Data management plan

This data management plan was created on September 9, 2014, for submission to the Office of Digital Humanities (ODH), National Endowment for the Humanities as required by ODH Guidelines in the interest of securing funding for this project under the NEH/DFG Bilteral Digital Humanities Grant program. This is the first version of the data management plan associated with this data.

Roles and Responsibilities
US Project Director Hayim Lapin will be responsible for implementing this data management plan in consultation with Tal Ilan, the Germany Project Director. They will oversee regular collection and curation of data over the life of the grant period. At the end of the grant period, the Directors will deposit the data in the digital repositories of the University of Maryland and the Freie Universität Berlin. At the final data management meeting of the grant period, the Directors will finalize plans for the permanent deployment of a live version of the application developed.

Types of Data
This project will produce a data set comprising transcriptions of manuscript witnesses to classical Jewish texts along with contextual information describing these texts and their scholarly interpretation. These data will be created by trained transcribers employed by the project or by external scholars cooperating with the project. In addition, the project will generate three data tables in collecting (a) alignment data for Mishnah witnesses and (b) for Tosefta witnesses, and (c) synopsis data for the Mishnah and Tosefta. Along with the text and alignments, information about the provenance and condition of manuscript witnesses, previous editorial interventions will be captured during the course of this project.

The project will also generate or refine algorithms for matching strings in corpora, programs for aligning texts in two distinct works, and the software for managing the web application

Data and Metadata Formats
Project data will be stored, managed, and archived in a custom XML format compliant with the Text Encoding Initiative (TEI) P5 Guidelines, version 2.6, for encoding of electronic texts. TEI is the most-widely adopted standard for scholarly representation of texts in digital form. This format is platform-independent and open source and freely-available. TEI XML is suitable for long-term archiving and preservation in its current form; no transformations are required for preservation.

Standard bibliographic metadata about the digital files will be stored in the header of the TEI files along with information about the provenance of the manuscripts on which the transcriptions are based. Additional metadata that captures in detail important facts about the transmission of the text, such as damage or editorial intervention, will be represented in specialized tags intended for these purposes throughout the body of the text.

The project's customization of the standard is documented in a TEI ODD file. This enables generation of prose documentation for the project's tag set and schema files for technical validation. Derivative files in HTML format, which capture various presentational aspects of the data will be managed and archived along with the XML data. The data set will also include programs written in the eXtensible Stylesheet Language (XSLT) to allow all derivative files to be regenerated as needed.

Other types of data may be in the form of C#, javascript, XQuery, SQL, Java, or HTML.

Access and Dissemination
All data and software will be available through the project's Github repository, which allows other researchers to download and build upon the work of this project immediately. Presentational versions of the data will also be available through a website designed by the Maryland staff. All data from this project will be immediately shared under the terms of open licenses approved in consultation with the University of Maryland's Office of Technology Commercialization.

Data Storage and Backups During the Active Life of the Project
At least two types of copies of the data will be actively managed and stored during the life of the project. One copy will be stored on servers managed by GitHub. The subscription supporting the storage of this copy will be paid by the Project Directors. A second, local copy of the data will be stored on shared server space managed by the College of Arts and Humanities and service the Division of Information Technology at Maryland, which has a proven record of and commitment to secure data archiving for the University.

Long-Term Preservation
Within three years from the end of the grant period, a copy of the data will be permanently archived with the University Libraries at Maryland and at the Freie Universität Berlin. Copies of the data may also be deposited with other suitable repositories as identified by the Project Directors. Data will remain publicly available through the two principle repositories.