

## 8. Data Management Plan (DMP)

Data description. We plan to generate, maintain, and distribute three types of information: (1) software, (2) research data, and (3) workshop materials. The software will consist of source code, configuration files, documentation for the application programming interface (API), and an installation and user guide. The research data will include 41 gazetteers in XML format from the Institute of History and Philology at Academia Sinica and structured datasets extracted from 2,000 gazetteers and collections of letters and notebooks. The workshop materials will be comprised of presentation files and in-depth tutorials.

Period of data retention. All source code will be accessible online through Github during system development. After the project has been completed, the source code will remain on GitHub. The workshop materials will become available on the project website after their conclusion, and some of the gazetteers and all of the extracted datasets will also become open-access after the final cleanup process. All of this data will be permanently preserved, maintained, and open to public on server provided by KCL. Extracted data will be made permanently available through its respective projects.

Data formats and dissemination. The source code, configuration files, and documents will be stored in plain text and HTML format on Github. We will also provide compressed copies on our project website for download.

Because the system is entirely web-based, we will not require users to download compiled, executable binary, unlike web servers such as Apache, or databases like Postgresql. Required third party libraries (free of intellectual property issues) will be made available on GitHub, along with all of the necessary download information. Through our arrangement with the Institute of History and Philology at Academia Sinica, the tagged and extracted data from 41 early gazetteers, together with their original texts, will be made available as XML files in order to demonstrate the viability of the platform.

The extracted biographical data from the larger corpus of gazetteers, for which we do not have distribution rights for the original texts, will be available on the China Biographical Database. Similarly, the extracted geographical information will be distributed through China Historical GIS and data from letters and notebooks through the Communication and Empire project. All data derived from gazetteers will also be stored as CSV (Comma Separated Values) files encoded in UTF-8 with metadata which details the primary sources; these files will be distributed through the World Historical Dataverse (<http://www.dataverse.pitt.edu/>) and the Dataverse Network (DVN) (<http://thedata.org>). The DVN will perform archival format migration, metadata extraction, and validity checks as well as enable on-line analysis, variable-level searching, data extraction and re-formatting, and other enhanced access capabilities.

The workshop materials will be disseminated as PDF documents, made available from both the project website and DVN to ensure long-term accessibility. All system files, research materials (excluding the original 2000 gazetteers) and workshop materials will be compressed in zip format, also downloadable from the project website.

System security. The DVN complies with Harvard University requirements for good computer use practices. The University has developed extensive technical and administrative procedures to ensure consistent and systematic information security. "Good practice" requirements include system security requirements (e.g., idle session timeouts, disabling of

generic accounts, inhibiting password guessing), operational requirements (e.g., breach reporting, patching, password complexity, logging), and regular auditing and reviews. The full University security policy can be found at <http://security.harvard.edu/>.

The project will also be supported by DDH in KCL. DDH provides 3 - 4 virtual servers (production, staging, and development) which share up to 1TB of redundant online storage. All DDH projects are hosted on industry standard VMware Vsphere server virtualization technology, which offers data redundancy, backup, and disaster recovery. DDH servers are located at the University of London Computing Centre, which is directly connected to the JANET backbone offering 99.999% availability. DDH's server infrastructure was completely replaced in September 2011 and a disaster recovery facility and duplicate filestore are now in place.

Budget. The cost of preparing data and documentation will be borne by the project, and is already reflected in the personnel costs included in the current budget. The incremental cost of permanent archiving activities of DVN will be borne by Harvard's Dataverse Network.

Privacy, intellectual property, other legal requirements. The system will be distributed as source code under the terms of the Apache License 2.0. Through our arrangement with the Institute of History and Philology at Academia Sinica, the 41 early gazetteers together with the original texts will be made available in the online system for academic use. The 2000 original gazetteer texts with intellectual property issues (including copyright, database right, license restrictions, trade secret, patent or trademark) will not be distributed to any party (including investigators, institutions, and data providers).

The data extracted from gazetteers are not subject to privacy and intellectual property concerns. All materials from system documents, extracted data, and workshop materials will be open to the public and distributed under the terms of the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported (CC BY-NC-SA 3.0) license.