IX. Data Management Plan

Indiana University's Cyberinfrastructure

Indiana University is a national leader in the deployment and use of advanced information technology and cyberinfrastructure in support of research and education. IU's Pervasive Technology Institute is the umbrella for six major centers that provide systems, tools, and services furthering IU's research mission. That includes the University Integrated Technologies Services (UITS)'s Research Technologies, which run IU's computing and storage resources and help build scholarly communities. All online development work we do runs on Indiana University Webserve and benefits from significant cyberinfrastructure: computing, data storage, archive and restoration, database, and network access and control resources. Major components of IU's cyberinfrastructure include the following:

Supercomputers: IU's supercomputers provide IU researchers with access to some of the most powerful supercomputers in the US.

Scholarly Data Archive: IU's massive data storage system can hold up to 42 petabytes of data. With mirrored tape silos in Indianapolis and Bloomington, this very secure storage system ensures data are stored securely and reliably. Backup and replication is done within SDA's High Performance Storage System (HPSS). Data are written to a fast, front-end disk cache and migrated over time to IBM TS3500 tape libraries on the two main campuses, providing highly reliable disaster protection. The SDA is backed up nightly. All data are retained permanently.

IUScholarWorks: This is an open access repository service provided by the IU Libraries for disseminating and preserving the intellectual output of IU scholars. The repository is designed to hold and deliver scholarly materials in digital form (text, data, image, etc.) that will not change over time and that are described with standard keywords and descriptors. IUScholarWorks makes scholarly research materials freely available at a stable URL, and maintains them over the long term.

IU Web Framework: IU provides at no cost a set of standard tools to build feature-rich web sites where contributors can create, maintain, and publish web content.

Box @ IU: The IU Box is a no-cost cloud storage and collaboration environment that provides a secure way to share and store non-critical files and folders online. Account quota is unlimited.

Data generated by our research

Envisioned as a scholarly community of editors and inquirers, the Peirce Project's production platform (STEP) will produce several categories of data.

(1) The software and other artifacts that drive STEP and its companion suite of applications STEP TOOLS.

(2) Real scholarly data, including digitized manuscript images, raw transcriptions, encoded transcriptions, rounds of proofreadings and corrections, rounds of critical editing and corrections, rounds of apparatus documents and corrections, rounds of layout documents and corrections, thousands of annotation documents, bibliographical data, associated databases, technical reports, associated non-confidential professional correspondence, all connected to the development, testing, incremental implementation, and refinement of the platform for later public release.

(3) Online project website, online video tutorials; digital publishing tools (including online display format); custom software code along with archiving of open source software tools integrated into the project to maintain compatibility; abstract custom digitized workflow algorithms; digital tools for generating scholarly content, documentation and end-user education material; relational databases.

(4) Scholarly contributions to the critical edition, peer reviews of contributed scholarly content, and high-quality large resolution digital image files of original material.

Data management plan

Open source, open access, and limitations

STEP development uses open source software released under various free licenses, along with the creation of new software. This allows for most data listed above under (1) to (3) to be released under an open

license approved by the Open Source Initiative and made available for download, modification, and use by any party, except where governed by other open source copyright licenses. Data listed under (4) will be open access except when subject to university press contracts, copyrights, or when release could result in an invasion of personal privacy.

Data retention

STEP and STEP TOOLS will be released under an open license compatible with CDDL license once stable release versions have been reached that meet our benchmarks. Such stable versions should (a) successfully process the creation of potential project data, (b) enable customized workflow, variable editing guidelines, and modifiable access-control lists, (c) facilitate the creation of structured data, (d) incorporate data backup and restoration solutions, and (e) provide a method for unique website theming. Produced content used as both pilot data and scholarly work will be made available after an initial STEP release once the content will have been approved for dissemination.

Using components of IU's cyberinfrastructure stated above and below, all data produced during STEP and STEP TOOLS development and pilot testing will be both accessible *as soon as produced* and protected from disaster, assuring present and future access. Redundant access configurations will afford constant access with emergency outages being covered by enterprise-level restoration processes, systems, and personnel. All data are automatically saved to UITS's cloud storage server as well as to a local server. Data include database files (SQL), XML files, image files (jpeg, gif, tiff), document files (MS Word, PDFs, html), TEI files, STEP-generated files, and zip files. Files generated by STEP users are stored wherever wanted, on IU servers or elsewhere.

Data formats and dissemination

The software data category will contain a mixture of file types, all with third-party open-source products available for their viewing and editing. Scholarly work data, aside from being both generated and rendered from within STEP and STEP TOOLS, will consist of digital image files and structured data marked up according to platform and/or edition requirements, including TEI-XML mark-up compliance when that applies (one virtue of that mark-up is to ensure long-term interoperability). All STEP-related code will be posted on GitLab.

Pilot scholarly work and other subsequent STEP users will be afforded the customization options necessary for control access to produced data, assuring any legal or private data is available only to those with permission while guaranteeing public access to available humanities collections.

Toward a cloud-based operation for data storage, access, and preservation

The Drupal framework and STEP components are currently installed on IU's Linux virtual machine (VM) server in IU's Intelligent Infrastructure, where a standalone server hosts all services (web server, database, and git); we also use one of IU's cloud storages (box.iu.edu) for extra storage and backup.

We plan to keep the current standalone web service in the IU VM server coupled with a separate cloud server to manage increased storage. We will thus be able to support multiple projects in the current system, projects that can share and integrate the data in the cloud anywhere around the world. No local data are stored in VM itself: all the data are stored in the cloud storage system *(see diagram in Appendix 4).*

By integrating the current system with an affordable cloud service (e.g. box.iu.edu, or Google Drive), we will ready ourselves to move later on to a globalized cloud solution (cloud + extra support for the standalone web server). Primary service will then come from a cloud interface similar to Google docs, Google sheets, and Google presentation, while global service will support multiple localized cloud services based on distinct languages or the various interfaces of preferred services. STEP will consist by then of a standalone TEI editor, and of a virtual STEP Platform management (workflows) and TEI online editor system. This will allow any number of editorial projects to use STEP. Saving all data to the cloud provides easy access globally, along with unlimited storage space, robust security, considerable scalability, and especially ease of maintenance and thus appreciable savings: only one person will be needed to maintain the whole structure instead of multiple maintenance staff in different places.