

# The Humanities in a Global e-Infrastructure: a Web –Services Shopping-List

Gregory Crane, Brian Fuchs, Dolores Iorizzo  
Perseus Project, Tufts University, Imperial College Internet Centre

This paper asks fundamental questions about requirements for any global e-infrastructure that supports large-scale text-based next generation services for digital libraries; including text retrieval and analysis; morphological, multilingual, and computational linguistic tools; as well as catalogue services, named entity services, customization and personalization services and structured user contribution services - including annotations, corrections and cross references. We also look at how such an infrastructure may affect the role and the impact of humanities within society.

What should a global e-infrastructure look like that supports text and web-based services in large-scale digital libraries? How will such an infrastructure affect the role and the impact of humanities within society?

These two questions, only a few years ago, would have seemed presumptuous; now they have become required reflection for anyone serious about the future of large-scale digital libraries that represent the collective knowledge of all human history. Members of the CHLT and Perseus projects have been committed to pushing the boundaries of technology in support of research driven scenarios in the humanities and now we have been tempted to read the tea leaves. What we present here is a basic 'shopping list' of must-have web-services that any global infrastructure for the Humanities would have to support, and that would surely apply to any text-based digital repositories in the sciences and industry.

Our starting point was funding to support the Cultural Heritage Language Technologies Project<sup>1</sup>, sponsored by the National Science Foundation and the European Commission, which offers proof that a modest investment in international collaboration can bring substantial gains to the global community. CHLT delivered on its original aims of providing access to rare and fragile resources aided by multilingual translation and information retrieval tools, so that specialist as well as non-specialist readers can now analyze and understand texts in difficult languages, thereby allowing new kinds of scholarship through the use of computational tools for the study of style, grammatical form, word-profiling, citation linking, and statistical

data, all integrated into a single reading environment which offers innovative visualization tools that cluster search results into conceptual categories. CHLT results led us to focus on much larger questions surrounding the service requirements for a global e-infrastructure which support semantic interoperability, authority services, annotation, and computational linguistics that would also have applications in the sciences, industry and society.

Of course there are many others who share our concerns. In the US, Lagoze points out in 'What is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL'<sup>2</sup> that many of the key elements essential to the development of digital libraries remain undeveloped, including a public key infrastructure, semantic interoperability and ubiquitous persistent identifiers. Borgman<sup>3</sup> reiterates that a public infrastructure is needed to support an active *collaborative* environment for users that encourages annotations, reviews and additions to texts/images in digital repositories, as well as providing a *contextual* expanding web of interrelations and layers of knowledge that extend beyond primary sources - expressing 'the wisdom of crowds'. The Global Grid Forum and HASTAC are also working away at the coal face. In Europe efforts are underway to build an e-infrastructure for the humanities, some of which build upon e-Science infrastructure such as TextGrid and DILIGENT, others begin with user needs specific to the humanities, such as ECHO and DARIAH.

The current state of play is that we lack sufficient resources in the humanities to create an e-infrastructure that will support future generations

of users. Even if we build upon the established e-infrastructure of the scientific disciplines, such as EGEE, bridging the gap between the needs of science and those of humanities will require substantial investment over many years. We are faced with the challenge of reinventing in digital form an intellectual life that has been long shaped by print. If we are to attract the funding required 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 an e-infrastructure that extends the intellectual reach to both expert and novice. Professional academics are good at promoting their own intellectual interests, yet however sophisticated we may be in our own specialties we will always be novices in most areas. If humanists claim to train students to think critically, then classicists, for example, should have the critical skills necessary to work with classical Chinese in an e-infrastructure that is supported by multilingual tools, applications and web-services.

We have good initial data to address the first question, 'What should an e-infrastructure for the humanities look like?'. Although we cannot predict what form a mature e-infrastructure will take over the coming years, we know very well some of the basic services that such an e-infrastructure must support. Our predictions are based on hind-sight: support from the IMLS, NEH, NSF and EC has allowed us to build versions of each service and make some of them available as standard features in digital libraries. What we propose to implement is technology that is tested and readily available to an audience that already exists. The services outlined below constitute a *minimal* set of services that should be a part of any respectable repository, and hence need to become part of an established global e-infrastructure.

