

Data Management Plan

Creating and publishing open data sets is what Scribe is intended to facilitate. As with other NYPL Labs projects (*What's on the Menu?*, *Map Warper*, *NYC Chronology of Place*), data is made available to the public almost in real-time to scholarly and technical users in a range of formats and services: exports, APIs and simple user queries. All data is multiply backed up on both internal servers and drives, and redundant storage in the cloud. Eventually our user-generated data sets will be stored in NYPL's Fedora-based digital repository.

Zooniverse projects make data available to the public in a number of formats via large Data Releases at the conclusion of a project⁶ representing all data contributed by human and machine classifiers to a project. For sharing data with research teams prior to a major Data Release, Zooniverse creates private shares using Amazon S3 with time expiring keys sent to the research teams to download pre-release data. They also have a number of automated backup and data processing tasks that run pre-defined data management routines.

While Scribe in no way mandates or can enforce responsible data management by projects that might adopt it, NYPL Labs and Zooniverse will continue to lead by example, hopefully instilling these best practices in the culture that springs up around the too.

Collection

The data to be produced by this project fall into two primary categories: software code (primarily) and code documentation (secondarily).

Code will be managed using Git (the versioning software that powers GitHub), and will track every change to the code, and will be hosted using GitHub.

Documentation will be treated the same as software code: as sets of text files, web pages, and written reports. We will also be including sample materials (i.e. historical documents) for tutorials and example code, which we will include as part of this set of documentation and will be tracked the same way. This documentation will include information on how we built Scribe, how to get started using it, technical documentation of its inner functions and APIs, design documents guiding developers and scholars through the mounting of a structured data transcription project, and explanation of how such projects within the software architectures of Zooniverse and NYPL.

Management

Overseeing data management will be the project's technical director, David Riordan of NYPL Labs. He will ensure that local copies of all data are retained and that all data project data is included in the git repository for the lifespan of the grant.

Sharing

The software that will be designed for Scribe will be open source (i.e. free to use, copy, distribute, modify) and will be made freely available to any user through Github.com. Subsequent releases of the project will also be published on GitHub. We are still evaluating

⁶ i.e <http://data.galaxyzoo.org/>, <http://www.milkywayproject.org/data>

different licensing options, but are likely to employ an Apache License (as is currently the case with the Scribe prototype).

Preservation

In addition to the open GitHub repository, NYPL will ensure copies of all project data will be backed up to a long-term offsite archival storage service such as Amazon Glacier. Local copies of the project will be onsite at NYPL and Zooniverse, as part of the code repositories on project developer computers.

A copy of the code and documentation will also be published to The Internet Archive upon the first Supported Release of the project in May 2015 for preservation.