From Cuneiform Archives to Digital Libraries: The Hermitage Museum Joins the Cuneiform Digital Library Initiative♣

Natalia Koslova
Hermitage Museum, St. Petersburg
nkoslova@online.ru

Peter Damerow
Max Planck Institute for the History of Science, Berlin
damerow@mpiwg-berlin.mpg.de

Abstract

The Hermitage Museum in St. Petersburg was one of the first museums to join the Cuneiform Digital Library Initiative (CDLI), an open source initiative aiming to create a digital library of cuneiform documents of the 3rd millennium B.C. The Hermitage collection, consisting of approximately 2,000 tablets of this period, has been digitized and became an integral part of the growing virtual CDLI library. This virtual library provides an unprecedented method of accessing the oldest written sources of mankind. The library contains images, standardized transliterations, and working environments based on language technology. To develop and implement a digital library of such unusual objects challenges both the traditional research standards of the highly specialized discipline of Assyriology and the community of information management specialists. It requires an intensive cooperation between IT specialists and scholars of cuneiform writing unparalleled in traditional research and development activities.

1 The cooperative project

The Cuneiform Digital Library Initiative (CDLI) represents the efforts of an international group of Assyriologists, museum curators and historians of science to make freely available through the Internet images and the contents of cuneiform tablets dating from the beginning of writing, ca. 3200 B.C., until the end of the 3rd millennium. The CDLI thus deals with the most ancient written documents in the history of mankind. A substantial collection of about 2,000 cuneiform texts from this period is kept at the State Hermitage Museum in St. Petersburg. As early as June 2000 the Hermitage Museum, represented by its director Prof. Michail B. Piotrovskij, agreed to the open access policy of the CDLI and decided to join this initiative. The digitization of the Hermitage Museum collection became one of the first museum projects to be supported by the CDLI [1].

The CDLI became the focus of a growing network of cooperation between research institutions, universities and museums. This network offers unprecedented research conditions for scholars of all scientific disciplines dealing with the cultural heritage of the ancient Near East. The CDLI is, in particular, associated with the Philadelphia Sumerian Dictionary Project, which is compiling the first comprehensive dictionary of the Sumerian language. This dictionary will be accessible in electronic form through the Internet. Its entries will be linked to the body of sources provided by the CDLI.

The CDLI was founded and is institutionally supported by the University of California at Los Angeles and by the German Max Planck Society. These institutions will ensure the long-term availability of the CDLI's outcome until other institutions start to realize the potential of information technology in establishing a new infrastructure for open access to the cultural heritage of mankind and guarantee its longevity.

2 The objects of digitization

2.1 Cuneiform writing

Given the unusual nature of the objects documented in the cuneiform digital library, it may be useful to describe them in some detail before addressing the technicalities of their digitization.

The cuneiform script was mainly written on clay tablets. It received its modern designation from the wedge-shaped stylus impressions composing its signs. Cuneiform writing was invented in the ancient Near East at the end of the 4th millennium B.C. It was then used for more than three millennia to write texts in almost all the languages spoken there at that time, above all Sumerian and Akkadian, two major languages of the ancient Mesopotamian civilization, the center of which was located in what is now modern Iraq. Clay is an extremely durable material and for this reason an enormous amount of texts have survived from all the lengthy periods in which cuneiform writing was used.

Proceedings of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia, 2003

2.2 The early history of cuneiform writing

The first documents, written on clay tablets in so-called proto-cuneiform script, appear at the end of the 4th millennium B.C. in the ancient city of Uruk, modern Warka, in the south of Mesopotamia. These texts are mainly administrative records revealing a developed system of retaining information and calculating goods, human and material resources.

On the whole we have about 5,000 proto-cuneiform archaic texts dating back to 3200-2800 B.C from different archaeological sites. After a gap of as long as two centuries, the next period from which texts dating approximately to 2600 B.C. survived is called the Fara period, relating to the modern name of the site where the first and very important texts of this period were found. The Fara period is represented by about 2,000 cuneiform documents now known definitively to be written in the Sumerian language.

The question of whether the earlier proto-cuneiform texts were also written by Sumerians, suggesting that the Sumerians invented the script in Mesopotamia, is still under discussion. It is characteristic of this incipient proto-writing that no representation of spoken language could be identified so far. But in comparing the proto-cuneiform archives with those of the Fara period, we can observe a number of continuities in the cuneiform system of writing and in the form and content of tablets which attest to the fact that at least the written tradition found in 3rd millennium Mesopotamia was uninterrupted.

