Plan Overview

A Data Management Plan created using DMPTuuli

Title: The Origins of Emesal

Creator: Krister Lindén

Principal Investigator: Krister Lindén

Data Manager: Tommi Jauhiainen, Aleksi Sahala

Affiliation: University of Helsinki

Funder: The Research Council of Finland (former The Academy of Finland)

Template: Academy of Finland data management plan guidelines

ORCID iD: 0000-0003-2337-303X

Project abstract:

One of the great remaining mysteries in the study of the Sumerian language is the nature and origin of its only known variety, Emesal, which made a somewhat counter-intuitive appearance in texts only after the extinction of Sumerian as a spoken language around 2000 BCE. For almost 150 years this mystery has been intriguing Sumerologists. With computational methods, we aim to trace the most likely temporal and local origins of Emesal, and by significantly increasing our understanding of the history of Emesal, we will improve the understanding of how Sumerian ceased to be a vernacular.

ID: 16429

Start date: 01-09-2021

End date: 31-08-2025

Last modified: 10-06-2021

Copyright information:

The Origins of Emesal

1. General description of data

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Additionally, give a rough estimate of the size of the data produced/collected.

1. Research data description

1.1. Research data: previous, produced and managerial

The main data set for our project is the electronic corpora of the **Innsbruck Emesal Dictionary (Innsbrucker Sumerisches Lexikon, ISL) Project (EDP)** which will be made openly available in the **Open Richly Annotated Cuneiform Corpus (Oracc)**. Additional data in Oracc will be reused. The project data is text. It is estimated to amount to a few gigabytes.

The datasets will be available in several formats: **JSON**, **VRT** and **Conll-U**. JSON is the native format of Oracc. All our data will be available in the Oracc JSON format to make the data easy to import back into Oracc.

The VRT format is the vertical annotation format used in Korp (korp.csc.fi), which is a corpus search environment provided by the Language Bank of Finland hosted by the FIN-CLARIN research infrastructure. VRT is a flexible and extensible format used internally in the Language Bank of Finland as a convenient way to store data in textual format.

The data will also be converted to Conll-U for ease of continued annotation with language technological tools, which may mean that some annotations are dropped for ease of processing in this format.

1. **Previously collected** existing data which are reused in this project:
    1. Emesal Dictionary Project (EDP) data in FileMaker format
    2. Open Richly Annotated Cuneiform (Oracc) project data in JSON format
2. Data **produced as an outcome** of the process:
    1. EDP data will be converted to Conll-U, JSON and VRT
    2. Oacc JSON data will be converted to Conll-U and VRT
3. **Managerial** documents and project deliverables:
    1. All scientific publications or managerial documents will be available in txt and PDF format.

1.2 How will the consistency and quality of data be controlled?

1.2. Quality control

Converted data are provided with mechanical consistency checks for the number of fields and field labels. Such checks will be applied to manually annotated data as well.

Training data for machine learning is annotated manually, and automatically annotated data samples will be checked manually.

2. Ethical and legal compliance

2.1 What legal issues are related to your data management? (For example, GDPR and other legislation affecting data processing.)

2. Ethical and legal issues

2.1. Legal issues

All data subjects are long dead, so no personal data is handled in the project related to data subjects.

2.2 How will you manage the rights of the data you use, produce and share?

2.2. Data rights management

The project data in Oracc is openly available with a CC BY license. The EDP data will also be made openly available with a CC BY license in Oracc and in the Language Bank of Finland by an agreement with the University of Innsbruck.

In addition, a rights transfer agreement with the project participants will be concluded with the project participants funded by the project.

3. Documentation and metadata

3.1 How will you document your data in order to make the data findable, accessible, interoperable and re-usable for you and others? What kind of metadata standards, README files or other documentation will you use to help others to understand and use your data?

3. Data documentation and metadata

The data will be documented with metadata records available through META-SHARE at (metashare.csc.fi) which is connected to various international harvesting end-points, e.g. VLO in CLARIN, Europeana, etc.

The metashare.csc.fi server is hosted by the Language Bank of Finland and has links to content documentation openly available in the Language Bank Portal. The Language Bank of Finland also provides the data sets with persistent identifiers and requires that the data is provided with README files and other documentation for distribution.

4. Storage and backup during the research project

4.1 Where will your data be stored, and how will the data be backed up?

4. Storage and back-up during the project

4.1. Storage and backup

Snapshots of the original and the processed data will be stored in the FIN-CLARIN Language Bank already during the project for ease of sharing between the project partners. The data is securely backed up in the Language Bank on a daily basis.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

4.2. Access to data

All the project data will be modifiable by the project partners in the Language Bank of Finland GitHub repository during the project for version and access control.

5. Opening, publishing and archiving the data after the research project

5.1 What part of the data can be made openly available or published? Where and when will the data, or its metadata, be made available?

5. Post-project publishing and archiving

5.1. Data publication

All the project data will be made openly available and downloadable through the Language Bank of Finland download service (www.kielipankki.fi/download) and the metadata will be available through the Language Bank metadata service (metashare.csc.fi). In addition, the EDP data will also be made available in Oracc.

5.2 Where will data with long-term value be archived, and for how long?

5.2. Long-term archiving

Long-term storage of data will take place in the FIN-CLARIN Language Bank with back-ups in IDA and Fairdata-PAS.

6. Data management responsibilities and resources

6.1 Who (for example role, position, and institution) will be responsible for data management (i.e., the data steward)?

6. Data management responsibilities and resources

6.1. Data management responsibility

Both Tommi Jauhiainen and Aleksi Sahala are familiar with the services and the data formats of the Language Bank. As project researchers, they will take care of the data preparation for the Language Bank and Oracc as part of preparing the data for project use.

6.2 What resources will be required for your data management procedures to ensure that the data can be opened and preserved according to FAIR principles (Findable, Accessible, Interoperable, Re-usable)?

6.2. Data management resources

The permanent staff at the Language Bank of Finland will ingest readily prepared data and metadata into the FIN-CLARIN services as part of their normal activities free of charge.