The Challenge of the Humanities to the World Wide Web: Perspectives from the Archimedes Project

Mark J. Schiefsky, Malcolm D. Hyman

March 12, 2004

1 The Challenge

It cannot be denied that the development of information technology poses a profound challenge to traditional scholarship in the humanities. The rapid development of computer technology and of the World Wide Web has opened up immense new opportunities for scholars working in historical disciplines. These opportunities also challenge the traditional ways in which scholarship has been practiced. Yet the challenges presented by humanistic scholarship to information technology and to the World Wide Web are perhaps even greater. In Part One of our paper, we briefly describe four fundamental limitations of the World Wide Web in its current form as a medium of scholarship in the humanities:

1. **Language technology.** Research in natural language processing has resulted in a proliferation of tools for automatic linguistic analysis of particular languages. There is, however, no way to bring the data produced by these tools into a browsing or editing environment in the absence of standard formats and protocols. What is needed is linguistic middleware that enables user agents to interact with heterogeneous sources of linguistic data.

2. **Semantic linking.** Scholarly research requires the creation and utilization of meaningful links between source materials. Ordinary HTML links are unidirectional and cannot differentiate between types of semantic relations. Digital library applications must be developed that achieve the ideal of a "semantic web"[1] by exploiting the potentials of relevant standards such as the XML Linking Language (XLink).[2]

[1] T. Berners-Lee, The semantic web, *Scientific American*, May 2001.

[2] `http://www.w3c.org/TR/xlink/`.

3. **Content creation.** Central to the notion of a digital research library is human analysis and annotation of source materials. No matter how sophisticated the algorithms used to generate them, automatically created links cannot take the place of scholarly analysis. The current paradigm of the Web distinguishes the creation and browsing of content as fundamentally separate activities. In a next-generation framework the browsing and creation of content must be more closely integrated.

4. **Distributed resources.** The free exchange of ideas is a crucial feature of research, whether humanistic or scientific. Any scholar anywhere who has access to the Web should be able to contribute materials to a distributed set of resources and to work with resources produced by other scholars worldwide. Yet the current client-server architecture of the Web limits the free flow of information. Digital libraries must move toward a more fluid distributed or peer-to-peer network model.[3]

2 Software Platform

We describe a software platform designed as a first step towards meeting these challenges. This software platform has been developed in the context of the Archimedes Project, an international initiative to create a digital research library for the history of mechanics funded by the National Science Foundation in the United States. Although our software has been developed with a view to solving the problems arising in the course of work on this project, it is not tied to the requirements of any particular area of historical scholarship. Our software platform has three principal components, all of which are freely available in the Internet at `http://archimedes.fas.harvard.edu`:

1. **The Pollux system** provides a unified means of access to dictionaries, or any other reference work that is organized by alphabetized headwords, in any natural language. The software is designed to make it possible for users to add new lexica with a minimum of effort.

2. **The Donatus system** provides a unified frontend to a variety of morphological analysis software and databases.[4] Morphological services are provided both through a Remote Procedure Call (RPC) interface that can be utilized by specialized user applications and through a CGI interface that is accessible in any web browser. Morphological data can be represented in XML, allowing them to be cached on client systems and to be processed by a wide range of software. Backend systems that have already been incorporated include Morpheus, a morphological analyzer for ancient Greek, Latin, and Italian developed by the Perseus Project;[5]

[3] P. Fox, P2P for grown-ups, *Computerworld* 37(12), March 24, 2003, pp. 22–3.
[4] `http://archimedes.fas.harvard.edu/cgi-bin/donatus/`.
[5] `http://www.perseus.tufts.edu/`.

the CELEX Linguistic Database for Dutch, English, and German developed by the Center for Linguistic Information of the Max Planck Institute for Psycholinguistics in Nijmegen;[6] and the Xerox finite-state morphological analyzer for Arabic developed by the Xerox Research Centre Europe (XRCE) in Grenoble.[7] Work is currently underway to integrate a morphological analyzer for Sanskrit that is being developed at Brown University and one for Sumerian being developed at the University of Pennsylvania. In addition to providing access to pre-existing linguistic data, Donatus allows for the dynamic extension of morphological datasets by a user.

