Making Data Count: Developing a Data Metrics Pilot

A Data management plan created using the DMPTool

Creator(s): Carly Strasser, Patricia Cruse, Stephen Abrams

Affiliation: University of California, Office of the President

Last modified: September 15, 2014

Making Data Count: Developing a Data Metrics Pilot

Types of data produced

Five kinds of data will be produced or exploited in the course of this project.
The first is social science data in the form of quantitative and qualitative survey results collected as part of Unit One. Given that these surveys will involve human subjects, we will ensure our research is compliant with Institutional Review Board policies and receive prior approval before conducting the surveys. These data will be collected via paper and electronic means. Paper survey results will be digitized by hand, and originals will be kept for the duration of the project.
The second is dataset usage metrics harvested from the DataONE network through mechanisms developed in Unit Two. Individual DataONE member nodes collect raw web log usage statistics. Since this data potentially could be used to identify individual data consumers it is anonymized before being aggregated by the DataIONE coordinating nodes that will be harvested by the DLM tool.
The third consists of software produced or modified as part of the project. The new DLM software created by PLOS will be maintained in a community Github repository during and after the project. Extensions to DataONE code will be maintained in the public DataONE Subversion repository. Fenner will be responsible for code related to the DLM tools.
The fourth is dataset citation data harvested by the DLM tool from traditional and alternative (e.g., blogs, Twitter, etc.) channels of scholarly communication as part of Unit Three activities.
The fifth consists of all other research products (e.g., additional end-user tools developed in Unit Four, analysis results from Unit Five, community feedback, and project status communications) that will be made publicly available via either peer-reviewed journal articles or the project's website, which will be maintained by California by Strasser.

Data and metadata standards

The DLM software produced by the project will conform to accepted community best practices, including version control (using git) with tagging of major releases, a permissive open-source license (Apache 2), public availability in a community repository (Github), inline comments, reference and tutorial documentation with download and installation instructions that is available from within the software and from a community website, and extensive test coverage.
As part of this project the DataONE development team will evaluate the utility and level of effort necessary to report data usage statistics in a form that is consistent with COUNTER standards. This would permit meaningful comparisons between metrics for data usage and other types of online resources, including traditional serial and monographic publications.

Policies for access and sharing

All five major data categories, as enumerated in Section 1, will be publicly available for review, evaluation, and use as they are generated during the project and after its completion. Announcements about software and data

availability will be made using a variety of channels (e.g., blogs, Twitter, email lists) targeting all interested stakeholder communities.
As part of this project the DataONE development team will evaluate the utility and level of effort necessary to report data usage statistics in a form that is consistent with COUNTER standards. This would permit meaningful comparisons between metrics for data usage and other types of online resources, including traditional serial and monographic publications.

Policies for re-use, redistribution

Similarly, all software products resulting from this project will be re-usable and redistributable both during the project and after its completion. The only restriction placed on redistribution of the software is that the copyright and license statement be kept intact as required by the Apache open source license. This software is expected to be of interest to publishers, data centers and repositories, individual researchers, and institutional administrators.
As part of this project the DataONE development team will evaluate the utility and level of effort necessary to report data usage statistics in a form that is consistent with COUNTER standards. This would permit meaningful comparisons between metrics for data usage and other types of online resources, including traditional serial and monographic publications.

Plans for archiving and preservation

In addition to managed in the community based GitHub repository, all major versions of the DLM software will be archived in CDL's Merritt repository with persistent identifiers supplied by the EZID service.
Research data and records will be maintained for as long as they are of continuing value to the researchers and project collaborators.
The Merritt Repository Service from the University of California Curation Center (UC3) has capabilities to manage, archive, and share digital content, and provides persistent URLs, search interfaces, and tools for long-term data management. Merritt relies on a highly fault tolerant micro-services architecture with significant redundancy of all computational and storage components. Currently managing over 15 TB of digital resources, Merritt has not experienced any data loss over its five years of production operation.
The NSF-funded DataONE project includes a significant strand of activity that is investigating a number of options for ensuring ongoing sustainability and ongoing organizational, financial, and technical continuity. The software underlying the complete DataONE infrastructure is managed in a public Subversion code repository and is supported by a large and diverse developer community. The DataONE infrastructure relies on a highly distributed fault tolerant architecture. Although all of the individual nodes on the DataONE network have undergone many periodic system upgrades and software refresh cycles, the global architecture has worked properly in all cases, automatically failing over to redundant system capacity on other servers. There has been no interruption of end-user accessibility to the DataONE network since it first went into production operation.