SOCIAL NETWORKS AND ARCHIVAL CONTEXT OPENREFINE PLUGIN

DIGITIZING CULTURE

An Undergraduate Thesis Portfolio
Presented to the Faculty of the
School of Engineering and Applied Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Jessica Xu

April 28, 2020

SOCIOTECHNICAL SYNTHESIS

The public, researchers, historians, and scientists now have more access to online materials than ever. Digital resources are more easily manipulated and changed than their physical counterparts. Thus, ensuring that digital information is both accurate and correctly represented is more important than ever. The technical research report summarizes the development of a standalone plugin for the archival institution, Social Networks and Archival Context, using OpenRefine, an existing open source project. The plugin has enabled participants to upload and consolidate large amounts of data on historical artifacts accurately and efficiently. The STS research paper investigates complications in attempting to digitize resources with a cultural heritage background. Both the technical and STS topics are heavily involved with digital archives, particularly with the accuracy of the information stored in them.

Social Networks and Archival Context (SNAC) is an archival institution that works with a variety of other institutions to build an archival collection, but each of these institutions has a different structure for storing records. Relationships between different entities, labels for certain types of data, and the hierarchy of the data itself are inconsistent from each outside institution. SNAC needs to reconcile the differences between the outside data and its own data storage structure before importing the data into its database. It is extremely impractical and time consuming to clean up the data manually. The goal of the technical project was to develop a simple and fault-proof interface for reconciling that would be easy for users to learn and use in the form of a plugin. This project will help improve the timeliness and accuracy of data insertion into the online repository.

The plugin has enabled participating institutions and individuals to upload and consolidate certain large tracts of data on historical artifacts. As originally intended, the project produced a stable product in the form of an extension to an existing open source project that solves most of the problems presented by the client. The actual users can upload their data through the tool, consolidate changes, and push them onto the official Social Networks and Archival Context (SNAC) database. Usually, inserting records into the database via the website interface takes a few minutes every entry, which leads hundreds of entries, which takes hours in a day. With the tool, hundreds of entries can now be processed and pushed to SNAC in a couple minutes.

The STS research paper focuses on the process of digitizing documents and resources that have a cultural background. It asks if the complications that arise during the process can be mitigated by viewing the situation through the lens of the Social Construction of Technology theory developed by Pinch and Bijker. The paper uses results from several different studies about the problems with digitizing cultural artefacts that expose the biases in the current digitization process. Several manuals and reports that detail the current digitization process were also analyzed in order to pinpoint the many entryways for biases within the process.

The analysis of the studies and reports led to the conclusion that many of the issues in the archiving process stem from the fact that a majority of archival institutions are based in western societies, and thus many of the archivists that work on cultural projects have a western perspective as well. The biases that exist in the digitization process can be disastrous for the integrity of information represented in a digital archive. Yet, having poorly formed data that does not comply with standards can also be disastrous to an archive. Only by having both a strong technical understanding and a strong cultural understanding of a resource, can an accurate and

efficient digitization take place. Currently, the digitization process is best modeled using the Linear Handoff Model. All participants in the process pass on their work to the next recipient. Instead of handing off the work step by step, the archivist, the archival institution, the indigenous community, and a local expert if needed should all work together throughout the entire process to

The pace of digital transformation is accelerating worldwide, and many physical books, papers, and other resources are being migrated online. It is crucial that the right technology and right processes are used for the digital transformation. The technical paper introduces one such technology that can aid the insertion of resources into an online repository while the STS paper seeks to introduce a new procedure for digitizing resources.

TABLE OF CONTENTS