

Repositories, Cyberinfrastructure and the Humanities

Gregory Crane
Tufts University

A report, funded by the ACLS and the Mellon Foundation, supports the creation of a cyberinfrastructure for the humanities and social sciences. Two interrelated questions now arise. First, what should such a cyberinfrastructure look like? Second, how will such a cyberinfrastructure affect the role – and the impact -- of humanities within society as a whole?

The second question is important because we in the humanities currently lack the resources to create the infrastructure that we will need. Even if we build on the infrastructure that the well-funded scientific disciplines create for themselves, bridging the gap between that infrastructure and the needs, both present and emergent, of the humanities will require substantial investment over the many years – we need to reinvent in digital form an intellectual life shaped by print. If we are to attract the material resources that we need to pursue our most advanced and challenging work, we need to redefine the relationship between academic humanists and society as a whole. Thus, even if our interests are wholly focused upon specialized research, we need a cyberinfrastructure that extends the intellectual reach of expert and novice. In this, professional academics serve their own intellectual interest, for, however sophisticated we may be in our own specialties, we will always be novices in most areas. If humanists argue that we train our students to think critically, then classicists, for example, should have the critical skills necessary to work with classical Chinese in a developed cyberinfrastructure.

We have good initial data to address the first question. We cannot predict what form a mature cyberinfrastructure will assume over the coming years, but we know very well some of the basic services that such a cyberinfrastructure must contain. Our predictions are based on hind-sight: support from the IMLS, NEH and NSF has allowed us to build versions of each service and made some of them available as standard features on a public digital library. What we propose thus reflects technology that is already available and for which an audience exists. The services outlined below need to shift from research and development and to become part of established infrastructure. They therefore constitute a minimal set of operations and should be a part of any repository that serves the humanities.

Four basic classes of service emerge: 1) catalogue services identify the discrete objects within a collection (editions of Vergil's Aeneid, books about Vergil); 2) named entity services identify semantically significant objects embedded within collection objects (references to Vergil or the Aeneid within other documents); 3) customization and personalization (given a particular passage of the Aeneid, what would be of interest to an intermediate student of Latin vs. a professional Latinist?); 4) structured user contributions (e.g., users tell the library that a particular word in a passage of Vergil has a particular sense or plays a grammatical role in the sentence). Summarization,

visualization, machine translation and other technologies all play roles within one or more of the service layers above.

1. Catalogue services:

Generations of librarians have provided a foundation on which to build but we must go further than traditional catalogues. The Functional Requirements for Bibliographic Records (FRBR) data model is an important step forward, for it provides an elementary framework within which we can begin to represent some of the basic knowledge structures that experts have developed to describe texts. A canonical work such as the Vergil's Aeneid has appeared in hundreds – and probably thousands -- of versions, all of which strive to represent a single edition (the text that Vergil left at his death) but errors crept into subsequent copies and each attempted reconstruction may differ from every other version ever produced. The Aeneid has been translated into dozens of languages, with each translation based on one or more editions. The Aeneid has attracted commentaries – documents that contain annotations about particular word, phrases and sections of the Aeneid.

The screenshot displays a digital library interface for Thucydides' *The Peloponnesian War*. At the top, the title and author are shown next to a small icon of a runner. Below this is a progress bar indicating the current position in the text. The main content area shows a text chunk starting with "86. 'The long speech of the Athenians I do not pretend to understand. They said a good deal in praise of themselves, but nowhere denied that they are injuring our allies and Peloponnesians. And yet if they behaved well against the Medes then, but ill towards us now, they deserve double punishment for having ceased to be good and for having become bad. [2] We meanwhile are the same then and now, and shall not, if we are wise, disregard the wrongs of our allies, or put off till tomorrow the duty of assisting those who must suffer today. [3] Others have much money and ships and horses, but we have good allies whom we must not give up to the Athenians, nor by lawsuits and words decide the matter, as it is anything but in word that we are harmed, but render instant and powerful help. [4] And let us not be told that it is fitting for us to deliberate under injustice; long deliberation is rather fitting for those who have injustice in contemplation. [5] Vote therefore, Lacedaemonians, for war, as the honor of Sparta demands, and neither allow the further

Navigation and reference options include:

- Table of Contents**: A sidebar menu with links for Book 1, chapters 1 through 11.
- Table of Contents**: A top navigation bar with a search box containing "Thuc. 1.86" and a "Table of Contents" link with left and right arrows.
- Translations and Notes**: A list of links for different versions, including Greek, English (Thomas Hobbes), English (Benjamin Jowett), Notes (E. C. Marchant), and Notes (Charles D. Morris), each with a "focus load" option.
- Places**: A section titled "Places (automatically extracted)" with a "hide" link. It includes a sorting instruction and a list of places: Sparta (Canada) (1), Peloponnesians (Greece) (1), Medes (Italy) (1), and Athens (Alabama, United States) (1).
- References**: A section titled "References" with a "hide" link. It states "Found 31 references related to this page." and lists cross-references to this page, including:
 - Herbert Weir Smyth, *A Greek Grammar for Colleges*, [DATIVE OF INTEREST](#)
 - Herbert Weir Smyth, *A Greek Grammar for Colleges*, [NEGATIVE \(PROHIBITIONS\)](#)
 - Herbert Weir Smyth, *A Greek Grammar for Colleges*, [VERBAL ADJECTIVES IN -τος](#)
 - Raphael Kühner, Bernhard Gerth, *Ausführliche*

Figure 1: Information about a canonical chunk of text: Thucydides, book 1, chapter 86. Note that display integrates access to content and a catalog of other versions of, or relevant to, the same chunk of text.

