Access to dead language resources: the contribution of the CDLI

Bob Englund, UCLA

The Cuneiform Digital Library Initiative (CDLI, <http://cdli.ucla.edu/>) is a joint US-German research project dedicated to the digital capture, preservation and dissemination of cuneiform collections that are spread throughout the world. Its host institutions are the University of California at Los Angeles and the Max Planck Institute for the History of Science (MPIWG), Berlin, and its work has in the past been funded, beyond these two institutions, by the National Science Foundation's Digital Library Initiative, and is currently supported by grants from the National Endowment for the Humanities and the Institute for Museum and Library Services.

This is a project that has "grown with its data". Work on the digitization of earliest cuneiform inscriptions dating to the 4th millennium BC began in the late 1970's on a punched card machine of the computing center at the Free University of Berlin, and since those days both the scope and means of data capture have progressed through the various generations of soft and hardware that are known to the participants of this conference. The digital resources of the CDLI currently include 1.4 TB of raw and archival data, and ca. 1.2 million lines of ASCII-formatted transliterations of cuneiform texts documenting nearly 140,000 cuneiform texts ("transliteration" is the technical term referring to a one-to-one transcription of signs in a cuneiform text to computer-readable text). Collaborating programming staff at MPIWG, Harvard, the University of Pennsylvania, and UCLA have written and posted to CDLI documentation pages (<http://cdli.ucla.edu/cdli_methods.html>) text describing our means of text and image acquisition that follow standards widely accepted in DL projects but that were tailored to facilitate the capture and online dissemination of Babylonian texts.

We are currently focusing our attention on three areas of lasting concern.

First and foremost, this project serves an academic community of researchers that, like our data, is very widespread. Like the formal collaborators of the CDLI, these users have varying degrees of technical competence, and work under varying financial constraints. We are nonetheless dependent on input from this disparate community to build and maintain a robust general repository of cuneiform, despite the possibly unsound data that we may receive. We are therefore testing online resources that put greater programming and data management burden on centralized and stable committed project partners, while at the same time facilitating the entry by collaborators of data into centrally monitored files. Our data are in three subsets:

1. metadata are maintained in a large, now rather cumbersome database file (Filmaker 7) remotely accessed by Los Angeles and Berlin project editors. The database file is at the same time the web server of CDLI.
2. transliterations in ASCII and XML format, and all interpretive texts that derive from these transliterations, including inline lemmatization (word glossary) markup and markup of composite text witnesses, are currently being processed for entry in a central, Zope-administered repository at our Berlin offices. This archival repository will allow

accredited users to download and lock current files, and upload corrected or augmented files. The user interface allows the following functionality (atf = ASCII text format):

Access

- Browse the archive
- Search the archive
- Show objects from list

Baskets / upload files (A basket is a single atf file which contains a collection of transliterations in atf)

- Show and download baskets (all)
- Upload an atf file / basket
- Show last uploaded baskets

Management

- Manage users (preliminary version, users can be added by using cdli's standard pw), select the role Manager for all new users.
- Show changes by author
- Show last changes by time
- Show locked files

Combining the strict versioning built into this repository with its ease-of-use, we expect that we will be able to substantially accelerate the rate with which limited project staff is able to collate and add to our current transliteration files. Like the metadata server, these text files will be automatically reformatted for upload to live web files.

3. Image files are stored in standard tif and, for vector graphic drawings of cuneiform texts, in Adobe's encapsulated postscript format. We have no current means to expand access to our raw and archival image files for image editors, due primarily to the cumbersomeness of moving large data packages across limited bandwidth. While a full download of all transliteration files, should an editor in Beijing wish it, would involve less than 100MB, a download basket of just ten or twelve image documents could well cross the GB barrier and thus represent a serious burden on many user systems. Thus CDLI servers are dump sites for raw images from digitization efforts in participating collections, and all final image processing and archival file stitching is performed by staff in Los Angeles.

