Journal of Open Humanities Data

Transliterated Cuneiform Tablets of the Electronic Babylonian Library Platform

COLLECTION:
REPRESENTING THE ANCIENT WORLD THROUGH DATA

DATA PAPER

YUNUS COBANOGLU
JUSSI LAASONEN
FABIAN SIMONJETZ
ILYA KHAIT
SOPHIE COHEN
ZSOMBOR FÖLDI
AINO HÄTINEN
ADRIAN HEINRICH
TONIO MITTO
GERALDINA ROZZI
LUIS SÁENZ
ENRIQUE JIMÉNEZ

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This work presents a corpus of transliterated cuneiform tablets from the *Electronic Babylonian Library (eBL)* platform, including a public API endpoint to download the latest version of the data, and a Python library to parse the transliterations in ATF format. As of the time of writing, the constantly growing dataset contains around 25,000 tablets with over 350,000 lines of transliterated text. This dataset is a sizeable addition to open-source cuneiform data and a major milestone for research within the fields of cuneiform studies and NLP.

CORRESPONDING AUTHOR:
Enrique Jiménez
Institute for Assyriology and Hittite Studies, Ludwig Maximilian University of Munich, Germany
enrique.jimenez@lmu.de

KEYWORDS:
cuneiform script; low-resource languages

TO CITE THIS ARTICLE:
Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., Hätinen, A., Heinrich, A., Mitto, T., Rozzi, G., Sáenz, L., & Jiménez, E. (2024). Transliterated Cuneiform Tablets of the Electronic Babylonian Library Platform. *Journal of Open Humanities Data*, 10: 19, pp. 1–7. DOI: https://doi.org/10.5334/johd.148

DOI: 10.5334/johd.148

The representation in Latin characters of signs of a tablet is called *transliteration*. A digital transliteration system that includes markup for all phonetic and graphic features, the *ASCII-Transliteration-Format* (ATF) (a name that is now anachronistic, since the format can use Unicode) was created by the CDLI in the late 1990s and early 2000s, and later modified by Tinney (2023) and named Oracc ATF.

CDLI, a database that contains records for more than 300,000 tablets, represents the largest publicly available digital collection of photographs and transliterations of cuneiform tablets. About 50 percent of the roughly half a million cuneiform tablets which have been excavated so far have not yet been transliterated or published (Streck, 2010).

The dataset in this paper was collected as part of the *Electronic Babylonian Literature* (eBL) project. The core of eBL is its online platform which provides easy access to an extensive collection of transliterations of cuneiform tablets along with tools that allow users to search the data and produce new transliterations of formerly unedited tablets. This way eBL seeks to offer a solution to the challenges posed by the fragmentary nature of the Mesopotamian written documentation. The eBL website and associated software projects are open-source and the individual records can be freely accessed through the browser (cf. Figure 1). The public API together with the Python code presented in this paper aims to make the entire dataset easy to access and process using our ATF parser.

2 METHOD

SOURCES

The catalogue of the eBL platform currently contains 262,717 records of cuneiform tablets, comprising the cuneiform collections of the British Museum, the Penn Museum, the Yale Babylonian Collection, the Hilprecht Collection, and the Vorderasiatisches Museum, among others. Of these, almost 25,000 are available in transliteration, and 52,105 in photographs; eBL is authorized to utilize and showcase the images with consent of the specific museum mentioned in each image. The initial list of museum numbers of the eBL platform was compiled using the catalogues of the CDLI, *The British Museum Digital Collections* (2023), *Yale Babylonian Collection* (2023) and other published and unpublished catalogues; the fields of the catalogue have been populated by the eBL team, who has also produced the transliterations. New tablets are constantly added and each document is subject to a careful revision process by the team before being entered into the database.

STEPS

Around 20,000 photos have been produced by photographers working for the eBL project. They cannot be reproduced without the explicit consent of the collections in which the objects are kept. The transliterations in the dataset have been produced by Assyriologists working at the eBL project, starting in 2018. In addition, transliterations have been entered by Assyriologists working at the projects *Edition of the Omen Series* Šumma *Alu* (Mittermayer, 2017–2021) and *Typology and potential of the excerpt tablets of* Šumma *alu* (Mittermayer, 2022–2023); *Introducing Assyrian Medicine: Healthcare Fit for a King* (Taylor, 2020–2023) *Reading the Library of Ashurbanipal: A Multi-sectional Analysis of Assyriology's Foundational Corpus* (Taylor & Jiménez, 2020–2023), and *Cuneiform Artefacts of Iraq in Context* (Jiménez, Sallaberger, & Radner, 2023–2046). Many of the over 25,000 transliterations have been produced solely on the basis of the photographs and have not been checked against the originals in museums. The transliterations are created using an online ATF editor that is part of the eBL platform. Once saved, the transliterations are parsed to a JSON tree using our ATF parser and saved in the database.