The history of the Mesopotamian written tradition in the second half of the 3rd millennium B.C. is conventionally divided into three periods:
1. Old Sumerian (ca. 2500-2400 B.C.) represented by about 2,000 cuneiform texts mainly from the ancient Sumerian city of Girsu (modern Telloh),
2. Old Akkadian (ca. 2350-2200 B.C.) represented by about 6,000 cuneiform texts from different sites written partly in Sumerian and partly in the Akkadian language,
3. Neo-Sumerian (ca. 2150-2000 B.C.) represented by about 100,000 cuneiform texts from different sites, again written mainly in Sumerian.

Although we also have school texts, so-called lexical lists, from as early as the archaic period representing the very beginning of writing, most of the cuneiform texts of the 3rd millennium B.C. are administrative documents, literary texts, and royal inscriptions.

2.3 The decline, disappearance and rediscovery of cuneiform writing

The last political state whose administration was conducted in the Sumerian language, the so-called third dynasty of the ancient city of Ur, collapsed around 2000 B.C. After that Sumerian was gradually forced from everyday use by the East Semitic Akkadian, and thus became a dead language. The cuneiform system of writing, however, continued to be used in Mesopotamia and neighboring regions and was adopted into other languages until it finally died out in the first centuries A.D. It fell into oblivion and was widely unknown in the Greek-Roman world, which would later determine Western culture.

Cuneiform writing was rediscovered in the 18th century by travelers who noticed inscriptions on rock faces which survived from the Assyrian and Persian empires of the 1st millennium B.C. Attempts to decipher such inscriptions started immediately but failed for a long time to come. Cuneiform writing was essentially deciphered only in the 19th and early 20th century when innumerable cuneiform texts were distributed throughout the world. While most of the languages of the cuneiform texts are well known today, the earliest of them, the Sumerian language of the 3rd millennium B.C., still raises many unsolved questions. The creation of a reliable dictionary of this language and the reconstruction of the development of the cuneiform signs from proto-cuneiform to standardized sign forms are some of the main challenges facing current studies, which call for innovative implementations of information technology.

2.4 The extant cuneiform sources

The cuneiform tablets to be digitized come from ancient archives that were either excavated by archeologists or, to a greater extent, plundered by dealers of antiquities and their "helpers." Although the total number of these tablets can only be estimated, it is surely in excess of one million.

In the current phase of the project the CDLI is limited to the cuneiform archives of the 4th to 3rd millennium B.C., which mostly contain Sumerian administrative records. The number of extant tablets inscribed during this period of early state formation amounts to approximately 120,000 tablets and tablet fragments. These tablets were discovered in archives unearthed by various excavations starting in the 1880s: the French excavations in Telloh (ancient Girsu), the German excavations in Warka (ancient Uruk) and Fara (ancient Shuruppak), the British excavations in Tell Muqayir (ancient Ur), the British-American excavations in Jemdet Nasr (ancient name unknown), and the American excavations in Abu Salabih (ancient name also unknown).

Between the regular campaign seasons thousands of clay tablets were plundered by local people. Some of the sites were excavated solely by clandestine diggers who then sold their booty to antiquity dealers. In this way thousands of cuneiform texts found their way into European and American museums and private collections and are now dispersed throughout the world.

2.5 The virtual reunification of ancient archives

One of the CDLI's primary goals is to reconstruct the ancient archives and join the widely dispersed but closely related documents in a virtual library. The state archive of the Old Sumerian Girsu, for instance, contained at least 1,600 tablets, which are today scattered in collections in Paris, Istanbul, Berlin, Baghdad, New Haven, St. Petersburg, and many other places. It therefore often happens that closely related tablets, which were kept in ancient times in one and the same basket,

are now dispersed in different places and have lost the textual context that made the content of the individual text understandable.

An example may illustrate the consequences and demonstrate the importance of easy access to different collections, even for studying only one particular text. Sumerian administrative documents of the third dynasty of Ur are usually dated according to special formulas at the end of the text, for instance "Year when Amar-Suen became king." However, there are some ambiguous cases. The formula "Year when the place of Simurum was destroyed" can refer to both the 25th year of the reign of Shulgi, the second and in many respects the most prominent ruler of the dynasty, and the 3rd year of the reign of Ibbi-Suen, the last ruler of the dynasty.

