

New digital tools for Mesopotamian cultural heritage preservation at CDLI

51st Rencontre Assyriologique Internationale (RAI 51)

Cale Johnson (CDLI / UCLA)

Chicago, July 20, 2005

[20 mins. + 5 mins. discussion, 15 slides]

[SLIDE 1: New Digital Tools for Mesopotamian cultural heritage preservation at CDLI]

Good morning. I'm here today to give a progress report on the work being carried out by the Cuneiform Digital Library Initiative—CDLI for short—and in particular on recent advances in both the coverage of our textual corpora, lemmatization work, and improved user interfaces of one kind or another. In honor of the completion of the CAD, which we celebrate at this Rencontre, I'm going to focus on questions of transliteration and lemmatization rather than, say, the construction of online catalogs or the preparation of archival images. And time permitting, I'll also try to sketch out the special role that the construction of transliterational corpora—which is by far the most important work that we do—plays in both lexicography and the preservation of Mesopotamian cultural heritage.

[SLIDE 2: Catalog, Image, and Transliterate]

These three words “Catalog!, Image!, Transliterate!”—taken in the imperative—form both a kind of mantra at CDLI as well as the set of basic operating instructions issued to its worker bees—such as myself—when we head into the “field” to document a collection of cuneiform tablets.

First we build a catalog, then we image every available surface of every single tablet—often using an ordinary desktop scanner—and finally—time permitting—we put together a rough and ready transliteration of as many documents as possible.

The photos on the slide are meant to represent what we typically start with, namely a pile of tablets that—if we are lucky—have each been assigned museum numbers. The tablets in these photos have not been so fortunate and, in a certain sense, we should always keep tablets like these in mind whenever matters of cultural heritage and its preservation are under debate: tablets like these are in a museum and, at least in the case of the two photos on the left, each box of tablets does have a museum number. We cannot say the same for the photo on the right—a room full of confiscated tablets—that will typically include tablets from every imaginable time period and in every imaginable state of decomposition. The important thing to keep in mind is that none of these tablets can be tracked through the markets if stolen; none have been properly documented, so if they do disintegrate we'll never know what they might have said; none can be read by any of us and they are all slowly turning into dust. And if you'll permit me to stand on my soapbox for a moment, the preservation of cultural heritage is not about Indiana Jones, rescuing the artifact from the unscrupulous simply to put it in a box and lock it away in a museum. It is all about documentation, curation and dissemination: the real work of the preservation of cultural heritage is about sitting in museums, documenting and transliterating tablets and building corpora.

[SLIDE 3: First-order markup (ATF > XML)]

Once the basic documentation is complete, we cook up transliterations of the tablets in ASCII Text Format—otherwise known as ATF—such as on the left hand side of this slide. Then we convert it—using a series of programs written by Steve Tinney—into the XML on the right hand side of the slide. At this stage in the process, the use of XML may seem gratuitous to some, but lemmatization and other kinds of second-order markup quickly grow far too complex for the relatively simple syntax of the ATF on the left. So, if nothing else, the conversion into XML lays the groundwork for other kinds of markup that link particular texts to corpora, dictionaries such as the PSD and ultimately prosopographical study and the localization of materials in terms of both time and place.

[SLIDE 4: First-order markup (XML > Lemmatized XML)]

On the left hand side of this slide, I have taken one line from the XML file on the previous page and rearranged it into a hierarchical representation of its structure. On the right hand side, we see that three lines have been added that qualify each “<w>” or “word” level unit: these are lemmatization entries and over the past few months—in cooperation with Steve Tinney and the PSD—we have lemmatized nearly all of the materials in the Ur III corpus. Lemmatized XML represents the culmination of first-order markup and once a text has reached this point, it becomes accessible through not only the CDLI website, but also through the PSD. It forms a self-contained piece of data that can be integrated into a variety of second-order corpora and/or projects such as a PSD or—since we make our DTDs freely available—through any other linguistic site that might want to include our data sets in its research focus, or alternatively into

the kind of collaboratively built Ur III prosopography that we hope to start working on in the near future.