Second, we and many other humanities digital library projects have enjoyed a certain cocoon existence with our modest data sets and closed user communities. Despite OAI metadata guidelines now widely in use that will be a point of discussion at the conference, the problem of making digital libraries a true part of the permanent library landscape is, so far as I see, largely unsolved. Any user worldwide should be able to take an ISBN number of a published book to the local library, or to his home computer, and begin the process of locating this resource. He may or may not have it in his hands in some reasonable time, but he can remain confident, or perhaps uneasy with the knowledge that the resource is out there. How do we put this fear of the divine librarian in the hearts of users of digital libraries? The cataloguing of cuneiform, or hieroglyphic texts shares much with that of publications, but like catalogues of works of art is relegated to an arcane specialist system of identifiers that may or may not map to current metadata standards. In published papers, I do not like to cite digital resources for fear of shooting at a moving target; surely, we must institute permanent web addresses for DL objects that are no less reliable and internationally accepted than the metadata of a particular paper publication. Yet in the end, the pathways to digital objects are more important than the quality of the objects themselves, that even if transitory still serve their purpose. We are currently in a testing program with the Digital Preservation Repository

group of the UCLA research library that will, as I hope, result in the transfer of our archival resources to a University of California-backed system of archival resource keys (ARK) mapped through METS metadata. Once permanent pathways to digital bundles that document physical objects (cuneiform tablets, Egyptian papyri, Ethiopian obelisks) are established, we make much easier the decision of humanities specialists to submit hyperlinked online publications—these are really the academic purpose of our project—an acceptable form of scholarly discourse.

Third, the CDLI is one of many cultural heritage projects that are anxious to offer their resources to the broadest possible audience of users, but particularly to such higher-level repositories as that envisioned by the organizers of this conference. The immediate need for the preservation *and dissemination* of shared world cultural heritage, in which the Middle East plays a pivotal role, is, in the wake of recent events in Afghanistan and Iraq, obvious to anyone who has thought about the problem. We must preserve for future generations in digital format all collections that are threatened with physical destruction by civil, political or economic strife, or by ideological fanaticism. In an active vein, the MEDL and any other cultural heritage DL should serve national and international policing agencies to deter artifact looting and trade, and to assist in the physical repatriation of unprovenienced antiquities, and in the digital repatriation of artifacts removed from the Near East under circumstances now deemed legal by the international community. But in the longer term, an MEDL will serve as a bridge of understanding among diverse communities often shielded from one another by systemic barriers. To be successful, such a library must be highly standardized; must offer innovative online tools that attract and serve the specialist and the informal learner alike; and must embrace the ideals of open access. In their own interest and the interest of their collections, administrators who assume primary curatorial responsibility for museum objects should reject the abuse of commercial copyright to shield cultural heritage artifacts from online dissemination. As the Berlin Declaration, October 2003 (< http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html>) states, “in order to realize the vision of a global and accessible representation of knowledge, the future Web has to be sustainable, interactive, and transparent.”

We follow in this educational component of the CDLI such highly effective sites as “Eternal Egypt” (<http://www.eternalegypt.org/>), and, most recently, the European Union sponsored “Museum with no Frontiers” (http://www.museumwnf.org/>). Given our limited funding, we plan to care for user investment in this effort by implementing a slightly restricted wiki-managed knowledge base of cuneiform studies (see preliminarily <http://cdli.ucla.edu/wiki/index.php/Main_Page>), relegating specific topics to select editors, but keeping user management fairly open. The particular strength of a modest project like the CDLI lies in the fact that we can link, within site, to a broad array of internally managed pages without fear of obsolescence. Achieving permanent Uniform Resource Identifiers (URI) for all cuneiform inscriptions in the project repository, URIs at the least as permanent as the University of California, puts this goal within reach.

Finally, the CDLI has been graced with educational technology partners and sponsors that it might not entirely deserve. Michelle Roper of the Federation of American

Scientists (<http://www.fas.org/main/home.jsp>) in Washington approached us two and a half years ago with a proposal to collaborate with their Learning Federation and with the Walters Art Museum, Baltimore, to develop and put in museum use an educational video game that targets visitors in the age range of 12-16 years. Some conference participants doubtless have teenagers at home who show much greater proficiency playing Halo 2 than finishing their history projects. Educational video games leverage this proclivity of young people to introduce a new learning experience into a technology that they will, with or without our blessing, embrace. This cooperation among FAS, WAM and CDLI was generously funded by the IMLS, and the game, entitled "Discover Babylon," is available in prototype for young visitors of the Walters, but is being expanded to a full version by a commercial game company located in Austin, Texas. It explores new ways to reassemble and restore the material culture of Mesopotamia now spread across many different museum and library collections, to contribute new research on information management, and to encourage interdisciplinary collaboration. We now plan an expansion of the goal institutions and partners in this educational research effort; one promising avenue of product extension lies in voice-over interchangeability. Thus high-expense video components of the game can be modified at marginal cost, and combined with, for instance, Arabic voice-over and Arabic text for kiosk use in Cairo, Damascus or Baghdad.