

Repositories, Cyberinfrastructure, and the Humanities

A 2006 report, funded by the American Council of Learned Societies (ACLS) and the Andrew W. Mellon Foundation, called for the creation of a cyberinfrastructure for the humanities and social sciences.¹ Before designing such a cyberinfrastructure, however, we need to consider the avenues of intellectual activity to be supported. Humanists, like all academics, want to enhance the collective understanding of their subject. But humanists are not producing scientific knowledge to address the problem of climate change or to eradicate dreaded diseases. Humanists have a particular obligation: to enhance the intellectual life of humanity as a whole.

Of course, there are fundamental and predictable needs—humanists need a network of repositories, for example. That being said, I argue that the greatest need is for networked repositories that can integrate collections and services distributed across the network in the short term and can then maintain these collections and services over decades. For this, the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) protocol (<http://www.openarchives.org/ore/>) provides the most promising intellectual framework, but we need robust repositories to provide the actual services.

The sciences have, to a large extent, assumed that their audiences will work in English. In the humanities, we have a fundamental obligation to make the human cultural record as physically and intellectually accessible as possible to every human being, regardless of their linguistic and cultural circumstances.

The contents of our libraries and museums must be accessible everywhere. The spaces through which we move should be readable: we should be able to view, in a form that best matches our immediate purpose, historical and cultural background for any location that we view. Such background must include overviews and links that lead to the full evidence for every statement that they contain.

Within the humanities, classicists are well positioned because they have been building digital infrastructure for a generation,² and they have a much more linguistically and culturally diverse audience online than they could reach with traditional print publications. The Greco-Roman world includes not only Europe but most of the Middle East: some Greek learning was reintroduced to Europe via Arabic translation, and a number of Greek texts survive only in the Arabic translations produced by Islamic scholarship. This heritage includes the classical world as well as the vast body of postclassical Latin and Greek. For centuries, these languages were a symbol and a component of that idea which has led today to the European Union. We thus have a natural audience that includes not only the traditional languages of classical scholarship (English, French, German, and Italian as well as Latin and Greek) but also Hungarian and Croatian, Arabic and Farsi. Still, in a globalized society, we need to think beyond the geographic

We have a fundamental obligation to make the human cultural record as physically and intellectually accessible as possible to every human being.

boundaries of the past and seek audiences who speak such widespread languages as Chinese and Hindi.

We already have the technologies whereby historical fields such as classics can begin building cyberinfrastructures that can reach audiences previously cut off linguistically and culturally. First, we need to optimize general machine translation for particular domains: for example, building language models that discern whether “case” describes a linguistic

category in a Greek grammar or a display cabinet in a museum catalogue. Second, classicists have begun to store basic information in emerging ontologies such as CIDOC CRM (<http://cidoc.ics.forth.gr>) and Functional Requirements for Bibliographic Records (<http://www.oclc.org/research/projects/frbr/>) to facilitate the problem of translation into multiple languages. Third, we are beginning to develop for our core canonical texts a network of linguistic annotations, resolving ambiguities of syntax (What is the object of the verb?), co-reference (What is the reference for a pronoun such as “he” in a particular text?), semantics (Is “bank” the edge of a river or a financial institution?), and other categories. Such data not only place our understanding of historical languages on a fundamentally new, quantitative footing and provide a new generation of reading support tools but also provide a knowledge base with which to generate better translations.

Three service layers help manage

these categories of data: (1) *catalogue services*, which identify the discrete objects within a collection (e.g., editions of Vergil's *Aeneid*, books about Vergil); (2) *named entity services*, which identify semantically significant objects embedded within collection contents (e.g., references to Vergil or the *Aeneid* within other documents), including semantic data (e.g., connecting a particular word in a particular text to a named dictionary sense); and (3) *customization and personalization services*, which identify ways to customize information for the user (e.g., given a particular passage of the *Aeneid*, what would be of interest to an intermediate student of Latin versus a professional Latinist, and into what language do Vergil's Latin and the various scholarly annotations have to be translated?). Summarization, visualization, machine translation, and other technologies all play roles within one or more of these service layers.

