Making Assyrian texts available to the public

Claus Povlsen[1,] Dorte Haltrup Hansen[2], Thomas Klitgaard Hertel[3]

[1, 2]Centre for Language Technology and [3]Institute of Cross-Cultural and Regional Studies
University of Copenhagen,
Njalsgade 140, DK-2300 Copenhagen
E-mail: [1]cpovlsen@hum.ku.dk, [2]dorteh@hum.ku.dk, [3]thertel@hum.ku.dk

Keywords: metadata, tei–header, ancient works.

1. Introduction

In recent years several initiatives have been launched aiming at establishing digital infrastructures for scholars in the humanities. DIGHUMLAB[1] is a Danish distributed research infrastructure that besides offering deposit of digital resources also facilitates research in terms of LT tools usage and browsing data by selection of annotated metadata (Henriksen et al., 2014). This paper concerns the considerations and actions taken in connection with upload and representation of an ancient dataset annotated with standard compliant metadata.

2. Description of resources

The Old Assyrian (OA) text corpus dates to c. 1950-1700 BCE and represents a subset of texts from ancient Mesopotamia written on clay tablets in the cuneiform script. At present the OA corpus consists of roughly 23,000 individual documents whereof app. 6,000 have been published. The texts have an average length of app. 100 words per text, the shortest less than 10 words and the longest app. 1,000 words.

The corpus is heavily biased in terms of chronology, geography and subject matter. Around 95% of the corpus dates to a window of some 30 years and stem from private archives owned by Assyrian long-distance merchants who lived in a foreign city in central Anatolia (ancient *Kanesh*, modern Kültepe), some 1,000 kilometers away from their mother city Assur, northern Iraq (Barjamovic, Hertel & Larsen, 2012).

The commercial background of the corpus penetrates the subject matter of the texts and the textual categories they represent: app. 40% of the material represents commercial letters exchanged between close associates and family members; another 40% consists of legal documents (e.g. economic contracts, debt-notes, and judicial records), while the rest represents a mix of private memoranda, notes, lists, international treaties, some incantations and a few pieces of literature (Hertel, 2013). Some clay tablets are preserved with envelopes that typically contain inscriptions as well as cylinder seal impressions (Fig. 1).

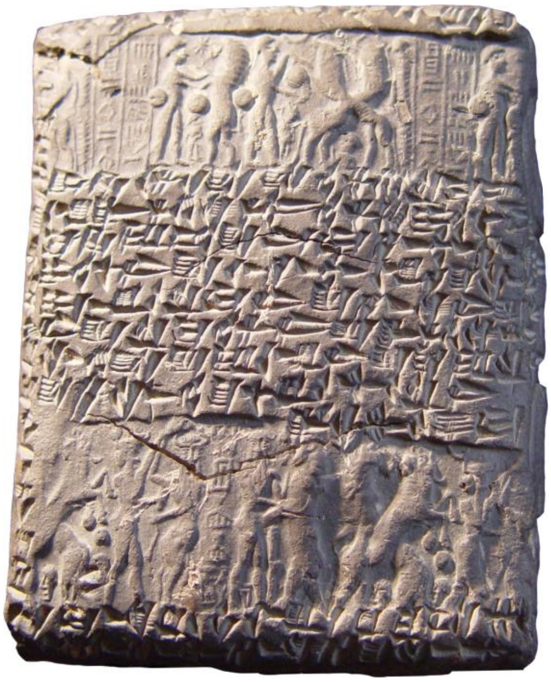


Figure 1: An Old Assyrian envelope with a clay tablet inside (AKT 6a, 197) showing cuneiform inscription and cylinder seal impressions (By courtesy of M. T. Larsen)

The current digital state of the corpus as a whole is in form of UTF-8 text files with transliterations and metadata regarding the physical format of the texts. For the purpose of the CLARIN-DK, a smaller controlled archive of c. 1,200 texts has been selected, which is currently published by Dr. Mogens Trolle Larsen, University of Copenhagen (Larsen, 2010; 2013). These files include specialist transliterations, English translations, digital photos of the physical clay tablets as well as text-specific metadata (ex. physical format, seal impressions, date, provenience, philological and contextual notes).

