Journal of World-Historical Information
JWHI

RESEARCH REPORT

Marieka Arksey and Ruth Mostern

The Data Hoover Project

The purpose of the Data Hoover Project (managed by CHIA PI Ruth Mostern, conducted by Graduate Student Researcher Marieka Arksey) is to study data use, creation and curation in the historical quantitative social science (HQSS) fields. Informed by an understanding of community needs and practices, the Data Hoover Project is intended to identify content to populate the CHIA Dataverse repository, inform repository and metadata design, and identify the opportunities and barriers that inform eager and active participation in CHIA on the part of HQSS data owners.

Surveying Data Owners and Users

The Data Hoover Project has been circulating a short survey about the use of digital data repositories in order to discover how scholars in HQSS use datasets in their current work and how they would ideally like to see datasets and data repositories evolve to support improved collaborative frameworks for data sharing. We began with a plan

Volume 2-3, No. 1 (2014-2015) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2015.26 | http://jwhi.pitt.edu

to use a Skype version of the survey for face-to-face interviews, but when we discovered how difficult it was to schedule sessions with informants, we moved to an online version. We have advertised it widely on listservs and social media, and we have received several dozen responses which are starting to yield consistent and interesting insights. The initial analysis of that data, along with a review of other scientific literature about data sharing, has allowed Arksey and Mostern to complete an article about the Data Hoover Project and its implications for the field, which is currently under review at the *International Journal of Humanities and Arts Computing.*

The Data Hoover survey is online at http://www.chia.pitt.edu/datahoover.html, and we invite participants to respond. In the coming months, we will conduct focus groups, based on the survey, at the Social Science History Association Annual Meeting in Baltimore in November, and at the University of California, Berkeley D-Lab. We are also planning to contact the NSF Social, Behavioral and Economic Directorate and ask for help in publicizing the survey to their contact lists, and we will continue to identify major HQSS centers and repositories worldwide and communicate with their affiliates.

Building a Collection

The Data Hoover Project has also been actively involved with dataset collection. We have focused on the theme of historical climate and environmental data. In particular, we have worked extensively with NOAA Historical Paleoclimatology data. Historical climate is a topic that is focused enough to enable a good test of what is available and what issues might arise. Historical climate is also a topic with plentiful publicly available datasets, and one that is high impact for this community, providing useful base data for many different HQSS projects.

The Data Hoover Project has the very specific task of collecting large datasets to populate CHIA's data repository. With this and our goal of global comparisons in mind, we have focused on climate and environmental data at the global or regional scale, large enough in scope, as well as temporal and spatial scale, to enable global comparison, and broadly covering the period 1500-1950.

Within this rubric, identifying and formatting specific datasets has been an interesting and informative exercise. It has involved identifying bibliographical information, communicating personally with dataset owners or repository managers, wrestling with languages other than English, and extracting information from tables embedded in print articles. Datasets often need to be reformatted extensively to fit the structure of the CHIA repository, and new metadata needs to be written, especially when the datasets themselves do not included detailed data definitions.

The process of working with these materials has helped us to understand what will be needed to scale up to a repository like the one that CHIA envisions at full buildout, why the Data Hoover role is such an important part of the process, and what kind of collection philosophy needs to be developed. As the CHIA infrastructure becomes more stable and as the ingest process and interface evolves, some of these issues will be resolved, and the Data Hoover Project's insights will help to inform such technical development.

This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.