One of the documents in the Hermitage collection bearing this ambiguous formula mentions an official named Ur-Ishtaran who received a number of sheep, probably in order to fatten them. The same person is acting as a shepherd in several texts dating back to the period between the 25th and the 46th years of Shulgi. Another document bearing the same formula was sealed by an official named Ushmu. This seal is attested in texts from the 8th year of Amar-Suen, the son and successor of Shulgi, to the 2nd year of Ibbi-Suen. Accordingly, the first document is supposed to come from the 25th year of Shulgi, the second one rather from the 3rd year of Ibbi-Suen. But the texts that allow for this conclusion are now kept in the British Museum and in a museum in Istanbul, in the universities of Yale and Princeton, in Rome, Leiden, Geneva and many other places. Using traditional methods, an enormous amount of scholarly work was necessary to identify the texts and to draw the conclusion. In contrast, the virtual library built up by the CDLI makes it possible to have all these texts on the screen within a few minutes of searching.

The virtual reunification of the ancient cuneiform archives not only facilitates philological research on the languages of cuneiform texts, but also makes them meaningful for scholars of other disciplines working on subjects related to their content, in particular for scholars who study the economic, social and political history of ancient Mesopotamia. Moreover, a well-designed digital library makes these sources useful for educational purposes and interesting to the public. Digitization also protects ancient texts from the misfortunes of modern history. Digital images cannot substitute the real tablets, but the extensive electronic documentation of the tablets makes each one easily identifiable. The recent tragic plundering of the Iraq Museum would have been a little less fatal had its collections already been digitized as part of an open "Web of Culture" representing the cultural heritage of mankind.

3 The cuneiform collection of the Hermitage Museum

The Hermitage cuneiform collection contains 1,945 administrative texts of the 4th to the 3rd millennium B.C. The majority of them (1,579 tablets) are Neo-Sumerian; the rest includes two proto-cuneiform tablets, 343 Old Sumerian documents from Girsu and 21 Old Akkadian texts from different sites. None of these tablets come from regular archaeological excavations.

The first acquired cuneiform tablets were purchased by the Royal Hermitage Museum at the end of the 19th century. In the museum archive there is a report from the chief curator of that time about the purchase in 1898 of "various remarkable antiquities" from a well-known French dealer named M. Sivadzhan. Among these antiquities were 78 tablets with cuneiform inscriptions from Babylonia, the majority without doubt Neo-Sumerian documents. The next acquisitions were made by the State Hermitage Museum after 1917. Some tablets were purchased, some of them simply confiscated from private collectors.

The main part of the Hermitage cuneiform collection comes from one of Russia's most famous private collections, that of N. Likhachev, a prominent Russian historian who was interested in all kinds of objects from Russian icons to ancient manuscripts, Egyptian papyri and Babylonian clay tablets. In most cases it is almost impossible to trace how the tablets found their way into Likhachev's collection.

According to some private notes in Likhachev's archive he purchased Neo-Sumerian texts from Telloh mainly from Sivadzhan in Paris at the end of the 19th century. Some remarks in Likhachev's private correspondence indicate that the so-called "messenger texts" from Djokha (ancient Umma, Neo-Sumerian period), that are lists of rations for officials travelling between the administrative centres, were also acquired in Paris from Elias Géjou between 1900 and 1914. Another source was Naaman in London. Since 1904 Drehem (a prominent site of the Neo-Sumerian time) was often mentioned in the letters of Géjou to Likhachev as the provenance of the tablets.

Before 1917, when Likhachev was the owner of the collection, it was open to scholars who worked on the decipherment of cuneiform writing. Two prominent Russian Assyriologists of that time, M. Nikolsky and V. Shilejko, worked on texts at Likhachev's residence. In 1915, Nikolsky published about 1,500 texts from this collection.

It is not exactly known what happened to Likhachev's cuneiform collection after the revolution. At the end of 1917 more than 1,300 tablets were moved to Moscow, among them nearly all tablets published by Nikolsky in 1915. In 1918 the remaining collection became public property and Likhachev's home in St. Petersburg was transformed into a museum under his direction, the State Museum of Paleography. Nevertheless, in 1919 Likhachev was still able to sell some cuneiform tablets to the State Department of Arts and Antiquities, possibly the tablets that were already in Moscow. Many of them later found their way into the Pushkin Fine Arts Museum in Moscow.