We envisage that the upload into the CLARIN-DK platform will have a significant impact on a variety of research communities—that it will not only aid as a research tool within the field of Assyriology but also encourage interdisciplinary research. For instance, (1)

[1] The DIGHUMLAB project constitutes the national consortium as the Danish part of the European infrastructure CLARIN ERIC.

scholars from the field of Assyriology, various branches of History (area studies, economic, legal etc.) and linguistics will be able to conduct both synchronic, diachronic and comparative studies drawing upon a controlled and public dataset; (2) studies in language technology will be able to include in their research and educational profiles structured data connecting signs in ancient original orthography, Assyrian language and English translation.

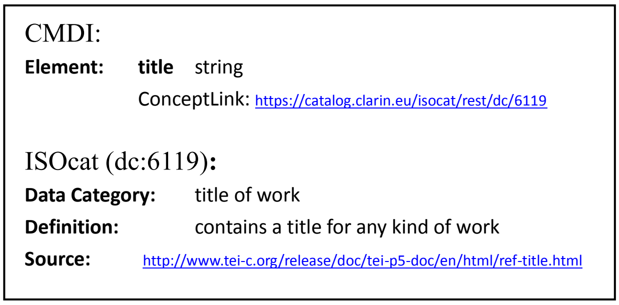
3. The TEI header concept

TEI P5[2] is a standard for representation of texts in digital form. It specifies syntax and semantics for metadata and for text in a very flexible way, allowing for a wealth of more or less fine-grained information. This flexibility gives a huge expressive power although at the same time making it difficult to reach agreement on one single common set of metadata.

To deposit written data in CLARIN-DK the data must comply with certain validation criteria; it must be embedded in xml, expressed in UTF-8, and encoded in TEI-P5 in accordance with a predefined header scheme[3]. The header scheme defines the valid structure and arity for a given subset of TEI-P5. The subset in question consists of 78 different TEI elements selected from the total amount of 705 different elements in TEI-P5. The TEI-header as the result of this first selection was then extended and implemented in CMDI[4] (the result of a co-operative effort with CLARIN Center Vienna and CLARIN-NL). The extension was based on the needs to cover description of single texts spanning from newspaper articles, contemporary literature to historical manuscripts. To give an example, the *msDesc* component (manuscript description) was added in order to account for historical text-bearing objects. The extended coverage resulted in an expansion from 78 to 88 different elements.

In the CMDI implementation process each TEI metadata element from the original TEI-header was preserved as well as the structure and the arity to the largest extent possible (see Hansen et al., 2014 for a further discussion). In this way the deposited documents encoded in TEI-P5 can be mapped automatically from TEI to CMDI in the repository.

In the TEI standard clear semantic is defined for each element. In CMDI, this information is represented in *conceptLink* linking to ISOcat where definitions from the TEI standard are stated as shown in Figure 2.



CMDI:

Element: **title** string
ConceptLink: https://catalog.clarin.eu/isocat/rest/dc/6119

ISOcat (dc:6119)**:**

Data Category: title of work
Definition: contains a title for any kind of work
Source: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-title.html

Figure 2: Semantics in CMDI expressed via ISOcat

The reason for CLARIN-DK to choose TEI-P5 and not a simpler metadata standard like Dublin Core or OLAC is that TEI is widely used among scholars and librarians to encode very detailed and specific information at both metadata and text internal level. The detailed information and the highly structured documents enable fine-grained search, textual processing and advanced visualization. The TEI-header is especially well-suited for describing historical texts. In the TEI introduction[5], it is thus stated that the guidelines: "… focus (though not exclusively) on the encoding of documents in the humanities and social sciences, and in particular on the representation of primary source materials for research and analysis".

4. Challenges and solutions

The main purpose of metadata annotation is to facilitate the users' search for relevant information and resources, often classified as resource discovery. In this context, it is considered important that the user in his search is able to distinguish between three layers or representations of the work in question, namely the collections of tablets, the transliterated version, and finally the translation into English.

