ELECTRONIC PUBLICATION OF ANCIENT NEAR EASTERN TEXTS

A conference sponsored by the Oriental Institute and the Franke Institute for the Humanities of the University of Chicago

October 8th and 9th, 1999

All sessions except Friday evening are in the Franke Institute in Regenstein Library, 1100 East 57th Street.
For further information, contact David Schloen at d-schloen@uchicago.edu or 773-702-1382.
If you plan to attend, please inform David Schloen if you have not already done so.

Friday October 8th

8:30–9:00 a.m. *Participants arrive at the Oriental Institute, 1155 East 58th Street, and are directed to the Franke Institute where coffee, tea, bagels, and pastries will be provided.*

9:00–9:10 Welcome and opening remarks (*Gene Gragg, Director, Univ. of Chicago Oriental Institute*)

9:10–10:00 "From Dictionary to Superdocument: XML, the Pennsylvania Sumerian Dictionary, and the Universe" (*Steve Tinney, Univ. of Pennsylvania*)

10:00–10:10 Questions and discussion

10:10–10:30 *Coffee break*

10:30–10:50 "The Ancient Egyptian Dictionary Project: Data Exchange and Publication on the Internet" (*Stephan Seidlmayer, Berlin-Brandenburg Academy of Sciences and Humanities*)

10:50–11:00 Questions and discussion

11:00–12:00 *Open discussion:* "The Current State of Electronic Publication: Problems and Possibilities" (*moderated by Charles Jones, Oriental Institute Research Archives, and John Sanders, Oriental Institute Computer Laboratory, Univ. of Chicago*)

12:00–1:30 p.m. *LUNCH BREAK* (*coffee, tea, and sandwiches will be provided, or participants may walk to nearby eating places*)
**For a tour of the new Egyptian gallery, meet Dr. Emily Teeter at 1:00 in the lobby of the Oriental Institute.*

1:30–2:20 "Creating, Integrating, and Expanding Electronic Texts in the Perseus Digital Library" (*Jeffrey Rydberg-Cox, The Perseus Project, Tufts University*)

2:20–2:30 Questions and discussion

2:30–2:50 *Coffee break*

2:50–3:40 "XML and Digital Imaging Considerations for an Interactive Cuneiform Sign Database" (*Sandra Woolley and Theodoros Arvanitis, School of Electronic and Electrical Engineering, and Tom Davis, Dept. of English, Univ. of Birmingham*)

3:40–3:50 Questions and discussion

3:50–4:50 *Open discussion:* "Editing, Disseminating, and Preserving Electronic Publications" (*moderated by Charles Jones and John Sanders, with panelists Patrick Durusau of Scholars Press, James Eisenbraun of Eisenbrauns Inc., and Thomas Urban of the Oriental Institute Publications Office*)

4:50–4:55 Announcements (*David Schloen*)

5:00–5:30 *Wine and cheese reception in the Director's Study in the Oriental Institute*

5:30–7:30 *DINNER BREAK* (*directions will be provided to nearby restaurants*)

7:30–9:00 p.m. *Presentations of electronic text publication projects in Breasted Hall in the Oriental Institute*

"The Achaemenid Royal Inscriptions Project" (*Gene Gragg and Matthew Stolper, Univ. of Chicago*)

"Egyptian Hieroglyphic Text Processing, XML, and the New Millennium" (*Hans van den Berg, Center for Computer-aided Egyptological Research, Utrecht University*)

"Using Encoded Texts at ARTFL: The Case for Simplicity" (*Mark Olsen, Project for American and French Research on the Treasury of the French Language, Univ. of Chicago*)