QUALITY CONTROL

A permission and revision system was implemented at the beginning of the project to maintain high quality of the data. Each transliteration is reviewed by another expert and changes are tracked, documenting the edit history of each document.

2.1 DATASET DESCRIPTION

2.2 DESCRIPTION

The dataset is a single JSON file which contains a list of objects (so-called "fragments", since most cuneiform tablets in the dataset are fragmentary). Each fragment contains an id (e.g. ND.5513) which can be used to find the fragment in the browser (see Figure 1), a short description, metadata such as the name of the collection, the museum and information on the publication history. There is additional information on the editors and the edit history of the transliteration, specified under "records", the genre, script type, pointers to external collections containing the item and many more properties. The transliteration of the fragment is saved as the "atf" property (as plain text, i.e. a *string*) which can be parsed into a JSON tree, as explained in detail below.

3 DOWNLOADING AND PROCESSING THE DATA

The *eBL fragments Python code* (see 1) can be used to download and parse all openly available transliterated documents using our public API. The eBL-ATF parser, which is an integral part of the eBL-API, has been made accessible as a standalone Python package in the *eBL fragments Python code*. Since eBL-ATF is a superset of standard ATF, the latter can be easily converted to eBL-ATF. For details on the parser and compatibility with Oracc ATF, the reader is referred to our documentation. The dataset at Zenodo contains all the fragments available on the 1st of September 2023. To get an up-to-date version, the *eBL fragments Python code* provided should be used.

OBJECT NAME

fragments.json

FORMAT NAMES AND VERSIONS

JSON.

CREATION DATES

2018-05-29 to 2023-08-31.

DATASET CREATORS

Sophie Cohen – Data curation

Zsombor Földi – Data curation

Ekaterine Gogokhia – Data curation

Aino Hätinen – Data curation

Adrian Heinrich – Data curation

Tonio Mitto – Data curation

Felix Müller – Data curation

Jeremiah Peterson – Data curation

Geraldina Rozzi – Data curation

Luis Sáenz – Data curation

Babette Schnitzlein – Data curation

Krisztián Simkó – Data curation

Henry Stadhouders – Data curation

Catherine Mittermayer – Data curation

Fabienne Huber Vuillet – Data curation

Kaira Boddy – Data curation

Jon Taylor – Data curation

Enrique Jiménez – Data curation, Project administration, Writing – review & editing, Funding acquisition

LANGUAGE

English.

LICENSE

eBL fragments Python code: MIT License

Data (fragments.json): Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Photographs: Reproduction of the images requires explicit consent from both the funding projects, the relevant institutions, as well as the institutions in which the cuneiform tablets are kept. Users are directed to review the conditions for image reproduction in the image captions.

REPOSITORY NAME

Zenodo, GitHub

PUBLICATION DATE

2023-08-31

4 REUSE POTENTIAL

For traditional philology the dataset is of enormous value, since it allows access to tens of thousands of cuneiform tablets previously unpublished. It has been estimated that all the tablets preserved in the world's museums as a whole contain about 10,000,000 words (Streck, 2010, 54–55): the dataset published here, compiled mostly from scratch, contains 350,000 lines that were previously inaccessible. This wealth of new data has already propelled the process of piecing back fragments for reconstructing fragments that were in a fragmentary state: alone in the compilation of the corpus 1,250 "joins" (i.e., fragments that belong together) have been detected. The dataset has also been used for easy identification of the content of fragments that would otherwise be difficult to identify.

NLP tasks for cuneiform scripts include, among others, generating automatic transliterations from signs to readings (Gordin et al., 2020), restoring damaged signs (Fetaya, Lifshitz, Aaron, & Gordin, 2020), matching fragments with their corresponding parts to reconstruct complete fragments, and machine translation from Akkadian to English (Gutherz, Gordin, Sáenz, Levy, & Berant, 2023). For an overview of different NLP tasks in Assyriology see Sahala (2021). The images can be used for semi-supervised or unsupervised OCR methods (Rusakov, Somel, Fink, & Müller, 2020). For recent advances in visual methods for cuneiform script see Bogacz and Mara (2022).

ACKNOWLEDGEMENTS

The photographs of tablets from The British Museum's Kuyunjik collection were produced in 2009–2013, as part of the ongoing "Ashurbanipal Library Project" (2002–present), thanks to funding provided by The Andrew Mellon Foundation. The photographs were produced by Marieka Arksey, Kristin A. Phelps, Sarah Readings, and Ana Tam, with the assistance of Alberto Giannese, Gina Konstantopoulos, Chiara Salvador, and Mathilde Touillon-Ricci. They are displayed on the eBL website courtesy of Dr. Jon Taylor, director of the "Ashurbanipal Library Project." The photographs of the The British Museum's Babylon collection are taken by Alberto Giannese and Ivor Kerslake (eBL Project, 2019–present). The photographs of the tablets in the Iraq Museum have been produced by Anmar A. Fadhil (University of Baghdad – eBL Project), and displayed by permission of the State Board of Antiquities and Heritage and The Iraq Museum. The photographs of the tablets in the Yale Babylonian Collection are produced by Klaus Wagensonner (Yale University) and used with the kind permission of Agnete W. Lassen (Associate Curator of the Yale Babylonian Collection, Yale Peabody Museum).