3. **The Arboreal user agent** is a powerful and flexible tool for content-based access to and annotation of XML texts. Arboreal includes special features for working with parallel versions of texts, morphology and terminology, and linked images. Integrated language support is currently provided for Latin, Greek, Arabic, Chinese, languages written in cuneiform, and major western European languages. Arboreal supports many standards and is designed as a cross-platform tool that can be used on many different computing systems. Distributions are currently available for Mac OS X, Windows, and GNU/Linux. We envision Arboreal as a prototype for the next-generation web user agent, which closely integrates content browsing and content creation. Arboreal allows for highly flexible navigation of any XML document, using two document views that are presented side-by-side. One pane displays a tree view of the document; the user can control the level of detail shown in the tree by expanding and collapsing nodes and sets of nodes. The other pane offers a detail view of the portions that are selected in the tree. Both views are customizable through a document description language (DDL), which we aim to extend in the next phase of the project. Powerful search capabilities are available, including regular expression searching, lemmatized searching (which takes advantage of morphological data generated by the Donatus system), XPath queries, and the ability to search in an orthographically normalized representation of the text (e. g. a query for the Latin word ‘uectis’ will also find ‘vectis’). Arboreal can be customized to work with any natural language, by supplying a description of the language in an XML langspec. This description makes possible language-specific features such as word detection and allows for various language-specific views to be defined (e. g. Romanization of a non-Roman script, or fully-voweled vs. non-voweled Arabic script). Close integration with the linguistic services provided by Donatus and the Pollux reference system is provided, allowing the user to access morphological analyses and dictionary entries for any word in a text.

[6] `http://www.kun.nl/celex/`.
[7] `http://www.xrce.xerox.com/`.

3 Application and Evaluation

We present a number of concrete applications of our software platform in the context of the Archimedes Project. First, this platform has proven invaluable in the basic work involved in building up a digital collection that now amounts to some 110 MB of text and 30 GB of associated images. For example, with the aid of the automatic morphological analysis provided by Donatus and the term annotation facilities of Arboreal, a user can easily highlight all words in a document that cannot be analyzed by the Donatus system, allowing for both the speedy correction of typographical errors and the improvement of our morphological datasets. Second, Arboreal's ability to apply arbitrary XSLT transformations allows us to create a range of derived files of immediate scholarly value from a single source text. Third, Arboreal has been extensively used in conjunction with external editing environments to produce scholarly metadata, in particular for the alignment and linking of parallel sections of different XML texts. Fourth, Arboreal's term annotation and editing facilities provide support for scholars attempting to produce consistent translations of source texts. Finally, we have taken preliminary steps towards integrating into our framework sophisticated NLP techniques for the discovery of technical terminology and establishing metrics for assessing the utility of these techniques.

A number of cooperating projects and institutions are also currently using components of our software platform. These include the Cuneiform Digital Library Initiative (CDLI),[8] the project *Gli anni della cupola 1417–1436: Archivio digitale delle fonti dell'Opera di Santa Maria del Fiore* (The Administrative Archives of the Cathedral of Florence),[9] European Cultural Heritage Online (ECHO),[10] and the Digital Sanskrit Library at Brown University.[11] Feedback from these projects has demonstrated the broad range of application of our tools and allowed us to create additional enhancements.

4 Further Perspectives

We describe further extension and generalization of our software platform to realize the goals stated in Part One, focusing on two areas in particular: (1) enhancements in core language technology, and (2) the development of tools for ontology creation and visualization.

[8] `http://cdli.ucla.edu/`.
[9] `http://duomo.mpiwg-berlin.mpg.de/`.
[10] `http://echo.mpiwg-berlin.mpg.de/`
[11] `http://sanskritlibrary.org/`.