, analyzing their similarities and visualizing the results to show patterns of change and influence in the history of our textual sources.

This work is based on the following core tasks:

- Automatically collating multiple editions
- Quotation identification
- Named entity identification
- Translation identification and alignment
- User contributions

3.1 Automatic collation of multiple editions

The first generations of primary texts have generally included the single best edition available for inclusion – they were, in essence, anthologies on a massive scale, excerpting the source texts (but not the introductory materials, textual notes or other scholarly apparatus) from individual editions. Efforts such as Google Books and the OCA, however, already have begun to include multiple editions of the same work and they digitize the entire book – front matter, textual notes, indices, etc. as well as reconstructed text. This raises challenges as well as opportunities. The following illustrate fundamental, and in some cases, connected tasks that are important for scholarship but that are beyond the general services that companies such as Google will provide:

- *Identifying multiple editions of the same work:* In many cases we will have cataloguing records with which to locate multiple editions of the same work but catalogue records are uneven for scholarly editions and some smaller texts may be embedded in larger anthologies or monographs. We should be able to locate multiple editions of the same work by using one copy as the source for one or more queries: most texts have sufficient unique forms and phrases that provide a signature with which to find similar documents (the same approach applied in plagiarism detection).[4]
- *Classifying OCR output into categories such as introduction, textual notes, headers, source text, indices etc:* Even if we know that a book contains an edition of a particular work, we need to be able to separate out the core text from the other components of the edition. To some extent, the fact that a few major series publish most editions in classics makes this a more tractable task. We can invest a substantial amount of labor tuning page segmentation tools for Teubner, Loeb and Oxford Classical Text editions. In this case, we can use precision and recall to measure our results.
- *Reference scheme analysis:* This is a special case of classifying OCR output but its importance in humanities scholarship is such that it warrants separate study. Standard OCR software is not designed to recognize numbers in the margins with which scholarly editions often mark the citation schemes for their texts. Floating

4 See, for instance, Zaslavsky et al. (2001).

numbers often become alphabetic strings (e.g., "15" in the margin of a text indicating line 15 often becomes "is"). Citation schemes are critical components for advanced scholarly services and we need to be able to measure how well we are able to identify citation markers and then apply these to the text (e.g., a marginal "6" indicates that the beginning of section six occurs somewhere in the line of text next to it, but leaves it to the reader to determine precisely where that break occurs).
• *Correcting and collating multiple editions against each other:* If we have multiple versions of the same text, we can compare them against each other to serve at least three goals. First, by identifying passages of text that are identical in two editions, we can help distinguish the restored text from notes, headers, etc. Second, where the restored texts of two editions differ and one of those differences does not generate valid Latin morphological analyses, then we usually have detected a data entry error. Third, if two editions differ but the differences in both editions generate valid morphological analyses, then we probably have intentional editorial variants.
• *Multiple editions and noisy OCR output:* In our work we focus on comparing the output from 19th and 20th century editions in clean print. We will, however, ultimately want to compare texts in many different sources that do not lend themselves to conventional OCR (e.g., early modern editions and manuscripts). We will need to explore the question of how well we can use a clean text to augment the results of noisy OCR. If, for example, we are able to extract 30% of the words from a page of Livy, how well can we align these with words in clean text?
• *Markup projection:* We have careful TEI markup for many editions. To what extent can we project precise XML markup from one edition onto another? The most important case of this problem involves citation schemes: if we have marked all the line or section breaks in one edition, we want to be able to find those lines/sections in many other editions. This particular task becomes especially important when there are multiple citation schemes for the same work and the original page images may not contain the line/section markers on which other reference works depend.
• *Automated textual histories:* If we can identify variants among editions, we can measure the extent to which different editions relate to one another: e.g., which editions as a whole or which readings in particular have been most influential? To what extent can we see texts stabilize over time (as editions converge) and/or texts that remain in flux?

3.2 Quotation Identification

Many quotations in secondary sources appear without recognizable citations. This is especially true in the vast body of literary and cultural texts outside of formal academic publications which cite and refer to ancient authors. Consider the following quote from a "tribute to Confederate heroes" in the *Southern Historical Society* papers.

> Do you forget the Pagan saying that reconciles so many readers of history to the fall of the noblest States and the defeat of the truest heroes, *Victrix causa Deis placuit, sed victa Catoni,* or the cynical paradox of the French Empire, that "Heaven is on the side of the bigger battalions?" *Southern Historical Society* 10 (1882) 562-563.