The FRBR data model allows us to identify and organize all editions, translations, commentaries, indices, and other documents focused on Vergil's Aeneid. But we need deeper granularity than FRBR's manifestations of expressions of a work. Scholars have

established canonical citation schemes so that they can describe the same chunk of text as it appears in many different editions. Few students and fewer scholars actually want all information about the Aeneid or any heavily studied canonical works of literatures – such works are almost fields unto themselves and no one can read, much less digest, all that has been written about them. In our day-to-day work, we examine subsets of these texts. We might adopt a breadth-first approach and examine a topic that runs through the text – e.g., a particular word or image or theme. Or we might focus in depth on a passage and explore many different themes relevant to it. In each case, we are looking at defined subsets of these documents.

Scholars have established canonical citation schemes as coordinate systems to map their texts. Figure 1 resembles a standard text display but it illustrates, instead, the results of a minimal catalogue to a modest collection. The user has not request information about Thucydides' History of the Peloponnesian War but about Thucydides, History of The Peloponnesian War, book 1, chapter 86. Notice that the text includes numbered sections as a third level of granularity. The users could drill down and select one of these sections as the object of interest. A mature system should be able to catalogue information about every word and every combination of words within and across each canonical chunk of text.

In the humanities, catalogues thus need to include not only books but the canonical documents within books. We need a catalogue that manages the canonical citation schemes and can extract from an open set of documents, versions of and information relevant to the same logical container. We also need intelligent version analysis and visualization within our catalogues: given N editions of a work, how does each edition relate to those which precede it? Which editions were most influential? What (if anything) is different in a new edition?

Data sources for cataloguing

Cataloguing thousands of citations in hundreds of editions and translations of canonical reference works by hand is not practical. We must depend upon automatic alignment, cross language information retrieval, and markup projection from one text to the other. To drive these processes we should have at least one carefully transcribed version of each canonical text in each major citation system. These base texts can then serve as the anchors around which to discover the many other editions, translations, and commentaries that will surface in very large, emerging collections and then to align these documents to a common citation scheme.

2. Named entity services

We may for the sake of argument assume that catalogues provide access to well-defined objects within a collection. We also need to be able to locate references to, and then summarize information about, named entities that appear within the contents of our collections.

Named entities can be documents (e.g., references to Thucydides' History of the Peloponnesian War), citations within documents, people, places, organizations, events

and the other topics for which we consult catalogues, encyclopedias and gazetteers. They also include linguistic topics as well: the word *facio* is a dictionary heading for the Latin word “to do, make” and is thus a named entity that integrates inflected forms such as *fecisset*, *factus* etc. Every word sense in a dictionary and linguistic phenomenon in a grammar is a separate named entity. Every subject heading or topic to which we assign a label is a named entity.

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Figure 2 -- Named entity identification: places from a chapter of a book about the US Civil War.

Figure 1 also includes a list of place names automatically extracted from the text on the left and linked to places in world. The results in this figure illustrate a technical challenge: three of the four places are incorrectly identified because proper nouns are semantically ambiguous (e.g., *Mede* – an ethnic name in Thucydides – is also a place name) and place names can describe many different places (there is a *Sparta* in Canada and an *Athens* in Alabama). In practice, place names are relatively easy to find and identify in classical texts (the normal success rate is c. 95%). Figure 2 illustrates place names extracted from a chapter on the American Civil War as plotted using Google maps. In January 2007, Google released its own service to map places from digitized books in Google book search. The goal must be for customize the Google results, making it possible to substitute more accurate services.

Data Sources for Named Entity Identification

These include language models calculated from unstructured articles about particular entities, structured data extracted from print gazetteers, machine readable dictionaries, and other existing knowledge sources, born digital resources such as WordNet, and labeled training sets (which may be lists of passages where named entities are tagged to a high degree of accuracy and which may in turn be mined from print indices). Reference works from print thus are capital resources in a digital library, providing the foundational data for many of the higher level services on which intellectual life depends. Automatic clustering and discovery of entities are crucial instruments but unlikely to provide the best results on their own. Converting print information about the past into machine actionable knowledge is the greatest task that the rising generation of humanists confront.