Several benefits can be associated with use of an already implemented TEI-header schema as a starting point for encoding metadata information. To mention just one, it contributes to a quick overview of the gross amount of TEI-P5 information types. A resource is however not always compatible with the selected TEI-header schema. As mentioned above, the intended coverage of the TEI-header spans from modern to historical texts. Consequently, the challenge here is to represent the desired information for the Assyrian ancient text resources based on a TEI-header schema covering historical works from the mediaeval period. Researchers often wish to record very detailed historical facts about written ancient scripts. In TEI-P5 these facts are recorded in the TEI *history* element. Even though this history element is not included in the TEI schema used in CLARIN-DK, information about the history of ancient resources can easily be represented by filling in other elements implemented in the TEI-header. The minor

[2] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
[3] https://www.clarin.dk/schemas/tei/TEIDKCLARIN.rng
[4]http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1380106710826
[5] http://www.tei-c.org/Guidelines/

drawback is that the description will be expressed in more general terms. To give an example the *acquistion* [6] information is alternatively represented in the *physDesc* sub-element of *msDesc*.

The encoding of bibliographic information about the transliterated (and digitized) version of the Assyrian texts is stored in the bibliographic element of TEI-P5 (i.e. *biblStruc*) which is also part of the TEI-header coverage. The bibliographic information stored is amongst others the author and editor of the transliterated version of the Assyrian texts. Metadata information on the translations of the transliterated text will be allocated its own header. The co-reference between the transliterated version and the translated version regarding metadata information will be established by use of common file names.

5. Future work

In this paper we have described some key issues in connection with annotating Assyrian texts with metadata information in accordance with TEI-P5. In this way it is ensured that users both in their search efforts and in their search results can benefit from relevant and well-structured metadata. In addition, the standard compliance opens up for the possibility of e.g. exchanging the Assyrian with other data written in Akkadian.

The next step in making the data in agreement with user needs and requirements would be to implement the display of both the transliterations and the translations in English in the same window. Linguistic research seen from a contrastive angle would benefit from such an additional display option. Implementation of parallel displays implies that the bi-texts involved are annotated with some kind of alignment information. Even though the Assyrian language operates within a sentence concept, the sentences are not marked up with punctuation information, meaning that sentence alignment would require a lot of manual work. As a first approach towards parallel display of the bi-texts, the solution will thus confine itself to alignment of texts.

An obvious next step would be to add facsimile representations of the data as a third display as done in the Deutsche Text Archive project[7]. Such an initiative would, without any doubt, improve the chances of enhancing the group of end users. Here the already existing metadata on the physical format of tablets and distribution of words per line will be the key for creating the alignment between transliterations, translations and facsimile representations of the ancient texts.

Finally, the users have expressed the wish to have all tokens of the transliterated version of the data annotated with morpho-syntax in order to extend the search possibilities. The future plan is to exploit machine learning techniques in order to make the morpho-syntactic annotation less resource demanding.

6. References

Barjamovic, G., Hertel, T.K., Larsen, M.T. (2012). *Ups and Downs at Kanesh: Chronology, History and Society in the Old Assyrian Period* (Old Assyrian Archives, Studies, Vol. 5). Leiden: Nederlands Instituut voor het Nabije Oosten.

Hansen, D.H., Offersgaard, L., Olsen, S. (2014). Using TEI, CMDI and ISOcat in CLARIN-DK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik, Island.

Henriksen, L., Hansen, D.H., Maegaard, B., Pedersen, B.S., Povlsen, C. (2014). Encompassing a Spectrum of LT Users in the CLARIN-DK Infrastructure. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik, Island.

Hertel, T.K. (2013). *Old Assyrian Legal Practices: Law and Dispute in the Ancient Near East* (Old Assyrian Archives, Studies, Vol. 6). Leiden: Nederlands Instituut voor het Nabije Oosten.

Larsen, M.T. (2010). *The Archive of the Šalim-Aššur Family. Volume 1: The First Two Generations* (Kültepe Tabletleri VI-a). Ankara: Türk Tarih Kurumu.

Larsen, M.T. (2013). *The Archive of the Šalim-Aššur Family. Volume 2: Ennam-Aššur* (Kültepe Tabletleri VI-b). Ankara: Türk Tarih Kurumu.

[6] *acquisition* is a sub-element of *history* and “contains any descriptive or other information concerning the process by which a manuscript or manuscript part entered the holding institution”.

[7] http://www.deutschestextarchiv.de