White Paper

Report ID: 2879359

Application Number: HD-229031-15

Project Director: Edward E. Baptist

Institution: Cornell University

Reporting Period: 5/1/2015-8/31/2016

Report Due: 11/30/2016

Date Submitted: 12/23/2016

White Paper
HD-229031-15

Freedom on the Move:
A Crowdsourced, Comprehensive Database of North American Runaway Slave Advertisements

Project Director: Edward E. Baptist
Cornell University
December 22, 2016

Freedom on the Move
NEH Digital Humanities Start-Up Grant
Final Report

Comparison of actual accomplishments with project goals:

The Cornell project staff began by conferring with our primary partners, Mary Niall Mitchell at the University of New Orleans and Joshua Rothman at the University of Alabama, to create common file and data field naming conventions and definitions of standard categories across our three collections, as planned. Working with these key partners, we have created and successfully tested a network infrastructure for data upload, adjusting the initial FOTM data model to support the demands of the new collections. We are in the process of integrating these two primary collaborator collections, which will add approximately 15,000 ads to our working database.

, We have created program connectors to automate bringing the collaborator files into the FOTM database. We proposed building a modular set of programs to map, validate and bring in data from disparate sources to the FOTM repository via Java and code and via command line. The first fully-functioning version of these programs has been created and is successfully being used to bring data into the crowdsourcing application. Future enhancements to this software will likely focus on improved validation and a more user-friendly interface.

- The FOTM mapping script will now automatically rewrite shared data to be FOTM-compliant. This script has been successfully tested with our partner collections and is fully functional, with the caveat that currently some irregular fields must be entered by hand by users. We expected to refine this over the second part of the project term.
- The FOTM mapping script has been written and used to partially transform ~10,000 ads to the FOTM standard.
- The Extract, Transform, and Load Script will take a project that is currently in the FOTM standard, write it to the FOTM database, verify the quality of the data, and deploy it to the production database.
- As of June, this script was not quite complete. When complete, it will deploy projects (once they are rendered compliant to the FOTM standard) via the Transactional Application Programing Interface (API).
- In keeping with our initial work plan, we scheduled a Spring workshop to introduce current and potential research and archival partners to the collaboration toolkit and data analysis possibilities of FOTM. This was held May 12-13, 2016. It was a success, and we have now built collaborative relationships with the participants, including Rob Nelson of URichmond and Graham Hodges of Colgate, Tom Costa at UVA-Wise, representatives of the Maryland State Archives, and Liz Pryor of Smith.
- We ran into two snags. The first was the unexpected departure of Jeremy Williams, CISER's lead programmer/developer, for a private-sector job at the beginning of May. This essentially froze the project in place for almost six months as CISER attempted to hire a new programmer/developer. When Brandon Kowalski was hired in the fall, it understandably took him some time to get up to speed on the code base for FOTM.
- The second snag then appeared, as Brandon discovered an error in the data model structure which unnecessarily reproduced category definitions (metadata) about each ad. This clogged the database as we added new data, slowing performance and creating

difficulty in indexing.

- Therefore, Brandon has now turned to fixing the specific error, cleaning up the database, and pushing us all through a much-needed overhaul of the data model. This will in turn enable the completion of the transactional API and the uploading of new data in 2017.
- Brandon also located and fixed security vulnerabilities.
- We established a newer, easier, and faster procedure for updating access to the crowdsourcing app, which will allow us to do more testing over the upcoming semester.

These developments have been documented by project staff (see appendices.) In the coming months, we will continue to refine explanatory user documentation and generic FOTM guidelines for future collaborators.

Throughout the project term to date, we have also been reaching out to new collaborators, expanding our sense of the complexity the FOTM infrastructure will eventually need to support. In addition to the automated processes described above, we have also created Python scripts to translate metadata from other common platforms to the FOTM data model. This development is the result of a collaboration with researchers at UNC-Greensboro, and the scripts are available as part of the growing FOTM collaboration toolkit. As another result of this new partnership, we have developed MOU documents guaranteeing appropriate citation credit to future project partners; these, too, will become template documents in the FOTM collaboration toolkit.

Remaining work for the project and adjustments to initial project timeline:

We are in the process of creating a FOTM index for optimized searching and query API. We initially proposed a target completion of 1/01/16 for this task; in light of shifts in personnel (described below) and expansion of our outreach and collaboration framework, and in light of the need to rebuild the data model, we are still not finished with this element of the project. We initially anticipated a completion target of 9/01/16.

Our initial work plan proposed building services (APIs) to query and download data from the FOTM database in a variety of formats: providing data in raw json, csv, and also offering downloads in formats that can be opened by popular quantitative and qualitative statistical software (SAS, SPSS, Stata, Atlas.ti). This is not complete due to the staffing turnover we have experienced, but given that there is nothing inherently difficult in this work, we don't anticipate difficulty completing it.

Changes to workplan or methodology:

A programmer (unnamed) was part of the original budget; we were able to hire a programmer whose skills were such that we could expand the level of effort associated with this position and Williams maintained his initially-proposed role of supervisory responsibility for all programming, data modeling, and software development to build the infrastructure for the project. Reducing his level of effort on the project allowed us to extend the temporary programmer's time to accomplish more tasks, under Williams' supervision. Shifting more of the programming effort to the hired programmer also made it possible for Williams to engage in

more project outreach and communication efforts.

However, Williams left CISER in May for a private sector position. Since October, we have been able to begin building a working relationship with CISER's new programmer/developer Brandon Kowalski, and his supervisor Janet Heslop. This has already been a productive partnership that has enabled us to address some issues with the FOTM platform that will make future work easier and more successful.

Role of automation and hardware, software, staffing problems:
Automating data upload and integration processes is crucial to the FOTM work plan; our efforts are detailed above. We have no significant issues to report related to hardware or software. Staffing complications are documented above.

APPENDIX A: DOCUMENTATION

Please see: http://www.ciser.cornell.edu/fotm/docs/etl/ for complete documentation. Description of the documentation is below:

Documentation of the code written as part of the Freedom on the Move project between October of 2015 and January of 2016. This code will be part of the solution to satisfy the following requirements:

1. Provide a standard method to ingest the data from any collaborator into the database without extra burden on a collaborator, preferably without requiring a collaborator to change anything to their current process.
2. Perform Optical Character Recognition (OCR) on the slave runaway ads as part of the ingest process, relieving this burden from the collaborators.
3. Add data from external repositories such as the one maintained at UNC Greensboro.

APPENDIX B: DIAGRAM OF SHARED-DATA INGEST ACTIVITY