Now for the shopping-list. As we see it, four basic classes of service are required: 1) *catalogue services* that identify the discrete objects within a collection (editions of Vergil's *Aeneid*, books about Vergil); 2) *named entity services* that identify semantically significant objects embedded within collection objects (references to Vergil or the *Aeneid* within other documents); 3) *customization and personalization services* (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 contribution services*, including annotations, corrections and cross references (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, this information then becomes part of the digital collection). Summarization, visualization, machine translation and other technologies all play important roles within one or more of the service layers above.

### 1. Catalogue services

Generations of librarians have provided a foundation on which to build collections but we must go further than traditional catalogues. The Functional Requirements for Bibliographic Records (FRBR) data model is an important step forward, since 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 words, phrases and sections of the *Aeneid*.

The FRBR data model allows us to identify and organize all editions, translations, commentaries, indices, and other documents focused on Vergil's *Aeneid*. Yet 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 then 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 well 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 of a modest collection. The user has not requested 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 at a third level of granularity. The users could drill down and select one of these sections as an 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 are most influential, both currently and historically? What (if anything) is considered 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 many other editions, translations, and commentaries that will surface in very large, emerging collections, which must then be aligned 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. Yet, 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 might 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 also a named entity.

*Figure 1* includes a list of place names automatically extracted from the text on the left which are 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%). We have extracted place names 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 should be to capitalize on Google results, making it possible to substitute more accurate services.

### **Data Sources for Named Entity Identification**

Data sources for named entity identification 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. The greatest challenge for the next generation will be to convert print information about the past into machine actionable knowledge.

### **3. Customization and Personalization**

Once we are able to identify most of the objects and named entities in our collections, we will want 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 next generation digital libraries machine translation and a host of multi-lingual tools should allow the novice with no knowledge of Greek to draw meaning from a text, yet the student and the expert will be able to interrogate the digital resource at a deeper level.

*Figure 2* illustrates a simple approach to customization of vocabulary. The user has developed a profile based on his or her text book in Latin. The system automatically compares that profile against the words it detects in a given page and then identifies which words the user probably has and has not encountered before. The example 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. This is an ideal use of customization that could be used across the board in a variety of applications in science and industry.

#### **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 employ 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 bought book A also bought books B and C”) but applied to academic users as well (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..

One of the most fundamental activities in humanities research is the close examination, commentary and interpretation of texts in multiple languages; understanding the meaning of a text requires a broad knowledge of its historical, cultural, geographic, and linguistic context. The greater the complexity, the greater the challenge of interpretation, and the multiplicity of competing points of view. The interpretation of a single text inevitably turns outward to other texts, each of which carries a network of relations, which in turn points to other resources and other relations, in an ever expanding network of discovery, annotation, comparison, and analysis. This iterative process underpins all interpretations and the social construction of meaning. Voting on which interpretation is correct will become a common feature of next-generation digital technology, and it is our responsibility to create an e-infrastructure that supports this contextual and interactive level of interrogation.

### **Conclusion**

How will a global e-infrastructure affect the role and the impact of humanities within society? Basic services that are necessary requirements for the development of humanities research, such as those outlined above, could provide huge benefits to science, industry and society in very practical ways; especially in the areas of information retrieval, multilingual tools, authority services, and personalization. It is also clear that Web 2.0 mash-up technologies will enter the humanities digital library space in the very near future, just as it has taken over social networks and popular websites, so we will need to have robust web-services in place that are embedded within a global e-infrastructure that can cope with large-scale web-based multi-service interrogation and re-combination of

newly created digital material. The advances in humanities web-services may just be attractive enough to the general public, students, scholars, scientists and industry, that there will be a market driven demand to provide an e-infrastructure for the benefit of society (*qua*

personal, intellectual, economic and social) and hence justify the expenditure public resources needed to get the job done. In this case the many may be smarter than the few, and the collective 'wisdom of crowds'<sup>4</sup> may yet win the day.

---

<sup>1</sup> 'The Cultural Heritage Language Technologies Consortium', J. Rydberg-Cox, D-Lib Magazine, May 2005, Vol. 11, no. 5.