Saturday October 9th

8:30–9:00 a.m.	*Coffee, tea, bagels, and pastries will be provided in the Franke Institute.*
9:00–9:50	"The Electronic Text Corpus of Sumerian Literature" (*Jeremy Black and Eleanor Robson, Univ. of Oxford*)
9:50–10:00	Questions and discussion
10:00–10:20	Comments on encoding cuneiform texts (*Miguel Civil, Univ. of Chicago*)
10:20–10:30	Questions and discussion
10:30–10:50	*Coffee break*
10:50–12:00	*Open discussion:* "Standards for Text Encoding and Markup" (*moderated by Gene Gragg and Steve Tinney*)
12:00–1:30 p.m.	*LUNCH BREAK* (*coffee, tea, and sandwiches will be provided, or participants may walk to nearby eating places*)
1:30–1:50	Proposal to form a "Working Group on Cuneiform Markup" (*Gene Gragg, Univ. of Chicago*)
1:50–2:20	Questions and discussion
2:20–2:40	*Coffee break*
2:40–3:30	"Texts and Context: Using XML to Integrate and Retrieve Archaeological Data on the Web" (*David Schloen, Univ. of Chicago*)
3:30–3:40	Questions and discussion
3:40–4:50	*Open discussion:* "What's It Good For? Uses of Electronically Published Texts" (*moderated by Matthew Stolper, Univ. of Chicago*)
4:50–5:00	Concluding remarks (*Gene Gragg, Univ. of Chicago*)

ELECTRONIC PUBLICATION OF ANCIENT NEAR EASTERN TEXTS

A conference sponsored by the Oriental Institute and the Franke Institute for the Humanities of the University of Chicago

October 8th and 9th, 1999

ABSTRACTS

"From Dictionary to Superdocument: XML, the Pennsylvania Sumerian Dictionary, and the Universe"

Steve Tinney, Babylonian Section, University of Pennsylvania Museum (stinney@sas.upenn.edu)

The ever-increasing importance of computers in gathering, storing, and presenting knowledge brings with it the need to respond to the challenge of exploiting the potential of electronic information management to the maximum. But while it is natural to view knowledge management from the viewpoint of eventual delivery or publication, whatever form it may take, the issue of information reusability is at least as important, and arguably more so. Reusability requires well-structured information as well as permission to reuse it and the means to access it. One of the promises of XML and its increasing circle of friends and relatives is the provision of a well-defined means of defining information structure and accessibility, not only as it relates to publication, but also as it relates to the storage and relational organization of interconnected datasets. Several aspects and implications of the above will be discussed in the present paper, particularly as they relate to the ongoing development of an electronic version of the Pennsylvania Sumerian Dictionary (ePSD). A brief orientation to relevant components of the XML world will be followed by a discussion of some key elements of the ePSD implementation. Some general considerations concerning the relationship between data, knowledge, dissemination and extant structures of publication and academia will also be offered.

"The Ancient Egyptian Dictionary Project: Data Exchange and Publication on the Internet"

Stephan Seidlmayer, Berlin-Brandenburg Academy of Sciences and Humanities (seidlmayer@bbaw.de)

Since 1993 the Ancient Egyptian Dictionary project has been housed at the Berlin-Brandenburg Academy of Sciences and Humanities in Berlin. It aims to provide up-to-date lexical information on the Egyptian language, supplementing and replacing the great *Wörterbuch der ägyptischen Sprache* by Adolf Erman and Hermann Grapow, which appeared in twelve volumes between 1926 and 1963, and which is outdated in important respects. As in the *Wörterbuch der ägyptischen Sprache*, work at the Ancient Egyptian Dictionary project is centered on compiling a comprehensive corpus of Egyptian texts, which in turn provides the basis of the dictionary. Both the corpus of texts and the dictionary are produced as a database and will be published in due course on the Internet. In this context, encoding Egyptian texts in XML will play an important part, because this standard supports long-term system-independent storage of the data, because it offers a common platform for the exchange of encoded texts and thus for international cooperation in the Ancient Egyptian Dictionary project, and because it opens up new perspectives for the publication of the material on the Internet.

