

6 Data management plan

Project Title: Enhancing Music Notation Addressability **Institution:** University of Maryland, College Park
Project Director: Raffaele Viglianti **Budget:** \$59,971
Beginning: 6/1/2014 **Ending:** 5/31/2015 **Duration:** 12 months

This data management plan was created on September 5, 2013, 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.

Roles and responsibilities

Project Director Raffaele Viglianti will be responsible for implementing this data management plan and for arranging long-term preservation of project data. It is expected that Viglianti will spend 2.5% of his allocated effort working on data curation activities. After data are deposited with the Haverford College Libraries, data curation staff from Haverford College will be responsible for maintaining access to the data.

Types of data

This project will produce two distinct types of data which we call "code" and "nanopublications". Management of the code will be MITH's responsibility; management of the nanopublications will be the responsibility of Haverford College and will be hosted as part of the current Du Chemin: Lost Voices website.

The "code" will consist of software packages written in Python or similar scripting language with accompanying documentation to be published on a software source code hosting platform such as GitHub.

The "nanopublications" will consist of RDF XML documents. They will be hosted and preserved at Haverford College.

Data standards and capture

The source code for software produced during this project will be open and freely available. We will be using an Apache 2.0 open source license. Documentation will be available on the world wide web in open formats including HTML.

Data and metadata formats

Software source code will be maintained as "plain text" files encoded using the UTF-8 character encoding standard. Nanopublications will be stored and managed as XML files, also encoded using the UTF-8 standard. The XML vocabulary used will be that for RDF/XML as defined by the W3C Recommendation (<http://www.w3.org/TR/REC-rdf-syntax/>).

Metadata or "documentation" for the software source code will consist of several components: inline comments in source code files, automatically-generated API documentation appropriate to the programming language chosen, and a README file explaining basic installation and usage (located in the root directory of the project). MITH will utilize packaging and distribution tools appropriate to the programming language selected to create basic descriptive and administrative metadata for the software packages.

For the Python language this consists of a plain text document with required values as specified in "PEP 314 - Metadata for Python Software Packages v1.1" (<http://www.python.org/dev/peps/pep-0314/>).

For nanopublications, the project will use established RDF vocabularies to provide metadata: the Nanopublication guidelines version 1.8 (<http://www.nanopub.org/guidelines/1.8/>) and the Open Annotation specification version 1.0 (<http://www.openannotation.org/spec/core/>).

Access and dissemination

All software source code created for this project will be made available on GitHub as soon as it is created. This mechanism allows other researchers to download and build upon the work of this project immediately. Source code will be made available under the terms of open licenses approved by the University of Maryland in consultation with the Office of Technology Commercialization. The project will use a license that supports cost-free re-use of the data without requiring specific clearance from the Project Director or other rights holders (e.g., Apache License Version 2.0).

Nanopublication data will be made available at the end of the project via the Du Chemin: Lost Voices website hosted by Haverford College. Nanopublications will also use a license that supports cost-free re-use of the data without requiring specific clearance from the Project Director or other rights holders for most purposes but discourages commercial exploitation of the data by outside parties to preserve the right of the project staff to seek all relevant publication opportunities for this data (e.g., Creative Commons Attribution, Non-Commercial).

Data storage, security, and backup during the active life of the project

At least two copies of the source code and documentation will be actively managed and stored during the life of the project. One copy will be stored on servers managed by GitHub. Since source code will be available under an open license, there is no cost for storage of this copy. A second, local copy of the data will be stored on servers managed by the University of Maryland College of Arts and Humanities. 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. Backups occur nightly. 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.

Similar arrangements are made at Haverford for the nanopublications through the Haverford College Libraries in "Triceratops" (<http://triceratops.brynmawr.edu/dspace/>), the tri-college digital repository. Triceratops provides secure, long-term preservation and web-based access to locally-generated digital content. See letter of commitment from Haverford College Libraries attached.

The specific volume of storage for this project is not anticipated to exceed 500 MB.

Long-term preservation

Within three years from the end of the grant period, at least one copy of the software source code and the nanopublication data will be permanently archived with the Haverford College Libraries in "Triceratops", the tri-college digital repository. Triceratops provides secure, long-term preservation and web-based access to locally-generated digital content. Copies of the data may also be deposited with other suitable repositories as identified by the Project Director. No data will remain on servers controlled by MITH. Data will remain publicly available through the libraries' digital repository. Haverford College has an interest in investing in the management of data created by affiliated researchers and in providing digital preservation services, such as file validation, integrity checks, and, if needed, format conversion.