This State Museum of Paleography in St. Petersburg existed until 1938. After its closure, the objects were moved to the Hermitage. After 1938 only a few more tablets were purchased from individuals.

4. The digital library

4.1 Overview

The digital library of the CDLI is a growing network of data representing cuneiform tablets, their structures, their contents, and results of scholarly work on them. The library provides a working environment, which facilitates scientific work on these data. All services are freely available on mirrored websites in Los Angeles and Berlin [2].

Currently, the following facilities are provided within the framework of the virtual library of the CDLI:

1. a comprehensive catalogue of cuneiform tablets,
2. digital images of original tablets,
3. scanned autographs of tablets,
4. text-coding standards for cuneiform script,
5. transliterations of tablets,
6. tools for handling the digitized sources,
7. publications with links to digitized tablets,
8. show cases for specific collections,
9. educational materials.

In the near future, ongoing work of the project will furthermore provide a comprehensive list of the cuneiform signs of the 3rd millennium B.C. documenting their development from proto-cuneiform to developed cuneiform writing. Furthermore, language technology is being developed for the morphological analysis of Sumerian and Akkadian grammatical forms in order to provide translation aids and automatic linking of the transliterations of cuneiform tablets to the forthcoming electronic dictionary of the Philadelphia Sumerian Dictionary Project [3].

4.2 The CDLI catalog

The CDLI has developed a format for a common catalogue of all collections, which unifies published catalogues into one publicly accessible database and completes and improves the catalogue entries during the work on the individual collections. This catalogue, which is continuously expanded, currently contains information about more than 75,000 tablets dating to the 3rd millennium B.C. It contains data on the archaeological provenience and chronological classification, on physical characteristics, the collection they belong to, and on the publication history of the registered tablets.

4.3 Digital images of original tablets

The display of digital images taken from the original tablets is an essential difference of cuneiform digital library to any traditional publication of cuneiform texts, which, at best, present some example photos together with the representation of the tablets by drawn copies and/or transliterations.

Since the original tablets are valuable and usually cannot be moved out of the collection, the images must be captured in the institutions where they are kept. This is the main reason why sophisticated technology such as three-dimensional scanning cannot be used as long as the required equipment is not easily transportable. The images are captured instead using rather common two-dimensional scanners, which are temporarily brought into the museums and collections by the CDLI.

The digitization of approximately 2,000 cuneiform tablets of the Hermitage Museum pursued in the year 2000 was one of the CDLI's pilot projects in the phase developing the applied methods.

Images are captured from all six sides of the tablets, usually with a resolution of 600 dpi, and stored as TIFF raw images. The post processing of these images includes combining all six images of a tablet in one composite archival image which is the starting point for any later processing such as enhancing the readability and converting the images into JPG compressed formats for display on the CDLI Internet websites. Before this time-consuming processing is finished, images of the obverse and the reverse of the tablets are already preliminarily displayed

Currently, about 1,300 final composite images are accessible through the Internet. Further 6,800 unprocessed images are accessible in preliminary representations of tablets. Some 2,000 further images are in the process of preparation for display.

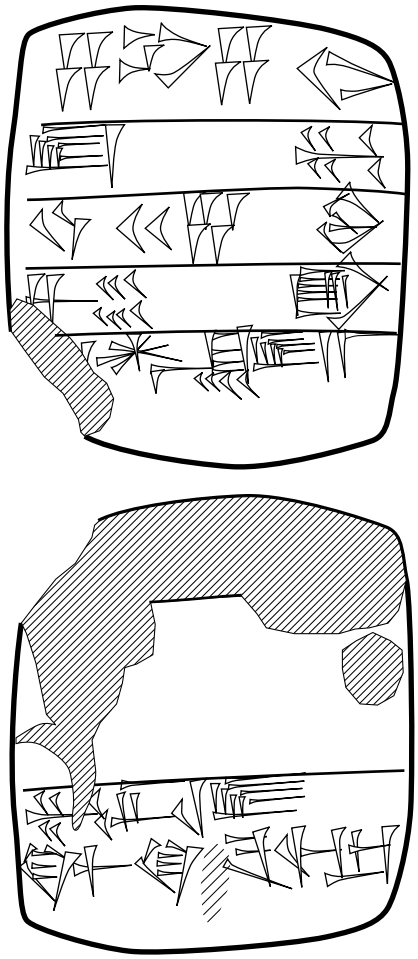

Figure 1: Drawing (2:1) of a slightly damaged cuneiform tablet [4].