Open discussion: **"The Current State of Electronic Publication: Problems and Possibilities"**

Moderated by Charles Jones, Research Archivist and Bibliographer, Oriental Institute (ce-jones@uchicago.edu) *and John Sanders, Head, Oriental Institute Computer Laboratory* (jc-sanders@uchicago.edu)*, University of Chicago*

Topics to consider include: (1) advantages and disadvantages(!?) of electronic publication as compared to traditional print publication; (2) good and bad examples of electronic publications that are available today; (3) the importance of cross-platform access based on non-proprietary open standards; (4) the limitations of HTML and the impact of XML/SGML; (5) the availability and effectiveness of software tools for producing XML-based electronic publications; (6) the distinction between markup of content or logical structure and markup of style or presentation characteristics; (7) the merits of facsimile reproduction versus transliteration of texts.

"Creating, Integrating, and Expanding Electronic Texts in the Perseus Digital Library"

Jeffrey Rydberg-Cox, Assistant Editor for Greek Language and Lexicography, The Perseus Project
(jrydberg@perseus.tufts.edu)

The Perseus Project (*http://www.perseus.tufts.edu*) is an evolving digital library of resources for the study of the ancient world and beyond. Collaborators initially formed the project to construct a large, heterogeneous collection of materials, textual and visual, on the Archaic and Classical Greek world. Recent expansion into Latin texts and tools and Renaissance materials has served to add more coverage within Perseus and has prompted the project to explore new ways of presenting complex resources for electronic publication. In this paper the data entry methods, archival formats, and initial tagging process used by the Perseus Project will be presented, followed by a description of how tagged information is used to create "interoperable" primary and secondary sources. These secondary sources include research tools such as lexica, commentaries, and morphological analyses, and plans are now being developed for the integration of geographical data and architectural reconstructions as well. Mention will also be made of the use by the Perseus Project of techniques from the fields of information retrieval and corpus linguistics in combination with structured data to add value to electronic works.

"XML and Digital Imaging Considerations for an Interactive Cuneiform Sign Database"

Sandra Woolley (s.i.woolley@bham.ac.uk) *and Theodoros Arvanitis* (t.n.arvanitis@bham.ac.uk),
Educational Technology Research Group, School of Electronic and Electrical Engineering,
Tom Davis (t.r.davis@bham.ac.uk), *Department of English, and*
Alasdair Livingstone (a.livingstone@bham.ac.uk), *Department of Ancient History and Archaeology,*
University of Birmingham

This presentation will summarize the work of a centrally-funded interdisciplinary team project at the University of Birmingham, working toward an interactive database of cuneiform signs. The project team comprises cuneiform specialists from the Department of Ancient History and Archaeology, digital imaging researchers from the School of Electronic and Electrical Engineering, and a forensic scientist from the Department of English Literature. The presentation will describe the objectives of the project, the findings of the first 12-month study, work in progress, and plans for future work. There will be a brief description of the basic principles of forensic handwriting identification, with examples. A proposed database format, issues relating to XML coding of the data, and plans to improve digital image representations of cuneiform signs will be presented.

From the website of the University of Birmingham Cuneiform Database Project (http://www.eee.bham.ac.uk/cuneiform):
The usual method of recording and publishing cuneiform material is through the time-consuming process of copying by hand, and this method is also used in the standard reference lists of cuneiform signs. Inevitably, the hand of the modern copyist comes between the hand of the ancient scribe and the eye of the modern scholar who uses the copy. The University of Birmingham Cuneiform Database Project seeks to apply the most recent research on digital representation and compression to the particular challenges posed by cuneiform texts. Relevant techniques include those developed for industrial inspection and medical imaging. One of the principal challenges involves three-dimensional visualization and computation, because a drawback of the traditional handcopying method is the fact that a three-dimensional script is represented in two dimensions on paper. A further interdisciplinary aspect of the Cuneiform Database Project involves the adaptation of existing techniques of handwriting analysis to the cuneiform writing system, providing a scientific system of description that will enable an objective categorization of scripts and script types.

