

6. Data Management Plan

Data description: BPS uses existing data in the form of corpora of text sources that are usually archived elsewhere, and generates prosopographical data from these corpora as a result of analysis and user curation. However, the data is produced by the actions of users of our system, rather than as a core activity of the grant activities. The software that supports the service is managed as essential data, although it has different characteristics.

Existing data: The existing data consists of the TEI of the demonstrator corpus of Uruk legal texts (HBTIN). Participants in the BPS implementation commit to transforming their prosopographical data (whether residing in connected text or in database format) to TEI. For this project, integration is not an issue of bringing data sets together but of transforming or annotating them to tool-kit input standards. The existing software will form the basis for ongoing development, and the new software will be managed directly as new versions.

Format: The data objects produced by BPS include assertions, expressed in XML, and prosopographical annotations that will be expressed in TEI. Any vocabulary/authority data that are produced will be exportable in a standard format, probably SKOS (an XML-based schema for simple ontology interchange). All of these are transparent formats, following interchange standards to the extent it is practical (i.e., to the extent standards exist). The formats can be easily compressed and so require very modest resources to archive and preserve. The software is stored as Java, HTML, CSS, PHP and related source code and associated resources, all readily accessible with appropriate tools.

Metadata: The TEI, SKOS, etc. namespaces declared in the XML provide the primary metadata, as these reference the standards that we follow. We will document the schemas used for the other information and/or any extensions to the TEI, as well as our usage of TEI. The software is well documented to allow re-use and facilitate maintenance.

Storage and backup: All input corpora and related input files (assertions and templates) are maintained on the file system of the cloud-based virtual machine (a.k.a., *VM* or *slice*) on which BPS is hosted. All analytic data will be stored in a database, running on the same VM, or an equivalent infrastructure. These cloud platforms are supported by established vendors in cloud computing, with professionals devoted to maintaining the automated backups (of storage and database) that are a standard part of the services. All software is in source code repositories local to developers' computers (backed up as part of UCB local infrastructure) and on central source code repositories. We are currently migrating from code.google.com to github.com, both of which provide secure infrastructure for open source projects.

Security: Our data stores no significant personally identifying information, no financial or health information, for any users. Users specify a login id that can be opaque, and only optionally specify a real name and website URL. This information is stored in the database and is accessible only by the users themselves at this point. All passwords are hashed in the database. The system currently requires new users to request an account to be added to the system so that we can filter unwanted automated abuse (e.g., "spam-bots" and "link-bots"). No security issues arise in the software, except that essential authentication information (e.g., database passwords) is not stored in the checked-in software, but is "injected" during the software build process.

Responsibility: Project Manager Pearce and Technical Lead Schmitz will be responsible for data management in the project.

Intellectual property rights: To the extent that results from the BPS tools are publishable (i.e., when workspace views are made public and shared), the workspace owner/s (the user/researcher) retain copyright over their own work. The assertion mechanism is designed to preserve the provenance chain of authorship, but each researcher chooses whether to publish their assertions and expose their workspaces, thereby retaining control over the fruits of their labor. If users request support for it, we can include copyright terms (e.g., a Creative Commons license) to be associated to content they publish. All software

is copyrighted by the Regents of the University of California, and available under an Apache 2.0 or ECL 2 license.

Access and sharing: The underlying data in the form of corpora are generally available in a public forum. Within the tool, any user can see any imported corpus and view the shared results of analysis on any corpus. BPS will make tools available to enable sharing, collaboration, and re-use of the data researchers generate. Individual researchers retain control over access to the analytic results in their workspaces, but there is a tradition of sharing and publication in these domains. Access to the software is freely available from the public repository.

Audience: The audience of secondary users of the data generated by BPS researchers/users will consist of researchers and students in the disciplines from which the data originates, researchers interested in exploring conceptual and methodological questions that draw on prosopographic data from a variety of disciplines. The potential audience for the software includes developers in the digital humanities domain, among others.

Selection and retention periods: BPS does not make any curatorial decisions about the corpora or resulting workspace data produced—such control remains in the hands of the users. All corpora and workspace results are retained in the system, and are projected to remain in the system unless and until users remove them. If the system is retired at some point, a reasonable attempt will be made to archive the published/frozen workspaces, so that they remain available in citation. All revisions to the software are archived in the public repositories.

Archiving and preservation: BPS does not currently include archival services for the data produced by our users. The export features to be added (described in the sections above) will enable easy and inexpensive archiving of generated results, as chosen and curated by the system users. The services-based architecture facilitates integration with archival services should that become a priority, but such integration is out of scope for the current phase. The files and database can be “dumped”, downloaded, and restored to another VM instance, should our agreement with the current hosting service be terminated. All revisions to the software are archived in the public repositories. These repositories support export of the software and revision history, should they discontinue their service.

Ethics and privacy: The website that presents the BPS tool-kit has a standard UC Berkeley privacy policy that is linked from every page (see: berkeleyprosopography.org/privacy). It notes that while information may be collected to run the services, personal information will not be disclosed without a user’s consent, except for “certain explicit circumstances in which disclosure is required by law.” As BPS does not gather sensitive data, this should not be an issue. No ethics or privacy issues arise with the software.

Budget: Documentation of the export formats that support individual users’ archival practices are included in the project costs as part of the development and documentation activity. Costs to host the platform with backups for the production service are included in the budget. The cost to archive the software on public repositories is zero—free hosting is provided for open source projects. The cost to back up local instances is included in the software development budget (as overhead).

Data organization: BPS organizes user data in corpora and workspaces, stored as separate entities in a database. Each has associated metadata about the user who “owns” or manages the entity, whether it is public, etc. The software is organized following current best practices for software development, with revisions (and authorship, etc.) tracked by the software repository.

Quality Assurance: End users are responsible for their own data, and any quality in this sense is outside the scope of BPS. Software quality is assured through user testing, as well as code reviews as part of the development process. The current grant includes resources to develop more automated testing to further ensure software quality.

Legal requirements: We are not aware of any applicable federal or funder requirements related to the data or software.