4.4 Digitized drawings

Copying tablets by drawing is the standard way of the primary publication in cuneiform studies. The drawn copies are not neutral reproductions of the tablet but rather the first step of interpretation. To draw what seems to be visible on a tablet requires decisions about the nature of any detail of the surface texture of the tablet, whether it represents a natural detail of its material surface, an effect of some kind of damage, or a trace of the impression of the stylus with which the tablet was written. Drawings of tablets can neither be replaced by photos, nor by digital images. This is the reason why the CDLI complements the scanned images of original tablets with scans of existing drawings of them.

Since it is technically much easier to scan drawings than to scan the original tablets, the number of scanned drawings already accessible on the CDLI websites is actually much higher than the number of scans of the original tablets. Currently about 32,200 composite images of drawings scanned with a resolution of 150 dpi can be accessed at these websites.

4.5 The definition of a document type description for an archival XML format for cuneiform text transliterations

The most important interpretative step in the studying of cuneiform documents is their transliteration. The difficulty of this step results from the specific character of the cuneiform writing system. In principle, the cuneiform script was a syllabic writing system. The graphemes represented certain phonemes consisting of consonants and vowels. But this representation was in many respects ambiguous. Cuneiform writing involved polyphony and homophony, that is, most graphemes had more than one phonetic value, and different graphemes could represent one and the same phoneme. Moreover, the same graphemes were also used as logograms representing entire words. They were further used for phonetic glosses and for determinatives. In this function they did not belong to the part of the text that represented spoken language. They were rather added in order to disambiguate the use of other graphemes, either phonetically or semantically. Furthermore, in the 3rd millennium elements of two completely different languages, Sumerian and Akkadian, were often simultaneously contained in the same text. The situation is made worse by the fact that usually the tablets show various types of damage that often make parts of the text unreadable. To transliterate a cuneiform text thus includes for each recognizable grapheme a decision about its actual function and phonetic reading in the specific context is was used.

A transliteration system able to document, on the one hand, all these decisions and, on the other hand, the actual graphemes on the tablet has to be correspondingly complex. Accordingly, the transliteration of cuneiform texts is based on sophisticated traditional system of transliteration conventions using indices and diacritics in order to distinguish homophones and specific signs and special characters for indicating what traces of cuneiform signs justify the specific reading proposed by the transliteration.

This system long resisted any attempt to be mapped into electronic data formats. When the advantages to use computer technology became too obvious to be further neglected, formatting capabilities of text processors and specifically developed fonts with strange characters dominated its application for representing transliterations of cuneiform texts. Such methods were suitable for adapting the display and printing of electronic transliterations according to traditional conventions, but they were of little use for any sophisticated text processing using language technologies.

Once this became obvious, attempts have been made to simplify the system of conventions in a way that pure ASCI text could represent the transliterations. This, however, made things even worse. The simplified ASCI transliterations lost so much of the precision of the traditional system that they could not be used any longer for traditional publications, nor did they provide favourable conditions for the development and application of sophisticated language technologies.

Right from the start, the CDLI has worked intensively on a more adequate solution to the problem of electronically coding cuneiform transliteration. The major outcome of this solution is an XML document type description, developed in cooperation with the Philadelphia Sumerian Dictionary Project, which suitably captures the traditional transliteration conventions and makes the often only tacitly applied rules explicit in a machine-readable form. At the same time this XML format standardizes the various transliteration systems applied by scholars according to personal and often somewhat idiosyncratic preferences. It makes it possible to integrate the work on a comprehensive and reliable Sumerian dictionary with the editorial work on the primary sources it has to be built on.

The XML document type description of the CDLI covers a wide range of different aspects of cuneiform writing such as:

1. the nature of the object carrying cuneiform writing,
2. the identification of the object by metadata,
3. the format and subdivision of the object,
4. the location of cuneiform writing on the object,
5. the location of damage to the object,
6. the graphemes on the object,
7. the actual phonetic values of the graphemes,
8. the role of graphemes in composite graphemes,
9. the words containing the graphemes,
10. the role of graphemes as glosses etc.,
11. the numerical notations composed of graphemes,
12. the actual values of the numerical notations, and
13. the errors of the scribes.