Open discussion: **"Editing, Disseminating, and Preserving Electronic Publications"**

Moderated by Charles Jones and John Sanders, Oriental Institute, University of Chicago
With panelists Patrick Durusau, Interim Manager, Information Technology Services, Scholars Press (pdurusau@emory.edu),
James Eisenbraun, Publisher, Eisenbrauns Inc. (jeisenbraun@eisenbrauns.com),
and Thomas Urban, Senior Editor, Oriental Institute Publications Office, University of Chicago (t-urban@uchicago.edu)

Topics to consider include: (1) the ease of "self-publication" on the Web and the role of peer review and editorial oversight; (2) maintenance and upgrading of delivery media, whether optical disks or Internet servers; (3) citation of electronic publications and the problem of permanence; (4) the economics of electronic publication and the fate of traditional publishers; (5) institutional responsibilities for the establishment of digital monograph series and journals.

"The Achaemenid Royal Inscriptions Project"

Gene Gragg (g-gragg@uchicago.edu) *and Matthew Stolper* (m-stolper@uchicago.edu),
Oriental Institute, University of Chicago

The aim of the Achaemenid Royal Inscriptions project (*http://www-oi.uchicago.edu/oi/proj/ARI*) is to create an electronic study edition of the inscriptions of the Achaemenid Persian kings in all of their versions: Old Persian, Elamite, Akkadian, and, where appropriate, Aramaic and Egyptian. The edition is to be accompanied by translations, glossaries, grammatical indexes, basic bibliographic apparatus, basic text critical apparatus, and some graphic apparatus (e.g., plans indicating provenience of the inscriptions, images of exemplars); the texts will be available for downloading and printing. The first stage of the project presents the inscriptions from Persepolis and nearby Naqsh-i Rustam, where the Oriental Institute of the University of Chicago carried out excavations between 1931 and 1939. Close study and accurate use of these texts calls for synoptic presentation of the versions. Yet no handy synoptic edition has replaced F. H. Weissbach's magisterial *Keilinschriften der Achämeniden* of 1911, because the development and divergence of scholarship on Old Persian and Old Iranian, Elamite, and Akkadian make replacing it with an equally compendious and authoritative printed edition a forbidding undertaking. On the other hand, the flexibility of the electronic media makes it possible to present useful working synoptic editions with apparatuses and illustrations that can be undertaken in stages, and that can be progressively enlarged, improved, and interconnected.

"Egyptian Hieroglyphic Text Processing, XML, and the New Millennium"

Hans van den Berg (vdberg@ccer.nl), *Center for Computer-aided Egyptological Research* (http://www.ccer.nl),
Utrecht University

Since 1985 there has existed in Egyptology a standard for the encoding of hieroglyphic texts for computer input, the so-called Manuel de Codage. This standard has been implemented in the three major hieroglyphic text-processing programs: Glyph, MacScribe, and Inscribe. The Manuel de Codage offers guidelines for alphanumeric (ASCII) and phonetic encoding of the signs, as well as the grouping of signs and layout of the text as a whole. Though an alternative system is now under construction by the Unicode consortium which makes a 16-bit set of character encodings available on any platform, the Manuel de Codage standard will no doubt keep dominating electronic hieroglyphic text processing for years to come. The main reason for this is the fact that producers of the established hieroglyphic text processing programs will not easily switch to another standard, in order not to confuse their customers, though ways may be found to make them compatible. Though standardized character sets help out when it comes to working on text publications and grammars, a problem lies in the growing desire of modern Egyptologists to do electronic epigraphy and palaeography. It is quite common that newly recorded texts yield previously unknown signs or character anomalies. To be able to record these the hieroglyphic text-processing software will have to find new and more flexible ways of working with character sets and character encoding. The coming of XML with its flexible element, attribute and entity declarations, and its flexible tag set, holds great promise for achieving this goal without giving up the established encoding system. With the use of graphic markup languages based on XML, all palaeographic characteristics or even whole character sets can be recorded and communicated, thus bringing a standardized way of hieroglyphic communication over the Internet within our reach. Forces are gathering in Egyptology right now to develop an Egyptian hieroglyphic markup language based on XML that will take hieroglyphic text processing and text exchange into the new millennium.

