Workset Creation for Scholarly Analysis Prototyping Project

Distributed Metadata Correction and Annotation

University of Maryland
Principal Investigator: Trevor Muñoz

Project Description and Purpose

As part of the Workset Creation for Scholarly Analysis (WCSA) project led by the HathiTrust Research Center (HTRC), the Maryland Institute for Technology in the Humanities (MITH) proposed to develop a set of services and interfaces that would allow scholarly research teams to pull metadata records from the HathiTrust APIs, correct and annotate these records using standardized vocabularies, gather corrections and annotations from other application instances, and export them in formats suitable for publication as linked data. MITH also proposed to produce a demonstration of an index service that would allow research groups to register their data publications in order to make them available to other groups through a discovery interface.

The purpose of this project was to prototype simple tools that could support correction and enhancement of HathiTrust metadata by research teams who are building worksets and collections from the larger Hathi corpus for their own projects. Presently, teams of researchers might maintain a spreadsheet or custom database of metadata corrections and annotations that would be used internally for search and analysis. This approach requires limited infrastructure and can be effective for individual projects in the short term, but when it is adopted by many projects over many years it becomes extremely inefficient. MITH's prototype demonstrates how a distributed system of data publishing, annotation, and sharing might be created around the HathiTrust by leveraging existing and emerging linked data standards such as Open Annotation (OA), and CSV on the Web (CSVW).

Summary of Project Accomplishments

MITH accomplished all project goals within the period of performance. Over the course of two cycles of iterative design and development the MITH team created an initial prototype application, tested this prototype with a group of scholarly target users, radically simplified the prototype design for metadata correction based on user feedback, and finally created a demonstration version of an accompanying web application for registering, discovering, and publishing enhanced sets of HathiTrust metadata.

MITH's partners in developing the distributed metadata correction and annotation prototype were members of the Foreign Literatures in America (FLA) project research team. The FLA project aims to challenge conceptions of "American literature" that turn upon the American citizenship of an author, since historically it is clear that foreign authored works—as well as works by immigrant authors who wrote in many languages and were not citizens of the United

States—have constituted an important part of the literatures and cultures of the U.S. Many of the primary materials that are the focus of the project are included in the holdings of the HathiTrust Digital Library, but discovery and analysis of these volumes can be challenging because of incomplete and inconsistent metadata. Transliterations of personal names and translations of titles can vary widely for a single work, as can the bibliographic fields used to identify translators, editors, and writers of introductions—all information that is essential to this kind of study of reception history. The non-standard and inconsistent encoding of places and dates of publication in library metadata also raises problems for the study of regional and temporal variation in the reception of foreign authors, which is an important focus of the FLA project.

Initial Prototype

The first phase of project work, which led to the initial prototype, focused on those portions of the overall workflow dealing with acquisition of HathiTrust metadata through data correction and enhancement. (Support for modeling, publishing, and distributing such changes was not part of this initial phase.)

Members of the FLA team built worksets of materials related to their research using the existing collection building tools provided by the HathiTrust Digital Library. The first version of the MITH team's prototype utilized a software library originally created by Travis Brown for the Princeton Prosody Archive to download and reformat bibliographic metadata for volumes contained in the sample public worksets created by FLA. This metadata was loaded into an instance of the Open Refine software running on a central server. Interactions with the Open Refine application were brokered by a web application created by MITH to provide functionality to log in different users for tracking provenance information about various data changes.

User Testing

The MITH team conducted a small user study of this initial prototype with the FLA team members who had created worksets. FLA team members were asked to interact with metadata for volumes in their worksets loaded in the Open Refine application. MITH staff provided a basic tutorial and walk through for using Open Refine. During the testing, a basic "speak aloud" protocol was used in conjunction with screencasting software to record both participants' impressions as well as on-screen actions. Users were asked to review metadata records, and, where appropriate, attempt to make corrections or additions based on their existing subject knowledge as well as any project-specific conventions (e.g. related to transliteration of certain names).