Catalogue Services

Catalogue services allow us to find intentionally labeled logical structures within collections. Humanists need catalogues that manage individual documents and their relations as well as the canonical citation schemes by which we address particular chunks of text: in actual practice, humanists typically want information about a named chunk (e.g., chapter 86, book 1) or a semantic unit (Pericles' *Funeral Oration*) more often than they want information about the whole (e.g., Thucydides' *History of the Peloponnesian War*). This requires not only library ontologies such as FRBR but also finer-grained methods such as those described in the Canonical Text Services Protocol (<http://chs75.harvard.edu/projects/diginc/techpub/cts>).

Named Entity Services

While catalogues provide access to well-defined objects within a collection, named entity services locate references and provide the basis to summarize information that appears within the contents of collections.³ Named entities can be documents (e.g., references to Thucydides' *History of the Peloponnesian War*), citations within documents, or people, places, organizations, events, and the other topics for which users consult catalogues, encyclopedias, and gazetteers.

Named entities also include linguistic topics: 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* and *factus*. Every word sense in a dictionary and every linguistic phenomenon in grammar is a separate named entity. Every subject heading or topic to which we assign a label is a named entity.

Although place-names can be extracted automatically from text and linked, using geographic information systems, to places in the world, this is not a precise art, and technical challenges remain. 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 locations (e.g., 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 about 95 percent. In January 2007, Google released its service to map places from digitized books in Google Book Search. The ability to customize Google results, substituting more accurate services, would be of significant value to the humanities.

Customization and Personalization Services

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 knowledge of Greek and should extend the capacity of those studying the language, at all levels, to draw meaning from the text.

Useful tools need not be complex or immensely sophisticated. A simple but successful approach supports vocabulary customization based on what a user already knows. The user develops a profile based on his or her use of a Latin textbook. The system then automatically compares that profile against words it detects as new to the user, thus identifying words the user probably has and has not encountered before. Even though the

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

Conclusion

Classicists have a generation of experience developing services and complex data sets. They already can point to existing services, available on an as-is basis from disparate projects, and these services, when given stable homes in large-scale systems, will provide a first generation of humanities cyberinfrastructure. The potential impact goes beyond the production of new humanities scholarship and allows us to begin designing a knowledge environment that can make the human record intellectually accessible to audiences of far wider linguistic and cultural backgrounds than was ever feasible before.

Notes

1. ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences, *Our Cultural Commonwealth* (2006), <<http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf>>. This built on a preceding National Science Foundation report, which established the term "cyberinfrastructure": *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, January 2003, <<http://www.nsf.gov/od/oci/reports/atkins.pdf>>.
2. Gregory Crane, David Bamman, and Alison Babeu, "ePhilology: When the Books Talk to Their Readers," in Ray Siemens and Susan Schreibman, eds., *A Companion to Digital Literary Studies* (Malden, Mass.: Blackwell Publishers, 2007), pp. 29–64, <http://dl.tufts.edu/view_pdf.jsp?urn=tufts:facpubs:grane-2006.00003>; Gregory Crane, "Classics and the Computer: An End of the History," in Susan Schreibman, Ray Siemens, and John Unsworth, *A Companion to Digital Humanities* (Malden, Mass.: Blackwell Publishers, 2004), pp. 46–55, <<http://www.perseus.tufts.edu/~grane/blackwells.final.Crane.pdf>>.
3. Alison Babeu, David Bamman, Gregory Crane, Robert Kummer, and Gabriel Weaver, "Named Entity Identification and Cyberinfrastructure," in *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries* (Budapest: Springer Verlag, 2007), pp. 259–70.



Gregory Crane is Winnick Family Chair in Technology and Entrepreneurship, Professor of Classics, and Editor in Chief of the Perseus Project at Tufts University.