

## Data Management Plan

**Project Title:** Topic Modeling for the Humanities    **Institution:** University of Maryland

**Project Director:** Jennifer Guiliano

**Budget:** \$24,807

**Beginning:** 05/01/2012

**Ending:** 04/30/2013

**Duration:** 12 months

This data management plan was created on September 12, 2011, for submission to the Office of Digital Humanities (ODH), National Endowment for the Humanities as required by ODH Guidelines in the interest of securing funding for this project. This is the first version of the data management plan associated with this data.

**Types of Data:** The data produced by this project will consist of presentations to be given by participants in the proposed workshop. These data will comprise slide presentations, audio recordings, video recordings, and text files documenting an interdisciplinary workshop on the application of topic modelling approaches to humanities data. Social media will be a component of participation in the workshops and postings on social networks such as Twitter will be collected. The researchers will also create a web site to publicize the event and provide access to materials (see "Access & Sharing" below).

**Data Standards and Capture:** Data for this project will be captured as part of the process of running the proposed workshop. Invited speakers will be asked to upload slides, documents and other supporting materials to computers provided by Maryland (as the host institution). Maryland will also perform audio and video recording of the workshop events. Posts from social media services will be collected using application programming interfaces (APIs) provided by these services and open-source tools which comprise PHP code and MySQL database software. For presentations given at the workshop, widely-adopted formats, such as Microsoft Powerpoint, will be used. Recent versions of these formats are at relatively open but, to the degree they are not, these formats are widely-supported across multiple technical environments and are likely to remain readable. Some conversion may be performed to normalize presentation files to PDF format as necessary for long-term accessibility. Audio and video of workshop events will be captured as uncompressed audio or video but will be converted and maintained in the form of widely-adopted open formats such as MP4 that are suitable for storing efficiently and sharing over the network. Social media content will be collected using published application programming interfaces maintained by these services. Plain-text, open-formats such as XML or JSON (JavaScript Object Notation) are the most common formats for such services. The project will rely on consistent naming conventions created in accordance with local policies at Maryland. Furthermore, as part of deposit with digital collections hosted by the University of Maryland Libraries, all items will be given unique identifiers according to the handle system, an open protocol published by the Internet Engineering Task Force. This system will support persistent citation of data created by this project.

**Metadata:** Descriptive metadata will be created as part of the process of depositing items with the Maryland Libraries. Metadata elements will be manually supplied using the Dublin Core Metadata scheme by project personnel collaborating with library staff to ensure consistent application of content conventions and controlled values. Embedded technical metadata will be maintained with all audio and video recordings. Dublin Core is the chosen metadata standard for this project because it can be created using existing workflows for depositing research products into Maryland's institutional repository and also because Dublin Core metadata can be easily reused as embedded data in web pages related to the workshop. This alignment with wider web practices will help enhance the discoverability of materials produced by this project.

**Legal Policy:** Participants in the workshop will retain their copyright and other intellectual property rights in material submitted for the workshop but will agree to license materials prepared for the conference under permissive open source licenses (such as Creative Commons) according to Maryland's standard policy or a similar arrangement agreed between participants and the workshop organizer. Contributions to the final white paper will be considered under a similar agreement. Participants will be asked to sign waivers giving their consent for audio and video of presentations to be shared. There are no other legal policy issues associated with data created or captured for this project.

**Data Storage, Security, and Backup:** Once collected from invited speakers or from digital audio and video recording devices, data will be physically stored on a password-protected server maintained by the Information Technology Division of the Maryland University Libraries. Servers are housed in a secure machine room with redundant power. No data will reside on any other portable or external media. Data is backed up incrementally through a service provided by the Maryland Office of Information Technology, which has a proven record of and commitment to secure data archiving for the university. In addition to backups hosted on university servers, data will be copied to tape and stored at a geographically-distant site through a service provided by Iron Mountain information services. The specific volume of storage for this project is not anticipated to exceed 1 TB.

**Access, Sharing & Re-use:** All data from this project will be made available for download from either the dedicated site hosted by MITH or from the Digital Collections of the University of Maryland within 6 months of the end of the workshop. There will be no additional permissions required to download or reuse data except for those specified above. Data from this interdisciplinary workshop will be useful to researchers in the fields of machine learning and related disciplines in information science and computer science. Also, data may be used by humanities researchers who are interested in applying computational methods to detect patterns in large collections of cultural heritage material.

**Long-Term Preservation:** Within three years from the end of the grant period, data will be permanently archived with the University Libraries at Maryland. No data will remain on servers controlled by the Maryland Institute for Technology in the Humanities (MITH). Data will remain publicly available through the libraries' digital collections. The University of Maryland has an interest in investing in the management of data created by researchers affiliated with Maryland and to providing digital preservation services, such as file validation, integrity checks, and, if needed, format conversion.