3. Customization and Personalization

Once we are able to identify most of the objects and named entities in our collections, we need to use this information to increase intellectual, as well as physical, access. In print libraries, a book in Greek is useless to a reader who has not studied Greek. In a modern digital library, machine translation and a host of translation aids should provide basic access to the novice with no Greek and to extend the capacity of those studying the language at all levels to draw meaning from the text.

Figure 3 illustrates a simple approach to customization of vocabulary. The user has developed a profile based on his or her text book of Latin. The system automatically compares that profile against the words it detects in a given page, then identifies which words the user probably has and has not encountered before.

C. Suetonius Tranquillus, *Caligula*
Maximilian Ihm, Ed.

[Study vocabulary in this passage.](#)

[Table of Contents](#) ↔

Click on a word to bring up parses, dictionary entries

This text is part of:
[Greek and Roman Materials](#)
[Latin Prose](#)
[Latin Texts](#)
[Suetonius](#)

View text chunked by:
[life](#) : [chapter](#) : [section](#)

Table of Contents:
[▶ Divus Iulius](#)
[▶ Divus Augustus](#)

[XML](#) ↔

Your vocabulary profile:
Wheelock (5th)
 Wheelock, Frederick M., *Wheelock's Latin (5th Edition)* (1990)

This passage contains **115** possible dictionary forms.
 According to your vocabulary profile, you have already learned **54** of
 This page displays the **61** remaining dictionary forms.

[Customize your vocabulary profile](#)

	Frequency	Dictionary Form	Short Definition
	2	contio	a meeting, assembly, convocation, gathering, audience
	1	acerbitas	bitterness, harshness, sourness
	1	armatus	armed, equipped, in arms
	1	armo	to furnish with weapons, arm, equip
	1	atrocitas	fierceness, harshness, enormity
	1	augustus	consecrated, sacred, reverend
	1	Augustus	a cognomen given to Octavius Caesar as emperor, his majesty
	1	circumdo	to place around, cause to surround, set around
	1	cogitatio	a thinking, considering, deliberating, thought, reflection, meditation
	1	confestim	immediately, speedily, without delay, forthwith, suddenly
	1	contrucido	to cut to pieces, cut down, put to the sword
	1	de	down (adv.)
	1	decedo	to go away, depart, withdraw, retire

Figure 3 – Customization. The digital library recognizes that the user has encountered 54 of 115 dictionary words in a given passage

The example above is fairly simple but the underlying principle is fundamental. The system asks (1) what it knows about its own contents, (2) what the user already knows, and then (3) customizes the results for the immediate needs of this particular user.

Data sources for customization and personalization

We need profiles with structured data representing what and when users have encountered particular topics. Named entities are a natural starting point because we already assume services to identify named entities. We also need log data from which we can identify usage patterns. We need recommender systems similar to those familiar from Amazon and other e-commerce sites (“users who book book A also bought books B and C”) but applied to academic issues (e.g., “readers who looked up words X, Y, and Z, also were interested in words M, N, and O”).

4. Structured User Contributions

We need not only new methods to acquire traditional publications but also much more granular contributions: e.g., “bank” in passage X represents “river bank” rather than “financial institution”; Washington in passage A is Washington, DC, but George Washington in passage B.

The screenshot shows a web interface for studying Latin text. At the top, it identifies the text as "P. Vergilius Maro, Aeneid" edited by "J. B. Greenough, Ed.". Below this, there is a search tool titled "Word Study Tool" with a search bar and a "Go" button. The search results for the word "saucius" are displayed in a table, showing various morphological analyses and their corresponding probabilities. The table is as follows:

Word	Part of Speech	Number	Gender	Case	Probability	Action
saucia	adj	pl	neut	nom	14.1%	[vote]
saucia	adj	pl	neut	voc	13.8%	[vote]
saucia	adj	pl	neut	acc	13.9%	[vote]
saucia_	adj	sg	fem	abl	13.9%	[vote]
saucia	adj	sg	fem	nom	14%	[vote]
saucia	adj	sg	fem	voc	13.6%	[vote]

Below the table, there are "Word Frequency Statistics" for the word "saucius" across different corpora:

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	33	0.54	18	0.30	Latin Poetry
3414041	99	0.29	71	0.21	Latin Texts
83620	8	0.96	7	0.84	Vergil
63770	8	1.25	7	1.10	P. Vergilius Maro, Aeneid

The interface also shows a "Table of Contents" on the left side, with links to "Book 1", "Book 2", "Book 3", and "Book 4". The main text area displays a passage from the Aeneid, with the word "saucia" highlighted in blue. The search tool also shows a "Word Frequency Statistics" table for the word "saucio" (to wound, hurt) with the following data:

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	19	0.31	4	0.07	Latin Poetry
3414041	42	0.12	14	0.04	Latin Texts

Figure 4 – Automated systems have enumerated all possible morphological analyses of a given form and then ranked their probability in a given context. Users can then vote on what they think the correct interpretation is.