The excerpt above illustrates Southern attitudes towards the defeat of the Confederacy. The passage quotes, but does not translate, a line of Latin and describes it only as the "Pagan saying." In fact, the quotation is line 128 from Book 1 of the Roman poet Lucan's *Pharsalia* and means "the victorious cause pleased the gods but the defeated cause pleased Cato." Cato died fighting Caesar and defending the Roman republic. The quotation, when recognized, reflects the idea that the Southern cause was defending republic virtue against the imperial despotism of Northern power. Identifying quotations like this allows us to track the influence of particular texts over time and also lets us include testimonia about a given work as part of its edition.

While in simple cases this is a problem of string matching, Classical languages in particular present several search problems:

- Since Greek and Latin are highly inflected languages, they have an intricate case structure, and a string which might appear (for example) in the accusative case in a source text (e.g., *filium Dei,* "son of God") may appear in the nominative case *(filius Dei)* as a quotation in another text;
- Greek and Latin's word order is relatively free, and the words in a quotation might appear in a different order (*iacta alea est,* "the die is cast") than in the source text *(alea iacta est)*.
- Either the source text or the quotation may have textual errors (especially given that both are likely the results of imperfect OCR).
- There may be minor editorial differences between the source text and the quotation. For example, Lewis and Short's Latin Dictionary cites Caesar's *B.G.* 4.22.4 as *naves … **ab** milibus passuum **octo** vento tenebatur* ("ships ... were held by the wind 8 miles away"), while the source text edited by T. Rice Holmes contains ***a** milibus passuum **VIII** vento tenebantur*.

3.3 Named Entity Identification

While specialists may be familiar with the people, places, organizations, technical terms and other entities in our particular disciplines, many research questions bring us into contact with documents that assume very different backgrounds from those that we possess. In the example from the Southern Historical Society papers, we needed not only an understanding of Latin but also a knowledge of which Cato (the Elder or Younger) was meant and how that Cato died fighting for the Roman

republic. Named entity analysis allows us automatically to identify, with increasing accuracy, which Cato a particular passage denotes. Once we can distinguish references to the Elder and Younger Catos, we can generate links to articles in online reference works and/or generate a list of passages that cite our Cato, organizing the search results into labeled clusters.

While very large collections will contain ever more passages referencing ambiguous names such as Antigonus, Smith, Alexandria, Washington, York, etc., we can turn that volume to our advantage by organizing Alexandrias in passages with similar vocabulary into distinct clusters, then looking for disambiguated referenes within those clusters (e.g., "Alexandria, VA," "Booker T. Washington") to determine the likely referents for the unspecified Alexandrias and Washingtons.

At present we have implemented scalable named entity services for English language documents, including translations of Greek and Latin texts.[5] We have supported automatic analysis of place names and dates within the Perseus collections since 2001. In the past several years, we have extended our coverage to include names, numerical quantities of various types (monetary, weight, volume, distance, etc), time of day, etc.

Automatic background information services include:

Term identification: Many organization names (e.g, "the New York Times"), technical terms *(patres conscripti)* and other phrases are sufficiently distinct that we can simply look them up in a list. Automatic linking has been a staple service in Perseus for a decade. In 2002, the National Science Digital Library provided Tufts and Johns Hopkins with a grant to create a scalable, open source version of this automatic linking service, which could quickly recognize in full text examples from very large, customizable lists of terms and then generate links from these to the relevant glossary, encyclopedia or other background articles. The Services for a Customizable Authority Linking Environment (SCALE) package was completed in late 2005.[6]

Semantic classification: Many proper nouns are semantically ambiguous – especially in American English, which makes relatively little use of semantic classifiers: thus we have more Jacksons than Jacksonvilles in the United States. For some purposes, classification is sufficient: we may be interested in analyzing the types of names conferred on people or places at particular periods.

Identification: If June is a month, then it is "June 1834," "June 1918" or some other year? If Washington is a place, is it Washington, PA, Washington, NC or another Washington? If 1.33 is a citation, is it "Thuc. 1.33," "Hom. Od. 1.33" or some other citation?

