Building Cultural Digital Libraries

Gregory Crane
Editor in Chief
Perseus Project
Professor of Classics
Winnick Family Chair of Technology and Entrepreneurship
Tufts University
gcrane@tufts.edu

Our work on the Perseus Digital Library has continued since 1987. Our initial goal was to create a critical mass of materials relevant to the study of a single discrete subject. We initially chose classical Greece (in part because the primary materials for this domain were limited and manageable) but our long term goal was, from the beginning to study general problems underlying digital libraries in general and humanities digital libraries in particular. We began branching out into Renaissance English and Roman materials in 1997. Current funded research projects include a DL on the history and topography of London and its environs, excavations of ancient remains at Giza in Egypt (with the Museum of Fine Arts Boston), the history of Mechanics from antiquity to the early modern period (with the Max Planck Institute for the History of Science in Berlin), the development of the new Greek Lexicon (with a project team centered in the UK), the form of the New Variorum Shakespeare Series (with the Modern Language Association), and the structure of various sections from the American Memory site of the US Library of Congress.

Our production Web site (www.perseus.tufts.edu) draws substantial traffic (up to 300,000 page impressions in a twenty four hour period during the spring of 2000) and allows us to test our strategies for structuring and representing data in real world conditions. Current research areas include:

Document Design: Digital Library system research often focuses on the problem of dealing with documents as is. We are exploring structures that will allow documents to take fuller advantage of the capacities of a digital library system. Problems range from disambiguation during the editing process (identifying "Springfield" as a place, not a rifle, and as a particular place — Springfield in Missouri vs. New Hampshire) to the design of virtual spaces with seamless links to the sources upon which the virtual space is based.

Data Visualization: Automatic document comparison and even text searching Time and space are obvious mechanisms for abstracting and visualizing the contents of documents. We automatically extract dates and toponyms, generating timelines and maps for documents, collections and the Perseus DL as a whole. Ultimately, we will be able to search the database according to time and space (e.g,, locate all contents that describe a Worcester County in Mass in the 1890s). Other features lend themselves to this technique as well: the value of money shifts over time and a DL that tracks monetary values could provide important tools for search.

Impact on Cognition: The impact of Perseus has been evaluated continuously for more than a decade. A postdoc with expertise in text comprehension and spatial cognition will be joining our staff in September. Our goal is to apply state of the art techniques from these fields of cognitive science to understand more fully the effect of various designs on users from K-12 through researcher.

All three of these research areas interact: document design enables data visualization; studying the impact on cognition informs visualization and, in turn, influences backend document design.

Collaborations on the Korean Digital Library

Many of the position papers overlap with our research. Sung Been Moon and Sam Oh's work on integrating human factors into DL design, Hae-Chang Rim's research on constructing bilingual resources for digital libraries and Sung Hyon Myaeng's cross-language federated searching are directly relevant to our core efforts. The Tiger Project proposed by Sung Hyuk Kim stands out in particular, as its overall goals and design are strikingly similar to our approach with the Perseus Digital Library. Virtually every task outlined has a parallel in our research. Three elements in particular stand out:

1) **The emphasis on outreach beyond the traditional community**. Perseus was designed from the outset to expand the set of people, from students to scholars, directly engaged with primary materials. This is a fundamental principle of design which many humanities DL projects ignore, focusing instead upon the traditional researchers in a field. Various consequences follow from this principle, affecting data collection, visualization, and, above all, rights policies. If the Tiger Project is to have its maximum impact, then the project must work with the Korean government to establish the most open rights policies possible. If the Tiger DL is designed to promote knowledge of Korean Culture, then it should be made freely available and with minimal rights restrictions. Here the American Memory Project of the Library of Congress provides an outstanding model.

2) **The Maintenance of a Bilingual Corpus**. Perseus maintains bilingual corpora of Greek and Latin source texts (c. 5,000,000 words) with matching English translations. These include a network of lexica, commentaries, grammars, and other reading tools. Machine translation is an important technique but it screens the reader from the original source language. In a developed environment, the reader proficient in English but with little or no knowledge of the source language (whether Greek or Korean) should be able to query every word in the source language and, with the aid of on-line tools, make effective use of the original source texts.

3) **Integration of heterogeneous data**: Spectacular 3D reconstructions, elegant translations and other elements of a cultural digital library have little significance if they are not interlinked with other components. The individual viewing an art object should, for example, be able to retrieve relevant textual documents. Libraries are spaces in which autonomously designed and created objects, from many different

sources, coalesce into a whole that is greater than the sum of its parts. A DL effort must challenge its specialist contributors to think beyond their individual focus.

Plan of Action

The Perseus Digital Library Project is willing to collaborate with one or more partners or teams from Korea and/or the US.

First, we propose to expand the evolving PDL system to incorporate Korean materials. Moving to a non-European language and culture would constitute a logical next step for our research agenda. A common document architecture stands at the core of this effort: if we can design documents that follow common standards then we will enhance the power of the DL systems that subsequently manage them. Such a convergence would require work on both sides. We certainly hope that our colleagues in Korea will be able to learn from our experiences. On the other hand, although we have from the beginning of our work struggled to make every element of the PDL as general as possible, we may also find that Korean materials raise issues that the European collections in Perseus have not. It may therefore be necessary to modify our existing collections to bring them into conformance with a more generic standard.

Second, we propose to collaborate in developing a truly multilingual version of the PDL. At present, the PDL is an English-language system with non-English source materials. We propose developing a Korean language version that incorporates non-Korean materials. The initial focus would be not only on Korean cultural materials but on creating an environment that would make non-Korean materials more accessible to a Korean audience. One strategy would be to take English collections (e.g., the Perseus materials, collections from the Library of Congress American Memory Project) and to tie these with a variety of language tools, including Machine Translation, automated dictionaries and, for select portions, Korean translations.