TALAR – Research Data Policy

Thorsten Trippel and Claus Zinn

Last Update: 28.04.2023

Abstract

The document describes the research data policy of the **T**übingen **A**rchive for **LA**nguage **R**esources (TALAR). The repository is intended for data generated by research activities conducted by linguists from both the University of Tübingen and from other public research organisations.

Research Data The TALAR repository predominantly accepts datasets that have reached the end of the research data lifecycle and are considered worthy of archiving and publication. In exceptional cases, the repository may also accept research data that results from earlier phases of the lifecycle.

Contents The repository is intended to accommodate the specific needs of the linguistics community at the University of Tübingen, Germany, but it also accepts research data from other institutions. The TALAR repository hosts data stemming from the Collaborative Research Centre 833 (SFB-933). Its main content foci are *treebanks* and *word nets*.

Formats Generally, the TALAR repository is advocating the use of open, non-proprietary file formats as proprietary formats usually require a commercial product, and hence reduce the potential of research data reuse. Research data stored in proprietary formats, *e.g.*, EXCEL for table-based data, should be exported to the CSV (comma-separated values) data format before submission.

Similarly, data formats should be preferred that are in common use by the linguistics scientific community, and for which there a good availability of open-source tools for data access and processing.

For TALAR, we advocate the use of XML-based formats. We strongly suggest the use of XML schema files allowing users to validate their XML-based data representations. For treebank data, the repository accepts data using the TCF (https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format) or the CoNLL-U (https://universaldependencies.org/format.html) format.

The repository also accepts research data in the TEI format (https://tei-c.org). In this case, TEI-based representations must be complemented by their respective TEI schema definitions to check the validity of all data before ingest.

In principle, other data formats are allowed, but must find the approval of the TALAR archivist before ingestion.

Data Providers Acceptable data providers for this community include researchers affiliated with the University of Tübingen, and academic researchers from elsewhere, provided that their research data fits the content coverage of TALAR.

Metadata It is mandatory that all research data in TALAR is described by CMDI-based metadata (https://www.clarin.eu/content/component-metadata). The metadata must contain a minimal set of administrative metadata, similar to DataCite (https://datacite.org), and must contain a minimal set of resource-specific metadata. The TALAR archivists have access to a range of different CMDI-based schemas that data depositors can use to describe their research data accordingly.

Data Curation / Curation processes Prior to archiving and publishing datasets in TALAR, all research data will go through a data curation process to ensure that it meets the minimal requirements for archiving set-out in this document.

Data Volume A limited amount of 20 GB storage for research data is available per submission. Data sets larger than 20 GB must be divided into smaller data sets. In exceptional cases, the data limit for an individual data set can be increased in consultation with the archive manager.

Data Quality Data providers must agree to the data repository's terms of service and provide relevant, accurate and complete data and metadata. All research data must adhere to quality assurance criteria, which often can be defined in terms of the format in which the research data is expressed.

For example, all XML-based data (including TEI) must be well-formed and valid. Each TEI dialect must be complemented by a schema definition. Data should also meet other quality standards. Data that is being described in a scientific publication and data that stems from a completed, publicly-funded research project meet these requirements automatically. When required, the TALAR quality board, potentially aided by external reviewers, will decide whether the research data meets the minimal quality standards and can be archived.

Languages The repository accepts data records primarily in German and English, but metadata must be provided in English to ensure that the data is easily discoverable and accessible to researchers worldwide.

Reuse of data / Licence Depositors must provide a license for their uploaded data, allowing for reuse and/or redistribution of the data. The TALAR community requires the use of Creative Commons licenses (CC-BY 4.0), which is widely recognized and accepted in the research community. If other licenses are to be used, a prior negotiation with the TALAR archivist is required.

Publication processes Users can initiate the submission of a dataset via the Bagman software (https://weblicht.sfs.uni-tuebingen.de/bagman/). Bagman helps users to describe their data with a minimal set of descriptors, and to package their data into a BagIt File Packaging Format (https://datatracker.ietf.org/doc/html/rfc8493). Once researchers submit their package, the TALAR archive manager is contacted and will process the package. Any open questions regarding the research data package will be discussed between archivist and submitter via conventional communication means (*e.g.*, via email, phone, or a video conference). Once all data and metadata is quality-tested and agreed upon, and the appropriate licence has been determined, the researcher(s) need(s) to sign a data transfer contract (German: *Datenüberlassungsvertrag*). Then, the archivist ingests all research data into the TALAR repository. Research data, once published, cannot be altered. Metadata describing research data can be altered, for instance, to keep administrative data such as contact information up-to-date. When research data is being altered, a versioning mechanism is being used.

Each data set in TALAR obtains a persistent identifier using handle.net or doi.org.

Access Restrictions Data providers may request an embargo period for their deposited research data with a maximum duration of 1 year. During the embargo period, the research data is not accessible to the public. After the embargo period has ended, the entire data record becomes accessible to the public. Access to sensitive or confidential data must be restricted permanently to protect the privacy and confidentiality of research participants.