REFERENCES

Bogacz, B., & **Mara, H.** (2022). Digital assyriology—advances in visual cuneiform analysis. *Journal on Computing and Cultural Heritage*, *15*(2), 1–22. DOI: https://doi.org/10.1145/3491239

Cuneiform Digital Library Initiative. (2023). Retrieved from https://cdli.mpiwg-berlin.mpg.de/ (last accessed: 29 August 2023).

Fetaya, E., Lifshitz, Y., Aaron, E., & **Gordin, S.** (2020). Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(37), 22743–22751. DOI: https://doi.org/10.1073/pnas.2003794117

Gordin, S., Gutherz, G., Elazary, A., Romach, A., Jiménez, E., Berant, J., & **Cohen, Y.** (2020). Reading Akkadian Cuneiform Using Natural Language Processing. *PLOS ONE*, *15*(10), 1–16. DOI: https://doi.org/10.1371/journal.pone.0240511

Gutherz, G., Gordin, S., Sáenz, L., Levy, O., & **Berant, J.** (2023). Translating Akkadian to English with neural machine translation. *PNAS Nexus*, *2*(5), pgad096. DOI: https://doi.org/10.1093/pnasnexus/pgad096

Jiménez, E., Sallaberger, W., & **Radner, K.** (2023–2046). *Cuneiform Artefacts of Iraq in Context (CAIC)*. Retrieved from https://caic.badw.de/dashttps://caic.badw.de/das-projekt.html-projekt.html (last accessed: 7 February 2024).

Mittermayer, C. (2017–2021). *Edition de la série divinatoire Šumma Alu*. Retrieved from https://data.snf.ch/grants/grant/175970 (last accessed: 29 August 2023).

Mittermayer, C. (2022–2023). *Typologie et potentiel d'exploitation des tablettes d'extraits de Šumma alu*. Retrieved from https://data.snf.ch/grants/grant/205122 (last accessed: 29 August 2023).

Rusakov, E., Somel, T., Fink, G. A., & **Müller, G. G.** (2020). Towards Query-by-eXpression Retrieval of Cuneiform Signs. *2020 17th International Conference on Frontiers in Handwriting Recognition*, 43–48. DOI: https://doi.org/10.1109/ICFHR2020.2020.00019

Sachs, A. (1976). *The Latest Datable Cuneiform Tablets* (B. Eichler, Ed.), Kevelaer/Neukirchen-Vluyn: Butzon & Bercker, Neukirchener.

Sahala, A. (2021). *Contributions to Computational Assyriology* (Doctoral dissertation). DOI: https://doi.org/10.13140/RG.2.2.15525.17127

Streck, M. P. (2010). Großes Fach Altorientalistik. Der Umfang des keilschriftlichen Textkorpus. *Mitteilungen der Deutschen Orient-Gesellschaft*, *142*, 35–58.

Taylor, J. (2020–2023). *Reconstructing a 2,500 year old medical encyclopaedia.* Retrieved from http://oracc.museum.upenn.edu/asbp/ninmed/project/index.html and https://www.britishmuseum.org/research/projects/reconstructing-2500-year-old-medical-encyclopaedia (last accessed: 29 August 2023).

Taylor, J., & **Jiménez, E.** (2020–2023). *What was Ashurbanipal's Library?* Retrieved from http://oracc.museum.upenn.edu/asbp/rlasb/index.html and https://www.britishmuseum.org/research/projects/what-was-ashurbanipals-library (last accessed: 29 August 2023).

The British Museum Digital Collections. (2023). Retrieved from https://www.britishmuseum.org/collection (last accessed: 29 August 2023).

Tinney, S., Novotny, J., Robson, E., & **Veldhuis, N.** (2023). *Oracc: Open Richly Annotated Cuneiform Corpus*. Retrieved from http://oracc.museum.upenn.edu/ (last accessed: 29 August 2023).

Yale Babylonian Collection. (2023). Retrieved from https://collections.peabody.yale.edu/search/Search/Results?lookfor=bc+babylonian+collection&limit=5&sort=title (last accessed: 29 August 2023).

TO CITE THIS ARTICLE:
Cobanoglu, Y., Laasonen, J., Simonjetz, F., Khait, I., Cohen, S., Földi, Z., Hätinen, A., Heinrich, A., Mitto, T., Rozzi, G., Sáenz, L., & Jiménez, E. (2024). Transliterated Cuneiform Tablets of the Electronic Babylonian Library Platform. *Journal of Open Humanities Data*, 10: 19, pp. 1–7. DOI: https://doi.org/10.5334/johd.148

Submitted: 01 September 2023
Accepted: 17 November 2023
Published: 15 February 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.