The results of user testing were clear—even with a very small sample. Participants reported being confused by the presentation of information from different metadata fields as the information appeared in the Open Refine interface. This presentation resulted from constraints in transforming MARC metadata, which, generally speaking, contains series of nested fields into flat rows and columns in Open Refine's spreadsheet-like interface. Participants were also

uncertain how to add additional data through Open Refine, which may have been due to insufficient training on the program's basic functionality unrelated to this test. Also, at this stage, the MITH team had not determined a solution for modeling added data so participants could not be given much specific guidance. User testing focused attention on the challenges—but also the vital importance—of moving from data representations suited to library use-cases (MARC serializations and/or complex RDF) to representations of data that researchers found more approachable—tabular, or spreadsheet-like formats.

Second Prototype

Given the user feedback and technical challenges arising from the architecture of the initial prototype (discussed further below), the MITH team decided to revise its prototype. The second prototype developed from work on the demonstration of an index service for corrected and enhanced metadata. Consequently, this prototype radically minimized the tooling related to acquisition and basic correction/enhancement of metadata.

The MITH team proposed a process for distributed metadata correction and annotation that uses existing HathiTrust Digital Library and HathiTrust Research Center tools as well as existing "off-the-shelf" applications such as Open Refine. New software developed for this project serves to create workflows for using such tools in concert, supported by open standards.

This workflow entails:

- Users create worksets through existing mechanisms provided by HathiTrust Digital Library and HathiTrust Research Center. Leveraging these functions means that basic user account management, authentication for protected resources, and large-scale full-text search are all already provided
- To support downstream components of the workflow, MITH proposed that HTRC could make workset metadata available for download in CSVW format. Presently, bibliographic metadata for materials in worksets can be downloaded in Comma Separated Value (CSV) format. For the purposes of demonstration, MITH used the HathiTrust Bibliographic API to download metadata and make it available in CSVW format (see https://umd-mith.github.io/fla-metadata/). CSVW format allows creators to provide a schema for CSV that associates columns of tabular data with semantic predicates thereby facilitating processing of tabular data into other linked data standards such as RDF.
- Users register this original metadata with the prototype web application (https://github.com/umd-mith/csvwww) MITH created for this project. This registration step captures basic administrative information—such as the source of data—used for basic provenance tracking. Though it falls outside the scope of the current prototype, this function could also support harvesting and discovery of enhanced metadata by HTRC or other research teams.
- In the second prototype, users are free to edit or enhance metadata however they see fit, using any application or custom programming. (For the purposes of testing, MITH

continued to use Open Refine to perform data corrections). The only requirement is that changed data be saved back to the original CSVW document.
- Users upload changed data back to the prototype web application which tracks changes to the data and creates a log. When enhanced or corrected data is uploaded users are asked to describe changes—in the manner of a "commit log."
- MITH's prototype application captures changes to data and accompanying commit messages and incorporates this information as annotations in OA format. These linked data annotations are henceforth included as part of the CSVW metadata and could be acted upon by future applications to display changes to metadata or potentially to merge data sets edited by different groups of users.

The second prototype was presented and demonstrated at HathiTrust UnCamp 2015, however, there was insufficient time in the period of performance for a second round of user testing with the original FLA team of data creators. Testing and refinement of a system based on the second prototype would be an important component of any future development.

Challenges Encountered

In developing a prototype workflow for distributed correction and annotation of HathiTrust metadata for this project, the MITH team encountered a number of challenges. The most significant challenges are listed here:

- Metadata for HathiTrust materials are currently available in MARC format. Linked data representations of HathiTrust metadata are under development but were not available to work with during the period of performance. The in-process state of metadata in linked data formats challenged the ability of the MITH team to precisely target annotations representing corrections or enhancements to metadata. In the prototype, MITH elected to use fragment identifiers tied to CSVW serializations of HathiTrust metadata to provide referents for annotations. Additional processing of CSVW data would be required to translate such annotations into forms usable for updating HathiTrust data directly.

- The complete information represented in bibliographic metadata is often overwhelming to non-library researchers interested in data for their own projects. In practice, this means that more prescriptive choices to be made in serializing HathiTrust metadata for correction and enhancement by researchers. Flattening complex hierarchical structures and eliminating irrelevant fields should be two important considerations but further work is needed on the usability of various metadata serializations for different use cases.

