

VII. Data Management Plan

Expected Data

This project will produce three kinds of data. First is the results of lexicographical, term-centered research. As described in the proposal, this consists of references to primary sources illustrating the use of a term, translations of the term in other canonical languages and in modern languages, references to secondary sources that substantially discuss the meaning of the term, and explanations added to BTW by staff researchers and outside scholars. Second is the open-source software being developed to assist in such research. This includes the software code proper, its documentation, and the auxiliary structures (e.g. XML and database schemas) required for the software to operate properly. Third are the databases which the software will rely on to provide assistance to the researchers.

Period of Data Retention

There is no foreseeable specific date beyond which the data generated in this project will become obsolete, hence the data generated in this project will continue to be preserved into the foreseeable future. In addition to using BTW's own server (described above), we are looking into using a backup to ensure its preservation, and have discussed this possibility with one university partner. The White Paper produced at the end of this grant will be stored on each server. Dissemination of BTW's product is further assured through MRC's commitment to the 'three opens': using free, open-source software for BTW, following open standards for storage, transmission, and dissemination of BTW's content and metadata, and making content available under open licensing arrangements whenever possible.

Data Formats and Dissemination

The results of BTW's research will be made available for other scholars and for the public on a public web server whose permanent address still needs to be determined. Initially, the data will be available only for consultation. A search engine and an index will be provided for users to find the information they may be interested in. In a later phase of the project, a vouching process will be put in place so as to allow external scholars to add to the lexicographical information generated by our team, and to create their own lexicographical entries if they so desire. The general public will not be allowed to enter data in the system, on the presumption that they do not have the scholarly training necessary to ensure quality; however, we will experiment with allowing a public comment feature. The entries will also be exportable as files marked up with the Text Encoding Initiative (TEI) markup standard, facilitating incorporation of research results into other media. Entering data into BTW will entail granting the project a non-exclusive license for non-commercial use of the data entered, and material will generally be available under a Creative Commons License to the fullest extent possible. The software will automatically record with each piece of information who holds copyright. It will also keep a history of edits made to entries.

The software generated for this project will bear a GPL license. It will be made available for outside researchers on github.com. It will be accompanied by documentation in TEI format detailing how to install, operate and extend the software. The git version control system will manage controlled software releases and possible contributions from external developers.

The databases generated as part of BTW will be made available to external researchers as allowed by law or by agreements with other researchers. These will be made available as SQL dumps from our own databases, thereby allowing reasonably easy replication of the whole of BTW on other sites.

Privacy or confidentiality issues do not arise in this project. Someone wishing to keep his or her data private will simply not make it available. Otherwise, everything contributed is public.

A login system will ensure that a) data entered into the system is sourced to a specific scholar, and b) the data contributed by scholars cannot be tampered with.

Data Storage and Preservation of Access

The results of research will be stored on the same server, which will give access to these results, as described above. In addition, the data will be backed up off site daily to a storage space controlled by MRC. The software and databases will be preserved on github.com, as described above.