[SLIDE 5: Current progress: the state of the corpora]

So where do we stand today? As you can see, we generally organize our work on the corpora according to the traditional major phases of Mesopotamian history. Two of the earliest corpora (proto-cuneiform and ED I) were already complete at the beginning of the project due to the efforts of the our PI, Bob Englund. The proto-Elamite corpus was put together a couple years ago by Jacob Dahl, who is currently re-editing these texts in the Louvre and will be presenting his findings here at the Rencontre. Our largest corpus and the one that has garnered the most attention from project staff is the Ur III corpus, now numbering over 44,000 tablets in transliteration out of the approximately 65,000 tablets that we have in catalog: this amounts to roughly 68% coverage with over 600,000 lines of transliteration, and slightly less than 2,000,000 words. This is roughly twice as large as the first corpora of Modern English such as the BROWN corpus compiled in the 60s and 70s, but significantly smaller than present-day English corpora such as the British National Corpus with over 100,000,000 words.

More importantly, however, over the past six months, the Ur III corpus has undergone a great deal of standardization prior to its recent lemmatization through the elimination of alternative renderings of the same word. Besides the proto-cuneiform and proto-Elamite corpora, it is by far the cleanest set of materials we make available.

I should also add that we continue to produce, curate and distribute archival images, currently amounting to approximate 640 gigabytes of data, but we expect to reach one terabyte, or in other words, 1,000 gigabytes of image data by the completion of the current phase of the project in 2006.

[SLIDE 6: Current progress:Excluding the Ur III corpus]

If we exclude the Ur III corpus from consideration for a moment and focus on the other corpora, roughly half of the tablets from the 3d millennium exist in transliteration at some stage of our production process: the Early Dynastic period is rather poorly represented with the important exception of the ED IIIb corpus, the Old Akkadian materials are extensive but still in rather bad shape, while the Lagash II materials are nearly complete. The advances are due in large part to a great deal of work that has gone into cataloging, imaging and transliterating materials from the ED IIIb and Ebla corpora, which I'll discuss in more detail in a moment.

[SLIDE 7: Ur III Lemmatization]

The Ur III materials have been the focus of our lemmatization efforts in recent months. As the two pie charts show, in terms of attested forms—for each of which one or more tokens exist—we have achieved 77% lemmatization, which roughly corresponds to the rate of successful lemmatization at ETCSL. In terms of attested tokens—individual words on particular tablets—we have successfully lemmatized approximately 98% of the Ur III corpus.

The next corpus of lemmatized material that CDLI will produce will be the ED IIIb corpus, which I've already begun working on, again, in cooperation with Steve Tinney. The first batch of

lemmatized ED IIIb files will be completed this summer, a second batch which is currently being standardized into ATF will be lemmatized in the fall and a third and final subset including the rest of the as of yet untransliterated ED IIIb materials should be complete by the end of the 2005. With the completion of the ED IIIb corpus, most of the major corpora that provide the basis for lexicographical work in Sumerian, namely the ED IIIb and Ur III materials made available through CDLI and the Old Babylonian corpora produced by ETCSL and DCCLT—Niek Veldhuis' lexical list project at UC Berkeley—will be essentially complete.

In essence we are involved in much the same task that was undertaken by the CAD, but the crucial difference is that in an age of dynamic corpora, we also need a new kind of lexicography, which Steve has been instrumental in organizing. The fundamental difference is that projects that focus on particular bodies of material such as CDLI, DCCLT and ETCSL can produce corpora that feed the Pennsylvania Sumerian Dictionary, while receiving in turn a degree of unification and ease of access that might not otherwise be available.