The document type description of the CDLI [5] characterizes a cuneiform text by an XML hierarchy of about 20 different elements and their attributes. The top-level element <texts>, representing a collection of cuneiform texts, and its tree of child elements <text>, <object>, <surface>, <column>, and the line element <l> determine unique IDs, classify the objects into types, and specify the location of each line of translit-

eration on the object which is usually a tablet, an envelope, or a prism. The content of the line element consists of parsed character data which are further structured by a hierarchy of inline elements such as <n> for numerical notations, <w> for words, <gloss> for determinatives and phonetic glosses, <g> for graphemes, <cg> for compound graphemes and so on. Furthermore, elements such as the element <noncolumn>, the non-line element <nonl>, and the non-grapheme element <nong> provide tools for standardized descriptions of peculiarities such as special rulings, empty space, or damages.

The transliteration of the first two lines of the cuneiform tablet depicted in figure 1 may serve as a simple example of the CDLI archival XML format:

```
<text n="P212341" id="P212341" xml:lang="sux">
<object type="tablet" id="P212341.1">
<surface type="obverse" id="P212341.1.1">
<column n="0" id="P212341.1.1.1">
    <l l="O0001" n="1" o="1" id="P212341.1.1.1.1">
        <n><w><g>4(disz) </g></w></n>
        <w><g>gu4</g></w>
        <n><w><g>4(disz) </g></w></n>
        <w><g>ab2</g></w>
    </l>
    <l l="O0002" n="2" o="2" id="P212341.1.1.1.2">
        <w><g>e2</g></w>
        <w><g>muhaldim</g></w>
    </l>
    <l l="O0003" n="3" o="3" id="P212341.1.1.1.3">
        <w><g>u4</g></w>
        <n>
        <w><g>2(u)5(disz) </g><g>kam</g></w>
        </n>
    </l>
    <l l="O0004" n="4" o="4" id="P212341.1.1.1.4">
        <w><g>zi</g><g>ga</g></w>
    </l>
    <l l="O0005" n="5" o="5" id="P212341.1.1.1.5">
        <w>
        <g breakage="damaged">ur</g>
        <g gloss="pre">d</g><g>en2</g><g>lil2</g>
        <g>la2</g>
        </w>
    </l>
</column>
```

4.6 The definition of an ASCI input format for cuneiform text transliterations

The XML document type definition captures in a machine-readable format the highly sophisticated traditional conventions of cuneiform scholars. It meets the requirements for future development of language technology for the Sumerian and the Akkadian languages, thus inevitably, however, becoming too complex to be used as an input format for Assyriologists transliterating cuneiform tablets.

For this reason the CDLI has defined a second transliteration format, which is an ASCI text format (ATF) that reconciles the requirements of processing data electronically with approved methods and conventions of cuneiform specialists. For all elements and attributes of the archival XML format, the definition of ATF contains corresponding simple, but strict and unambiguous rules on how to type the transliterations. These rules are derived from common habits used to express the complex syntax and phonetic of cuneiform writing, replacing, however, conventions such as font styles and non-ASCI characters, which are incompatible with pure ASCI coding. Special ASCI characters, which are not required for the transliteration itself, are used to mark information complying with elements of the archival XML format. The complexity of the XML format is reduced by suitable defaults in accordance with tacit rules of traditional transliteration conventions.

The cuneiform tablet of figure 1 may illustrate the feasibility of the defined format. A traditional publication would render its lines in the following way:

CMAA 002-C0005 (=P212341)
obv.
1) 4 gu_4 4 ab_2
2) e_2 muhaldim
3) u_4 25-kam
4) zi-ga
5) ´ur`- $^{d}en_2$-lil_2-la_2
rev.
1) [ ]
2) (blank)
3) mu-us_2-sa ki-mash^{ki} ba-hul

The corresponding ATF transliteration defined as ASCI standard by the CDLI reads:

```
&P212341
#CMAA 002-C0005
@tablet
@obverse
1. 4(disz) gu4 4(disz) ab2
2. e2 muhaldim
3. u4 2(u) 5(disz)-kam
4. zi-ga
5. ur#- {d}en2-lil2-la2
@reverse
$ broken
$ blank
1. mu-us2-sa ki-masz{ki} ba-hul
```

While this rendering is obviously close to traditional conventions, it is entirely explicit and strictly standardized so that it can be automatically converted into the archival XML format from which the ATF rules are derived. This conversion is realized, in cooperation with "The Archimedes Project" [6], by a web service [7]. The transliteration file has to be uploaded using an Internet browser. The file is then automatically converted into an XML file, which can be downloaded. This web service is complemented with an XML validation procedure, which helps to correct formatting errors of input data and to disambiguate and correct legacy data that have been reformatted with errors into ATF. This validation procedure is again realized by a web service [8].