- The area of data correction and enhancement suffers from both immature tools and tools that may be reaching their end of life without clear prospects for upgrades or support. CSVW is an important emerging standard that offers great promise for translating between tabular and graph-based representations of data but it is an emerging

standard—currently only one implementation exists for "distilling" CSVW to RDF. Likewise, tools to support the production and consumption of Open Annotation linked data have not been adapted to this data pipeline use case. There are also newer entrants into this class of tools—for example, Dat, a version-controlled, distributed data tool (http://dat-data.com/), but it is too early to judge their potential application. Finally, the MITH team remains concerned about the sustainability of Open Refine, the most user-friendly data correction and enhancement tool available, now that it is once again a community-maintained project with relatively low levels of active development.

Project Dissemination and Deliverables

All software source code produced for this project is available on GitHub at the following locations:

https://github.com/umd-mith/csvwww
The main prototype for this project, umd-mith/csvwww, is a web application that lets implements a workflow for distributed metadata correction and annotation. Users load in data sets in CSVW format, and publish changes to them using Web Annotation standards.

https://github.com/umd-mith/fla-metadata
A set of sample metadata derived from the HathiTrust Digital Library Bibliographic API and formatted as both CSV and CSVW documents.

https://github.com/umd-mith/hathitables
A command line utility for generating CSVW formatted data for worksets built through either the HathiTrust Digital Library or the HathiTrust Research Center collection build. Used to create the sample metadata found at: https://github.com/umd-mith/fla-metadata

https://github.com/umd-mith/hathilda
The underlying Python library used by the umd-mith/hathitables utility. This library contacts the HathiTrust Digital Library Bibliographic API, downloads volume metadata, and serializes this metadata as JSON-LD.

All software source code is licensed for redistribution and reuse under Open Source Initiative (OSI) approved licenses.

Principal Investigator Muñoz presented on the Distributed Metadata Correction and Annotation prototype at the HathiTrust Research Center UnCamp 2015, held March 30-31, at the University of Michigan. Presentation files (as well as interim and final reports from this project) may be found in the University of Maryland institutional repository at:
http://hdl.handle.net/1903/14717

Changes to Personnel and Project Management

The key personnel of the Maryland team changed substantially between the initial proposal and the period of performance.

Original Principal Investigator Travis Brown and MITH Software Architect James Smith both left the University of Maryland to pursue other opportunities shortly after the prototyping grant was awarded. Brown was to have been responsible for overall direction of the project and the development of project software, data models, and data sets. Smith was to have provided consultation on data modeling and conducted reviews of software code.

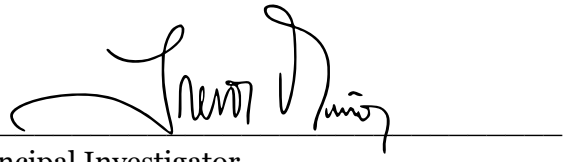
Trevor Muñoz, Associate Director of MITH, assumed Principal Investigator duties. Muñoz was responsible for the overall direction of the project and consulted on data modeling. Development of the first iteration of project software was led by Raffaele Viglianti, MITH's Research Programmer. Development of the second iteration of project software, as well as of project data models, and data sets was led by Ed Summers subsequent to his joining MITH as Lead Programmer in September 2014.

As initially projected, Stephanie Sapienza, MITH's Project Manager, scheduled and managed project meetings and tracked progress toward deliverables. Sapienza also led a key user testing session to evaluate the first iteration of the project prototype with Humanities Consultant, Dr. Peter Mallios, and members of the FLA research team.

Projected vs. Actual Expenses

Budgeted expenses for this project were $39,690. Grant award was $36,690. Actual expenditures amounted to $36,689.01.

Actual expenses varied from projections in the initial project budget in the following ways: 1) Personnel salaries were redistributed after the departure of Brown and Smith (described above); 2) Due to personal circumstances, Peter Mallios's time was curtailed due to personal circumstances and funds projected for his stipend were reallocated; 3) Expenditures for domestic travel for Muñoz to present on the project at the HathiTrust UnCamp 2015 were not included in the original budget; 4) Demand for Amazon Web Services computing resources was less than anticipated and the project did not make use of funds for this purpose.



Principal Investigator

06-15-2015

Date