

Data Management Plan

The data produced by this project will be mostly text and will be made publicly available under Creative Commons Attribution 3.0 Unported.²⁶ There will be open source code, documentation for the algorithms used, and the resulting dataset, which will include raw output (e.g., list of musicians and their dates), relationships among the musicians and events (stored as serialized RDF/XML). Ontologies standard to this research field commonly applied to representing RDF/Named graph data, such as FOAF²⁷ and Event Ontology,²⁸ will be used to standardize the vocabularies. If the original public-domain sources are not already OCR'd, we will OCR them and store the output as HTML or XML. (Non-public domain sources will not be OCR'd.) Conference reports (PDF) and presentations (PPT) will also be made readily available publicly. An estimated overall storage requirements for the database and supporting document will be no more than 200MB.

All of the data and documentations will be made public on our project website to be hosted at McGill University maintained by the co-director (Fujinaga). Final project data and documentation will also be publically accessible online through the JHU Data Archive, maintained by JHU Data Management Services. Most of this data is already available and maintained on github.com.²⁹ The new data and modification to the GATE³⁰ information extraction software generated at JHU will be backed up daily to a copy at JHU (on an external drive) and another in a file server at McGill University. The data files generated are in the commonly used and open-source RDF format, which is designed with standards for interoperability and accommodating future format migration. All linked data (RDF) will be maintained and accessible for public search using the SPARQL³¹ query language, which allows for flexible search of the historical data. Also retained are the public-domain sets of text as linked RDF sources (provenance) and all analysis code. We expect the data to be fully accessible by near the end of the grant (summer 2014).

The expected users of this data include music historians and students, researchers in Computing Science and Information Science working on RDF and Semantic web projects and others in related humanistic disciplines involved in sociology, prosopography, and social network research.

The project website at McGill University will be maintained long after the granting period, and since the data will be made available on the web, copies made by other users of the data will ensure preservation and longevity. Furthermore, all archiving of digital data and code for this project will be managed by the Johns Hopkins University Data Management Services (JHU DMS) using the JHU Data Archive systems (with a fee of 2% of the Direct Costs). The archive was developed by the NSF-funded Data Conservancy, and designed according to Open Archival Information System to provide data preservation and web-based access with data citation. JHU DMS will also provide consultative support to select and prepare data files for ingesting, sharing, and preservation for five years after completion of the project with the option for an extension.

²⁶ <http://creativecommons.org/licenses/by/3.0/> with the attribution given to NEH for its support.

²⁷ <http://www.foaf-project.org>

²⁸ <http://motools.sourceforge.net/event/event.html>

²⁹ Github is a third-party public source code depository with over 100,000 users and over 2 million repositories.

³⁰ <http://gate.ac.uk>

³¹ <http://www.w3.org/TR/rdf-sparql-query/>