[SLIDE 8: New front page and single-line search]

I'd now like to turn to some advances in the user interface and search capabilities. This is the newest iteration of the CDLI front page. Besides the merely cosmetic changes, the most important difference in function is the single-line search window at the lower right. The goal of this search window is to act as a universal point of access: the idea is that it takes museum numbers, publication information, author, whatever you want to throw at it and it searches all the major fields in the catalog looking for references. It is a simplistic and rough-hewn search

mechanism that is complemented by an Advanced Search page that allows the user to search particular fields in the catalog.

[SLIDE 9: Graphemic search]

As the name implies, graphemic search allows the user to type in a sign name like IGI and receive results that include not only different readings of the sign, but also complex signs in which IGI is a component. These can be ranked, for example, by frequency within a specific corpus as in the sample of 10,000 Drehem texts that was used for these examples. Ultimately it is hoped that graphemic queries will allow for a variety of advances such as the automatic lemmatization of unfamiliar orthographies, or at least the development of an intelligent system that will aid in the reconstruction of the many, many fragmentary tablets in our corpora. It comes down to a rather simple procedure: defining textual atoms, namely graphemes, within particular domains: IGI followed by UR doesn't usually mean "the face of the dog," and in the Drehem subcorpus it can—in all likelihood—automatically be assigned the reading HUL without a second thought.

[SLIDE 10: Boolean/User-defined domain]

The last new and improved search functionality that I would like to introduce is "boolean" and "user-defined domain" searches. These new mechanisms allow for search terms to be linked with Boolean operators such as AND, OR and NOT and, within the same search field, it also allows the user to define the domain in which the search terms should occur. In the first example at the top of the slide, the search is for mu AND du3 WITHIN 3 LINES, yielding a result such as the one on the right. In the second example, if you were looking for texts in which someone named

Geme-Enlila, for example, is described as a sister, nin9, or a spouse, dam, a Boolean search would allow you to find both the these results, even though neither contains all three terms.

[SLIDE 11: Recent initiatives and cooperative efforts]

Now, the part of the project that I myself mostly work on is the building and extension of new corpora, so I would like to say something about our recent efforts in this area before moving on to questions of cultural heritage. Work on Ebla began with a week-long cataloging session, where Alfonso Archi, other members of the ARET team and I put together a digital catalog on the basis of a variety of sources; this catalog formed the basis for further work and was incorporated into the CDLI database last summer. Recently we have also reformatted approximately 1400 transliterations derived from the ARET volumes. These now form the core of our Ebla corpus, which we will be returning to the ARET team this summer for correction and expansion.

Unlike the Ebla materials which are almost entirely managed by the Italian mission, the Old Sumerian materials from the ED IIIb period represented a much more difficult challenge. Work on both ED IIIb and Old Akkadian catalogs began in the summer of 2002 and extended over a series of meetings in 2002 and 2003 between myself, Aage Westenholz and Walter Sommerfeld in Copenhagen and Marburg. At roughly the same time, we began the reconfiguration of the transliterations of over 1,800 Old Sumerian tablets provided by Gebhard Selz, Bram Jagersma, and Remco de Maaijer. These formed our first subset of ED IIIb transliterations, which I'm now in the process of lemmatizing. In January, a second subset was put together based on the Old Sumerian royal inscriptions harvested from RIME 1, which Douglas Frayne contributed to the

effort, as well as a variety of transliterations produced by, among others, several graduate students at UCLA. As I mentioned already, we expect to have all available image and transliteration files available by the end of 2005.