5 See Crane and Jones (2006) and Babeu et al. (2007) for more on this work.

6 `http://nils.lib.tufts.edu/scale`

3.4 Translation identification and alignment

We want to be able to identify and align as many translations of Greek and Latin source texts as possible. Multiple translations are useful for human readers and will be especially helpful to researchers from outside of classics who are working with Greek and Latin. A cyberinfrastructure, however, should be able to identify and analyze translations for at least two major services:

- *Automatic glossing and phrase translation:* While we may not have access to enough parallel text for advanced machine translation, we can use parallel text analysis of source and translations to determine probable English equivalents to particular words in particular contexts: e.g., whether the Greek word *archê* corresponds to "empire" or "beginning" or one of its other meanings. Such services can be especially useful for the vast body of Latin that has not and never will be translated into English.
- *Semantic analysis:* If we can identify particular passages where Greek and Latin words have distinct senses, we open new methods with which to generate customized glossaries and to analyze broad patterns of word usage and content across very large bodies of text. Classics has a long history of producing word-level commentaries for its major source texts, but the presence of translations will allow a cyberinfrastructure to automatically create them using techniques borrowed from computational lexicography.[7]

We have done initial work on parallel text analysis and word translation for Greek and for Latin to create bidirectional sense inventories, such as that shown in Figure 5.

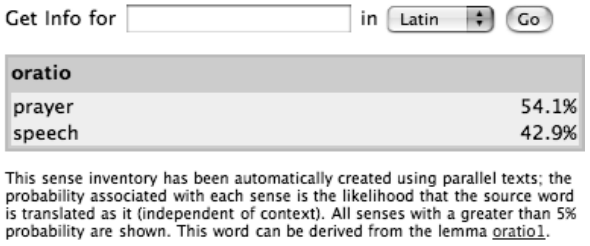

Figure 5: Translation equivalent for the Latin word *oratio*.

This type of parallel text analysis can be seen as a cascading sequence of alignments: first you identify that one document is a translation of the other, then align them on a section level, sentence level and finally word level. If one text is found to be a translation of another, alignment at the sentence and word level can be accomplished quite readily with open-source tools (e.g., Moore's Bilingual Sentence Aligner (Moore 2002) at the sentence level and GIZA++ (Och and Ney 2003) at the word level).

7 See Bamman and Crane (2008).

Our main focus now is on that first step: translation detection. This is, in turn, a special case of the cross language information retrieval problem with entire documents as the initial query. We anticipate that should be able to find translations of large documents with a high degree of success, which will then let us progress to the further steps of sentence and word level alignment.

Finding translations of individual poems or passages is a harder version of the quotation detection problem and will be more challenging. Our earlier work in developing translation inventories for language pairs will significantly help in this task, however: one variable to include in such a system will be the probability that a word in a possible quotation is a translation of a word in the source text, which (as Figure 5 demonstrates) we are able to calculate given our existing word alignments.

3.5 User contributions

In the course of the core tasks above, we are examining three ways in which this cyberinfrastructure can interact with users, for two main reasons: users – both expert textual editors and decentralized contributors – can help initially create and then continuously update these editions, and we must also be able to present these works to them in a way that is tailored to their specific individual needs. While we have not done as much work on these user contributions as on the services above, we have created substantive prototypes for all three.

- *Mining pre-existing knowledge bases:* In the simplest case, we will sometimes have access to a carefully edited electronic text of which we have multiple other versions available as part of a large collection of image books with uncorrected OCR. In this case, the "user" is the original textual editor, and we need to see how well the transcript in one version of a text allows us to correct multiple other editions that may have editorial variants. At a more complex level, we need to measure our ability to project markup from one edition to another: if we have speakers marked in one edition of Plautus or proper nouns in an edition of Tacitus, how well can we match that markup with other editions? Even if we can perfectly project markup from text 1 to text 2, how often do editors differ on such annotation? Other pre-existing knowledge bases, both born digital (e.g., training sets for morphological analysis) and derived from print sources (e.g., digitized lexica and encyclopedias) have more complex data that can be mined and exploited for error correction, named entity analysis and other services.

- *Attracting decentralized contributions in structured formats:* We want to be able to accept contributions from individual users resolving problems of named entity identification or other readily computable classification tasks. We have already developed systems to collect such annotations for linguistic data, such as morphological

tagging, sense disambiguation, and syntactic parsing.[8] We need to evaluate more closely the results that we are already receiving and refine/expand these methods.
• *Automatic customization and personalization of data:* Our work on tailoring information for individual users has focused on two functions, both primarily pedagogical in their initial implementation. First, we have developed models of linguistic knowledge based on various textbooks for Latin and Greek and used these models to identify new terms in unseen texts. Second, we have used existing weblogs as data for a recommender system to support reading. Once a reader has asked questions about four words, we can predict two thirds of all subsequent questions in a given passage.[9] Both of these areas of research point towards more sophisticated applications to support scholarship.