"Using Encoded Texts at ARTFL: The Case for Simplicity"

*Mark Olsen, Project for American and French Research on the Treasury of the French Language (ARTFL),
University of Chicago* (mark@barkov.uchicago.edu)

> XML does allow anyone to design a new, custom-built language, but designing good languages is a challenge that should not be undertaken lightly. And the design is just the beginning: the meanings of your tags are not going to be obvious to other people unless you write some prose to explain them, nor to computers unless you write some software to process them. (*Tim Bray, coeditor of the XML specification*)

Extensible Markup Language (XML) is generating considerable enthusiasm in many quarters. It is hailed as both a remarkable extension of Hypertext Markup Language (HTML), the current text-markup specification for World Wide Web documents and applications, and an equally important simplification and rationalization of Standard Generalized Markup Language (SGML), allowing users to more easily define text-encoding specifications. The goal of XML, to allow users to define and create "self-describing" documents and to use them, with appropriate stylesheets, directly in Web applications (e.g., popular browsers like Netscape Navigator and

Internet Explorer), is both laudable and important. The recent explosion of XML languages (XML "document type definitions" or DTDs) bears witness to the ease of use and expressive power of XML. XML is indeed a better SGML. XML is, however, like the SGML it replaces, a metalanguage: a formal way to specify a tagset that can be used to encode documents. The much-maligned HTML specification is itself an SGML DTD, an encoding specification written "in" SGML. One may specify encoding schemes in XML of widely varying complexity, from the relatively simple and flexible HTML-like (called Voyager) schemes to very complex specifications like the Text Encoding Initiative's DTD. In a very real sense, XML does not radically alter the problem of encoding specifications, because the real issues are found in the DTDs or "languages." Tim Bray, one of the designers of XML, writes that XML "lays down ground rules that clear away a layer of programming details so that people with similar interests can concentrate on the hard part—agreeing on how they want to represent the information they commonly exchange. This is not an easy problem to solve, but it is not a new one, either."

Since the problems are not new, it is relevant to this conference to examine the experience of humanities computing projects in using a variety of SGML DTDs and other encoding schemes, in order to identify past problems and to use this experience to help formulate XML languages. As Bray notes, creation of an XML language or DTD is complicated and constitutes only the beginning of the effort, which must include support such as software development. Evaluation of the utility of an encoding specification is broader than simply its internal, formal consistency. The ARTFL project has developed many large textual and multimedia databases using a variety of encoding schemes expressed in SGML and other less formal schemes. Drawing on examples from current ARTFL work, I will argue that development of relatively simple encoding specifications—devised in XML, SGML, or more informally—has a number of benefits for humanities computing projects. These include dramatically lowered costs in the development of sophisticated software for making these databases available to a wide range of users, the possibility of automated tagging, rapid training of graduate assistants, much more consistent tagging among individual students and employees, and most importantly, concentration on getting results. Text-tagging is a means to an end and not an end in itself. This should be clear to researchers in the humanities. I fear, however, that the example of the Text Encoding Initiative's declaration that it is a "new research community," presumably in text tagging, will suggest that only very extensive and complicated tagsets can be sufficient to represent humanities research materials. But the costs of complex encoding schemes are significant at all levels, from staff training to software development. Not only is it unclear that such extensive tagsets have practical use, particularly given the poor state of software development, but they ignore the fact that many, even most, of the problems entailed in processing humanities data have little bearing on many, if not most, of the problems posed by *automatic* processing of humanities data.

Simplicity, or economy of tagsets, can be expressed in XML.