<sup>2</sup> D-Lib, Nov. 2005, Vol. 11 no.11.

<sup>3</sup> 'The invisible library: Paradox of the global information infrastructure', C. L Borman, *Library Trends*, 2003, Vol. 51, no 4.

<sup>4</sup> *The wisdom of crowds; why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. J. Surowecki, 1st ed. New York: Doubleday, 2004





**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](#)

Prius quam provincia decederet, consilium atrocitatis legiones, quae post excessum A mouerant, contrucidandi, quod et patrem ducem et se infantem tunc obsedissent, u cogitatione reuocatus, inhiberi nullo mode uelle perseueraret. uocatas itaque ad contum etiam gladiis depositis, equitatu armato ci uideret suspecta re plerosque dilabi ad resu arma, profugit contionem confestimque u deflexa omni acerbitate in senatum, cui ad dedecorum rumores palam minabatur, que fraudatum se iusto triumpho, cum ipse pa honoribus suis ageretur, etiam sub mortis

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



P. Vergilius Maro, *Aeneid*  
J. B. Greenough, Ed.

Your current position in the text is marked in blue. Click anywhere in the line to

book: \_\_\_\_\_  
line: \_\_\_\_\_

**This text is part of:**  
[Greek and Roman Materials](#)  
[Latin Poetry](#)  
[Latin Texts](#)  
[Vergil](#)  
[Vergil, Aeneid](#)

**View text chunked by:**  
[book : line](#)  
[book : line](#)

**Table of Contents:**  
[Book 1](#)  
[Book 2](#)  
[Book 3](#)  
[Book 4](#)  
[line 1](#)  
[line 31](#)  
[line 54](#)  
[line 90](#)

**Table of Contents** ← →

Click on a word to bring up parses, dictionary entries, and statistics

At regina gravi iam dudum saucia cura  
volnus alit venis, et caeco carpitur igni  
Multum viri virtus animo, multumque re-  
gentis honos: haerent infixi pectore vo-  
verbaque, nec placidam membris dat c-  
Postera Phoebea lustrabat lampade te-  
umentemque Aurora polo dimoverat t-  
cum sic unanimam adloquitur male sa-  
"Anna soror, quae me suspensam in so-  
Quis novus hic nostris successit sedibu-  
quem sese ore ferens, quam forti pecto-  
Credo equidem, nec vana fides, genus  
Degeneres animos timor arguit: heu, c-  
iactatus fati! Quae bella exhausta can-  
Si mihi non animo fixum immotumque  
ne cui me vinco vellem sociare iugali,  
postquam primus amor deceptam mor-  
si non pertaesum thalami taedaeque fi-  
huic uni forsitan potui succumbere culp-  
Anna, fatebor enim, miseri post fata S-  
coniugis et sparsos fraterna caede Pen-

Word Study Tool

Get Info for \_\_\_\_\_ in Latin

**saucius**  
(Show lexicon entry in [Elem. Lewis Lewis & Short](#)) (search)

saucia	adj pl neut nom	no user votes	14.1%	<a href="#">[vote]</a>
saucia	adj pl neut voc	no user votes	13.8%	<a href="#">[vote]</a>
saucia	adj pl neut acc	no user votes	13.9%	<a href="#">[vote]</a>
saucia_	adj sg fem abl	no user votes	13.9%	<a href="#">[vote]</a>
saucia	adj sg fem nom	no user votes	14%	<a href="#">[vote]</a>
saucia	adj sg fem voc	no user votes	13.6%	<a href="#">[vote]</a>

Word Frequency Statistics ([more statistics](#))

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

**saucio** to wound, hurt  
(Show lexicon entry in [Elem. Lewis Lewis & Short](#)) (search)

**saucia\_ † verb 2nd sg pres imperat act** no user votes 16.6% [\[vote\]](#)

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the correct form. ([More info](#))

Word Frequency Statistics ([more statistics](#))

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
609375	<a href="#">19</a>	0.31	4	0.07	<a href="#">Latin Poetry</a>
3414041	<a href="#">42</a>	0.12	14	0.04	<a href="#">Latin Texts</a>

Figure 3 – 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.