4.7 Computer-assisted translation and interpretation

Cuneiform tablets document the culture of early cities and empires in the Near East better than other known textual sources do for any other early civilization in the world. The contents of these tablets range from mundane receipts and running accounts of a hypertrophic bureaucracy, to the incomparable verbal art of various literary and religious genres, to scribal exercises representing the earliest manifestations of scientific thinking. Such contents are of substantial interest to scholars of many disciplines as well as to the public in general.

At present however, the digital library of the CDLI primarily addresses the demands of cuneiform specialists. The main reason for this current bias towards their needs is the language barrier between modern communication and the ancient cuneiform documents written in Sumerian and Akkadian.

This is particularly the case for administrative documents. While the literary texts aroused the interest of scholars as well as the general public early on and translations are widely available even through the Internet [9], the administrative documents have been long considered trivial or at least less interesting.

The administrative texts are, in fact, philologically simple and, isolated from their original context, not very informative. The translation of the same tablet depicted in figure 1 may illustrate this:

obv.
1) 4 oxen 4 cows
2) (for the) household (of the) cook,
3) 25th day,
4) booked out:
5) Ur-Enlila

rev.
1) [ ]
2) (blank)
3) Year after: "Kimash was destroyed".

The text is in many respects characteristic of administrative documents. Literary texts represent spoken language. Consequently, they are philologically complex. Administrative documents, in contrast, represent economic activities. They widely use technical terms and standardized designations for objects, agents, and actions. Many of these terms are not used in inflected form. Philologically more complex phrases such as year names are often stereotyped. The conditions are therefore favourable for the development of computer programs, which are able to analyze administrative documents and link them to dictionaries, thus making the content of such documents accessible also for scholars other than cuneiform specialists.

There are two ways to establish such language support. The first is to create a program that analyses inflected forms using grammatical rules in order to determine the root so that the words can automatically be linked to dictionary entries. The second is to collect the empirically inflected forms of as many words as possible. Both methods complement each other. The second method provides the basis for the construction of effective rules for programs realizing the first method. This is true in particular for texts such as the cuneiform administrative documents, which display specific arrangements of technical terms rather than typical structures of spoken language. It turned out that, in fact, simpler rules apply to the formation of these texts than to that of literary texts.

The Hermitage collection serves as a model case for realizing such a combined strategy. Traditional publications of transliteration are often combined with carefully produced glossaries (see e.g. [10]). The work on such glossaries for the publication of transliterations of texts from the Hermitage collection is pursued within a computer-assisted working environment, which registers all individual decisions about inflected forms. Thus a data set of relations between roots, inflected forms and translations is built up which serves as a basis for the derivation of rules for electronically classifying administrative documents and determining their content.

4.8 The online publication of interpretations

The digitization of cuneiform tablets and of transliterations of their content not only provides easier access to these sources than traditional publications but also the possibility of linking them electronically with interpretations. In order to realize this potential in an effective way, the CDLI has set up an electronic journal and a bulletin with direct links to the digitized sources [11]. *The Cuneiform Digital Library Journal (CDLJ)* is a non-profit, refereed electronic journal for cuneiform studies. It seeks substantive contributions dealing with the major themes of the Cuneiform Digital Library Initiative, that is, text analyses of 4th and 3rd millennium documents (incorporating text, photographs, data, drawings, interpretations), early language, writing, paleography, administrative history, mathematics, metrology, and the technology of modern cuneiform editing. *The Cuneiform Digital Library Bulletin (CDLB)* publishes short notes that deal with specific topics, collations, etc., and do not attempt to offer synthetic treatment of complex subjects.

5. The CDLI Website

The cuneiform digital library is accessible through a website at the University of California at Los Angeles, which is mirrored at the Max Planck Institute for the History of Science in Berlin. Like other websites of institutions and projects, the website offers extensive information about the project, including the aims and activities of the initiative, the associated scientists who contribute to its work, the methods applied, reports about the outcome of meetings etc.