4. Conclusion

As new texts are added to the Perseus Digital Library, they are subjected to a variety of automatic processes – a morphological analyzer inspects each source word and presents a list of possible parses, while a tagger selects the most probable one based on the other texts in our collection; a named entity analyzer that has been trained on these texts does the same for all proper names found therein. All Greek or Latin source words are linked to their respective dictionary entries, and all canonical citations are linked to their source text. Every time a new text is added, it is analyzed by systems that have been trained on the texts that are already there, and it becomes part of the infrastructure itself.
The million book projects that are now emerging have the potential to significantly transform these processes by their sheer volume alone. We have been able to make great progress with a Classical collection of nine million words but we stand to go much further with a collection several hundred times that size. These projects are large but general and focus upon generic services. Our work lies in the gap between these generic services and the needs of advanced research. By using the structured resources we already have in our digital collections as training material for automatic processes, we can exploit these emerging million book collections in a manner that addresses the needs of the scholarly community.

Bibliography

ACLS Commission on Cyberinfrastructure. 2006. *Our Cultural Commonwealth: The Final Report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences.* http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

Babeu, Alison — David Bamman — Gregory Crane — Robert Kummer and Gabriel Weaver. 2007. Named Entity Identification and Cyberinfrastructure. In: *Proceedings of*

8 See, for example, Bamman and Crane (2007).
9 Both of these technologies are described in Crane et al. (2007) and Crane et al. (2006).

the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), pp. 259-270.

Bamman, David and Gregory Crane. 2007. The Latin Dependency Treebank in a Cultural Heritage Digital Library. In: *Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH)*, pp. 33-40.

Bamman, David and Gregory Crane. 2008. Computational Linguistics and Classical Lexicography. *Digital Humanities Quarterly* (forthcoming).

Blackwell, C. and N. Smith. 2005. A Guide to version 1.1 of the Classical Text Services Protocol. Digital incunabula: a CHS site devoted to the cultivation of digital arts and letters. `http://chs75.harvard.edu/projects/diginc/techpub/cts-overview`.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, Philadelphia.

Crane, Gregory. 1987. From the Old to the New: Integrating Hypertext into Traditional Scholarship. In: *Hypertext '87: Proceedings of the 1st ACM Conference on Hypertext,* pp. 51-56. ACM Press.

Crane, Gregory. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing* 6(4), pp. 243-245.

Crane, Gregory. 1998. New Technologies for Reading: The Lexicon and the Digital Library. *Classical World* (4): 471-501.

Crane, Gregory and Alison Jones. 2006. The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries,* pp. 31-40, Chapel Hill, North Carolina.

Crane, Gregory — Robert F. Chavez — Anne Mahoney — Thomas L. Milbank — Jeffrey A. Rydberg-Cox — David A. Smith and Clifford E. Wulfman. 2001. Drudgery and Deep Thought: Designing Digital Libraries for the Humanities. *Communications of the ACM,* 44(5), pp. 34-40.

Crane, Gregory — David Bamman — Lisa Cerrato — Alison Jones — David M. Mimno — Adrian Packel — David Sculley and Gabriel Weaver. 2006. Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. In: *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL),* pp. 353-366.

Crane, Gregory — David Bamman and Alison Jones. 2007. ePhilology: When the Books Talk to Their Readers. In: Ray Siemens and Susan Schreibman (eds.), *Blackwell Companion to Digital Literary Studies* (Oxford: Blackwell, 2007).

Gietz, P., et al. 2006. TextGrid and eHumanities. In: *E-SCIENCE '06: Proc. of the Second IEEE International Conf. on e-Science and Grid Computing,* pp. 133–141. IEEE, Wash., D.C.

Moore, Robert C. 2002. Fast and accurate sentence alignment of bilingual corpora. In: *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation,* pp. 135–144.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29.1: 19–51.

Porter, D., W. Du. Casse — J. W. Jaromczyk — N. Moore — R. Scaife, and J. Mitchell. 2006. Creating CTS collections. *Digital Humanities,* pp. 269-274.

Salmoné, H. Anthony. 1889. *An Advanced Learner's Arabic-English Dictionary.* Librairie du Liban, Beirut.

Zaslavsky, A. — A. Bia, and K. Monostori. Using copy-detection and text comparison algorithms for cross-referencing multiple editions of literary works. In: P. Constantinopoulos and I. T. Solveborg (eds.), *Proceedings of the 5th European Conference on Digital Libraries (ECDL),* pp. 103–114, Darmstadt, Germany, 2001. Springer.