From the website of the ARTFL project (http://humanities.uchicago.edu/ARTFL):

In 1957 the French government initiated the creation of a new dictionary of the French language, the *Trésor de la Langue Française*. In order to provide access to a large body of word samples, it was decided to transcribe an extensive selection of French texts for use with a computer. Twenty years later a corpus totaling some 150 million words had been created, representing a broad range of written French, from novels and poetry to biology and mathematics, stretching from the seventeenth to the twentieth centuries. It soon became apparent that this corpus of French texts was an important resource not only for lexicographers, but also for many other types of humanists and social scientists engaged in French studies, on both sides of the Atlantic. The result of this realization was "American and French Research on the Treasury of the French Language" (ARTFL), a cooperative project established in 1981 by the Centre National de la Recherche Scientifique and the University of Chicago. The ARTFL project has focused on three objectives over the past eight years: to include a wide variety of texts in order to make the database as versatile as possible; to create a system that would be easily accessible to the research community; and to provide researchers with an easy-to-use but effective tool. At present the corpus consists of nearly 2000 texts, ranging from classic works of French literature to various kinds of nonfiction prose and technical writing. In most cases, standard scholarly editions were used when converting the text into machine-readable form, and the data contain page references to these editions. The ARTFL database is one of the largest of its kind in the world. The number, variety, and historical range of its texts allow researchers to go well beyond the usual narrow focus on single works or single authors. The database permits both the rapid exploration of single texts and intertextual research of a kind virtually impossible without the aid of a computer. With the introduction of ARTFL on the Web, researchers have a new and easier way to access this database.

"The Electronic Text Corpus of Sumerian Literature"

Jeremy Black (jeremy.black@oriental-institute.oxford.ac.uk) *and*
Eleanor Robson (eleanor.robson@wolfson.oxford.ac.uk), *Oriental Institute, University of Oxford*

The Electronic Text Corpus of Sumerian Literature (*http://www-etcsl.orient.ox.ac.uk*) is a three-year project underway at the University of Oxford. Its aim is to make accessible, via the World Wide Web, over 400 literary works composed in the Sumerian language in ancient Mesopotamia during the late third and early second millennia B.C. In this talk we shall describe our methodology, focusing on the creation of SGML document type definitions, the development of an Operating Procedure for the project, and issues of transliteration and translation. We shall also discuss some of the problems we have encountered and our plans for the future.

Open discussion: **"Standards for Text Encoding and Markup"**

Moderated by Gene Gragg, University of Chicago, and Steve Tinney, University of Pennsylvania

Topics to consider include: (1) prospects and procedures for adopting encoding standards; (2) the purpose of encoding and the need for flexibility and reusability of encoded texts; (3) deciding what is to be encoded or tagged and what is not; (4) how to represent variants and critical apparatus; (5) grammatical parses, lexical glosses, and grammatical and lexical indices; (6) dictionary and grammar markup; (7) common standards for marking up relatively fixed "content," including the structure of texts and links between texts, versus the diversity of downloadable stylesheets for sharing highly variable modes of presentation of that content.

"Texts and Context: Using XML to Integrate and Retrieve Archaeological Data on the Web"

David Schloen, University of Chicago (d-schloen@uchicago.edu)

Many ancient Near Eastern texts are found in excavated sites or standing monuments, and so share with other archaeological artifacts a determinate geographical, architectural, and stratigraphic context. But to represent effectively the physical context of texts and other artifacts requires an appropriate standardized data model that can be readily implemented on the World Wide Web. In order to integrate information from many different archaeological sites we need a standardized data model that is not overly rigid or prescriptive but still provides a rigorous underlying framework for archaeological data interchange. In this paper a hierarchical "item-based" data model will be presented that is easily implemented as an XML tagging scheme. Using this tagging scheme, dubbed "ArchaeoML," any kind of archaeological data may be represented and delivered to Web browsers in a cross-platform, standardized fashion. The ArchaeoML tagging scheme and the data model it implements will permit seamless dynamic integration and joint querying of archaeological datasets derived from many different sources on the Internet. The benefit of such an approach is that it will make it much easier to retrieve and compare data across diverse sites and regions.

Open discussion: **"What's It Good For? Uses of Electronically Published Texts"**

Moderated by Matthew Stolper, University of Chicago

Topics to consider include: (1) types of linguistic, literary, and historical analyses enabled by electronic publication; (2) how to devise markup schemes to maximize the potential for reusability of data for different purposes; (3) the potential for multisource automated retrieval and analysis (i.e., on-the-fly integration of data from multiple sources on the Internet), as opposed to manual browsing or single-source retrieval.