

8. Data Management Plan (DMP)

As a member of the Computation Institute, the ARTFL Project has access to several high performance computing systems to support this project: PADS is a petabyte-scale high-performance online storage server capable of sustained multi-gigabyte per second input/output performance, tightly integrated with a data analysis cluster capable of 9 teraflop/s computing. PADS enables the reliable storage of, access to, and analysis of massive datasets by both local users and the national scientific community. BEAGLE is a Cray XE6 system a total of 17,856 compute cores on 744 nodes, each connected to 32 GB of memory, for a total of 23 TB. Teraport, a cluster that provides a grid-enabled analysis platform to the University of Chicago research community. Teraport is a grid computing research testbed for the Open Science Grid (OSG) and its member organizations. It is used to demonstrate and implement multi-grid interoperability by acting as a portal to the TeraGrid and other grid fabrics. The cluster comprises 128 dual 64-bit AMD Opteron processors with 12 TB of storage.

All data, tools and documentation generated in this project will be deposited and stored on the Bodleian Libraries' Digital Asset Management System (DAMS) at the University of Oxford. Object management capability and the objects stored will therefore become subject to the Bodleian Libraries' general digital preservation processes. The DAMS aims to provide a sustainable long-term repository for digital data, which can facilitate the necessary digital preservation processes. The University and the Libraries are committed to preserving and maintaining access to the digital collections in their care.

Once loaded onto the DAMS, the DataBank system stores the data and metadata unaltered as a canonical source of metadata, however it also transforms the metadata into RDF triples in order to allow for added discoverability and interoperability of the records. Furthermore, "context objects" have been generated to represent entities which can have significant identities in their own right (authors, significant figures, places, dates/events) so that they are amenable to annotation and the addition of further metadata. These context objects are essential for the re-use of this data over time. Object relationships are expressed using RDF. The object model accommodates multiple metadata streams for each object and as a result, items can in the future be readily augmented with comments, additional data, attachments and external links without requiring architectural changes to the storage system. Layered over the object store are a set of tools and services which provide full text indexing and faceted search (Apache-SOLR), XML query capability (eXist), an RDF triple-store (Mulgara) along with administrative tools such as virus scanning, text extraction and job scheduling. Data sharing protocols such as OAI-PMH, PAI-ORE, Atom and RSS are also catered for.

A Drupal website for the project will be developed within the existing CMS environment at the University of Chicago. This will be maintained for at least five years after the end of funding by the ARTFL Project and Computation Institute. Web-based information about the preserved digital outputs of the project will be guaranteed for at least ten years by the Computation Institute/ARTFL Project.

New primary and secondary data and metadata from the project will become a part of the corpus of preserved digital collections in the Bodleian Libraries and the corresponding repository at the University of Chicago. It is not intended that this include large-scale licensed or harvested data from existing established repositories used in this project. However the published research findings from the project will be linked to the data to which they relate, and the methodologies and analytical tools made publicly available. How access to the relevant data will be organized will be negotiated with content providers: either through APIs, or through the incorporating the analytical tools developed into the original data resource, or through the preservation of derived datasets that the project will generate.

The architecture of the Bodleian's DAMS ensures that the data and the interface remain structurally separate. This means that users can have access to the data through multiple means (through APIs, through aggregation websites) beyond the initial project website. The benefit of this approach is that access to the data can be guaranteed beyond the life of the interface.

Software created as part of the project will, wherever possible, be contributed to repositories appropriate to the discipline, in order to ensure maximum reuse. However, the digital humanities support at the institutions will also expect to provide ongoing training and support in their use.