[SLIDE 12: Monitoring the text-artifactual record]

So why do we go to all the trouble? Or, in other words, what does what we do have to do with the preservation and reconstruction of cultural heritage? As a partial answer to these kinds of questions, let me present two cases that illustrate what the tools we make available can do. The first example took place a couple years ago, when Bob Englund noticed several proto-cuneiform tablets that were being auctioned by Bonham's in London. It may strike some in the audience as odd, but Englund and staff at CDLI do regularly monitor not only the traditional markets, but also new media of circulation such as eBay. Given the images of the tablets published in the auction book, Englund did a search, located the tablets in our database and also noticed that the same tablets had been offered for sale in Amman in September of 2000. It is thus likely that these were a few of the growing number of Late Uruk, Early Dynastic and Old Akkadian tablets, presumably from Umma and its vicinity, that have been flooding the markets since the 1990-91 Kuwait War. The important thing in this case is that only because of such basic tools as catalogs, archival images and transliterational corpora now in place was it even possible to know which tablets these were, and where they had been until recently.

[SLIDE 13: Rebuilding scattered archives]

The other example of the real power of the corpus comes from the Ur III period. It has been known since early work by Struve that the larger Ur III administrative documents were often

based on dozens if not hundreds of single transaction receipts. Now one of the more interesting things we like to do whenever we come across a little receipt like the one in the lower right of this slide—this is one of the hundreds of Banks tablets that were sold throughout the US in the first few decades of the twentieth century—one of the first things we like to do is to plug its lines into the data set and see if we can find any summary tablets that were based on it. In this case, it turned out to be one of the fourteen known receipts that formed Erlenmeyer 152. For those of you interested in the case, it is fully documented in the Cuneiform Digital Library Journal 2003:1, which is available on our website.

Although this is a rather simple example of the power of the corpus, it does offer a concrete example not just of the preservation of cultural heritage, but of its virtual reconstruction. This kind of corpus-driven reconstruction only really becomes possible for those of us who cannot keep 600,000 lines in memory when relatively full data sets are publicly available in a human readable form. Simply put, that is our goal. We want every single tablet on earth in catalog, archival image and transliteration freed from the various constraints on both scientific work and the use of cultural heritage and made available through the internet.

[SLIDE 14: Lemmatization and second-order markup]

Only once the primary data are freely available do the really interesting kinds of second-order markup become feasible. Once the transliterations are lemmatized, then combinations of lexeme, subcorpus and various kinds of catalog data allow for all manner of higher-order corpora and uses to be effectively implemented.

In the near future, we at CDLI hope to make significant advances in several areas. Our next major initiative—analogue to the efforts we have put into the ED IIIb and Ebla corpora over the last couple years—will be the development of corpora of syllabically written Semitic languages covering the first millennium of the attested history of the Semitic languages from Ebla to the end of the Old Babylonian period. Two areas of second-order markup in which we hope to make substantial progress in the coming years are the development of a mechanism for collaborative work on the prosopography of the Ur III empire and the integration of the morphosyntactic parsing that is being applied to Sumerian materials at the PSD in cooperation with the Penn Treebank into, again, a collaborative system for the description of Sumerian morphosyntax and the various grammatical theories that have been applied to Sumerian over the hundred years since Poebel. Currently we are investigating the use of Wiki-technologies best known from Wikipedia and other online collaborative projects, which allow anyone to edit any particular page of content, while maintaining a history of changes that editors can use to roll back or “undo” changes that are malicious or uninformed.

[SLIDE 15: Last page]

In conclusion, I’d like to say that we as Assyriologists face nearly the same choice that linguists face on an almost daily basis: do we pursue our own private research agenda, or do we document some of the languages that—within our lifetimes—will no longer exist. In our terms, do we push a theoretical agenda in the hope of invigorating the field, or do we head into the trenches with our laptops and scanners? Well, no surprise here, I think we are obligated to do both. In my view, the best scenario is to continue to require that both graduate students and their mentors spend time in the trenches building and extending these corpora, but the real intellectual content of

Assyriology should not solely reside in the publication of primary sources. If the field of Assyriology is to have any relevance in this new century, we must as a profession come to value synthetic treatments of particular historical or linguistic topics over the mere publication of raw materials, materials that will hopefully come to exist in publicly available corpora maintained and extended by projects such as CDLI.