4.5 Data Management Plan (DMP)

Data collection

The data that are used in this project comes in the form of standardized text in “Canonical ASCII Transliteration Format” (C-ATF).[1] The text itself is encoded in UTF8 and the original language transliterations are restricted to simple ASCII characters. This notation system has been in use for 15 years and because of its simplicity and high level of standardization, all research projects that use large quantities of cuneiform texts will base their work on the Cuneiform Digital Library Initiative (CDLI) database which hosts these texts, or will use a derivative of C-ATF. The ATF notation created by the CDLI is the widest-used standard in the field. In the case at hand, the project will use approximately 24,000 lines of text and their translation, which will be augmented and pre-processed for use as a training data set. Because the CDLI is a long-lasting initiative, there are already quality checks and versioning systems in place. Each time a change is saved in one of the texts, a backup copy of the previous version is saved in the database. There are also a number of tools in place which are used to verify the quality before commit, such as compliance to the C-ATF standard and to a list of preferred sign readings (each cuneiform sign can have more than one reading), depending on genre and time period.

Documentation and Metadata

As a companion endeavor to the Cuneiform Digital Library Initiative there exists a wiki[2] which documents all aspects of the CDLI in a range of articles on history, specific inscribed artifacts, and genres, as well as discussions of processes and data acquisition. On the CDLI website, there are also articles discussing the museum collections holding the physical documents[3] and also the terms of use of the data[4]. These tools will be used to help document the project and its outcomes. Individual texts in the CDLI have an Open Context ark number assigned. Moreover, we will collaborate with a French research group[5] for the alignment of the texts’ metadata with the CIDOC-CRM ontology. The linguistic and semantic information generated in the translation process and by information extraction will be linked with linguistic open linked vocabularies. The software created will be thoroughly commented and a github Jekyll website will serve as a documentation hub for each new software module.

Licensing

New software and derivative data generated by the project will be both released to the public domain by using the Creative Commons CC0 license “Public Domain Dedication” (CC01.0).[6]

Storage and Backup

During the research, GitHub will be used as a versioning system for the code base of the project. The sample text will also be joined to the code, exceptionally as all text of the CDLI is usually backed-up daily in SQL, text and disk image formats. This is in order to keep a controlled sample since the CDLI data change every day. The Center for Digital Humanities at the University of California in Los Angeles gives us technical support and external backups that increase the security and recoverability of the data in case of a problem. We also have a mirrors of the servers both through the Max Planck Institute for the History of Science, Berlin (MPIWG; and through them to the Max Planck Society’s persistent storage hub in Göttingen) and through the University of Oxford. Teams in Toronto and Frankfurt will each use a development server of which the relevant data will be periodically sent to the CDLI and the code on GitHub.

---

[1] <http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html>

[2] <http://cdli.ox.ac.uk/wiki/>

[3] Take for example the page of the British Museum <http://cdli.ucla.edu/collections/bm/bm.html>

[4]< http://cdli.ucla.edu/?q=terms-of-use>

[5] <http://triplestore.modyco.fr:8080/ModRef/>

[6] <https://creativecommons.org/publicdomain/zero/1.0/>

Preservation
By renewing periodically our agreements with the Center for Digital Humanities, the MPIWG-Berlin and the University of Oxford, we are convinced that the CDLI offers optimal storage security and web server longevity; CDLI is in fact a model of data persistence—the longest lived digital humanities project in the field of Assyriology, with its predecessor the Uruk Project at the Free University of Berlin now 26 years in existence. Since the new data produced answers to a need in the study and teaching of cuneiform cultures, its usage will only increase. For the eventuality of any risk to the preservation of the software or the data, we will put copies of our work in official repositories to maximize their preservation.

Data Sharing
The code and data produced by this project will be released in the public domain and we will encourage anyone to use, modify and reuse any of their components. Our code will be available on Github at all times, the data will be viewable and searchable on the CDLI website, it will also be accessible to download in full as an archive from the same website. There will be no intermediary between the user and the data, no account verification or login. With our strong communications plan, we expect that a large proportion of people who might be interested in our results will hear of us one way or another. Because the process of translation involves the internal tagging of the text, it is possible to leverage these text notations and export them into a variety of formats, and due to the high level of standardization of the data, it will be possible make them compatible with other projects like ORACC, but it will also be provided in an RDF edition compliant with Linked Open Data (LOD) principles. An important contribution of this LOD interface is facilitating interoperability from other philology portals for which LOD-compliant components are currently being devised,[7] as well as with museum collections in the LOD cloud[8], it also will help discoverability from search engines.

Responsibilities and Resources
Because the CDLI has been running for many years, we are fortunate most of the lab material is already in place. We are planning on updating the actual software and on building upon it to be able to host and serve the new data produced by the project, but the costs involved in these upgrading and maintenance tasks are comprised in the workload of those participating in the project. We are, for example, using a public Github code repository as opposed to paying for private repositories. The IT team of the MPIWG-Berlin, the Frankfurt Team, the Toronto Team and the Center for Digital Humanities (CDH) at UCLA will be responsible for maintenance and backup of their respective servers. Once the project is completed, the CDH and the Berlin mirror will maintain the CDLI server copies. Any translation generated by the project that attains our quality standards will be merged with the current CDLI data into their respective text entry in the database and thus available to view on the CDLI website at all times, and also downloadable in part or in total at all times. The translations and the various information extracted from the analyzed texts will also be available to consult on the web interface that will have developed to this effect.

[7] E.g., Homer Multitext <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/blackwell-smith/>, Perseus <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/almas-babeu-krohn/> and SAWS <http://www.ancientwisdoms.ac.uk/method/ontology/>

[8] E.g., the British Museum (<http://collection.britishmuseum.org/>