This information about the CDLI, however, accounts for only a small part of the total information offered by the CDLI website. The site's main function is to provide free access to the growing number of catalogue data, images and transliterations, which constitute the library, and to assure the longevity of this access

with stable links to the sources beyond the time perspective of current CDLI activities.

In its present state, the site already shows substantial complexity. Nevertheless, it is still under construction. On the one hand, the number of registered tablets as well as the amount and quality of the data representing them are continuously increased. On the other hand, the facilities of the working environment it represents are enriched and new levels and functions are implemented.

The core of the digital library access system is the display of data concerning an individual cuneiform tablet or other object with cuneiform text. The standard display offers essential catalogue data, composite images of the six sides of the tablets displayed to scale with regard to their real sizes, the transliterations of their cuneiform text content, and links to further information such as full catalogue entries, 600 dpi high-resolution images, or images of hand copies. The display can be toggled between two layout versions, one for optimal screen display, the other for printing.

Moreover, as a service to collections digitized by the project, the CDLI offers the opportunity to have their collection additionally displayed using templates, which preserve their corporate identity (see the entry page to the collections [12]). These collection-specific display environments also provide particular access facilities supporting the administration of the collections

Access to the display of individual cuneiform documents is offered by search facilities and by catalogue listings of individual collections. The preliminary search environment, which is presently implemented, allows searches in essential fields of the CDLI catalogue and for text searches in the transliterations. Search results can be displayed in different ways. They can be accessed by simple lists of links to the standard display, by lists containing essential catalogue data and images, or by browsing the standard display through the search results.

The search facilities will be continually expanded in the future so that other types of data such as extended sets of metadata, glossaries, paleographic differentiations etc. can be included into search requests.

At the beginning, the CDLI followed a simple strategy to achieve a robust display of the data. A PERL script was used to create a system of static HTML pages. This strategy made it possible to distribute the data not only through the Internet but also by means of storage media such as compact discs.

The increased volume of data hosted by the CDLI forced the project to give up this simple strategy in favour of a dynamic solution. At present, a Filemaker Pro web server provides the data for the standard display of the cuneiform tablets, essentially using three files with data, the CDLI catalogue file, another file containing the transliterations in a display format generated by a conversion program from the ATF transliterations, and a file which contains the URLs to related high resolution images for each cuneiform text.

The following list of presently implemented units gives an overview of the current state of the CDLI website:

1. the CDLI home site with static pages containing information about the project,
2. the standard display system of the digital library,
3. the search unit of the digital library (preliminary version),
4. specific display systems for individual collections (to be continually extended),
5. access pages to philological tools (preliminary version),
6. access pages to technical tools and services (to be continually extended),
7. access pages to the Journal CDLJ and the Bulletin CDLB,
8. educational pages (to be extended).

References

[1] The Hermitage Cuneiform Collection. http://cdli.ucla.edu/links/hermitage.html

[2] Cuneiform Digital Library Initiative. A joint project of the University of California at Los Angeles and the Max Planck Institute for the History of Science in Berlin. http://cdli.ucla.edu (mirror site: http://cdli.mpiwg-berlin.mpg.de)

[3] The Pennsylvania Sumerian Dictionary. http://psd.museum.upenn.edu

[4] R. K. Englund. The Ur III Collection of the CMAA. In *CDLJ*, 2002.001. http://cdli.ucla.edu/Pubs/CDLJ/2002/CDLJ2002_001.html

[5] DTD for CDLI XML. http://cdli.ucla.edu/methods/de/dtd.html

[6] The Archimedes Project – Realizing the Vision of an Open Digital Research Library for the Study of Long-term Developments in the History of Mechanics. http://archimedes.mpiwg-berlin.mpg.de

[7] ATF to XML converter http://archimedes.fas.harvard.edu/cgi-bin/cdli/atf2xml

[8] SGML Parser http://archimedes.mpiwg-berlin.mpg.de/cgi-bin/sgml/parse.2.cgi

[9] The Electronic Text Corpus of Sumerian Literature. http://www-etcsl.orient.ox.ac.uk

[10] N. Koslova. Ur III-Texte der St. Petersburger Eremitage. Wiesbaden, Harrasowitz, 2000.

[11] CDLI Publications. http://cdli.ucla.edu/pub.html

[12] CDLI Digital Library http://cdli.ucla.edu/digitlib.html

♣ The CDLI is funded by the Digital Libraries Initiative (NSF/NEH), United States, and the Max Planck Society, Germany.