Undergraduate Thesis Prospectus

Archival Institutions and the Contribution of Records to SNAC

(technical research project in Computer Science)

Libraries and the Promotion of Archival Data

(STS research project)

by

Sandra Gould

December 10, 2019

technical project collaborators:
Charles Chang
Grace Wu
Jessica Xu
John Perez
Mark Jeong
Peter Tran
Victor Shen

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

signed: _____ date: _____

approved: _____ date: _____
Peter Norton, Department of Engineering and Society

approved: _____ date: _____
Ahmed Ibrahim, Department of Computer Science

General Research Problem

How can access to collections be improved? Libraries, archives, and similar institutions hold documents and other items for public use. Yet most archives offer digital access to only a small fraction of their materials due to technology limits, selections of resources to digitize, copyright law, and finances limiting the scope of digital preservation (Kastellec, 2012). Other archives are not adequately equipped for the overhead of digitizing their records and the upkeep of preserving them. While countries in North America, Europe, and Australia have developed reference frameworks for the management and long-term preservation of digital materials, in Latin American countries there are only a few isolated attempts (Voutssas, 2012).

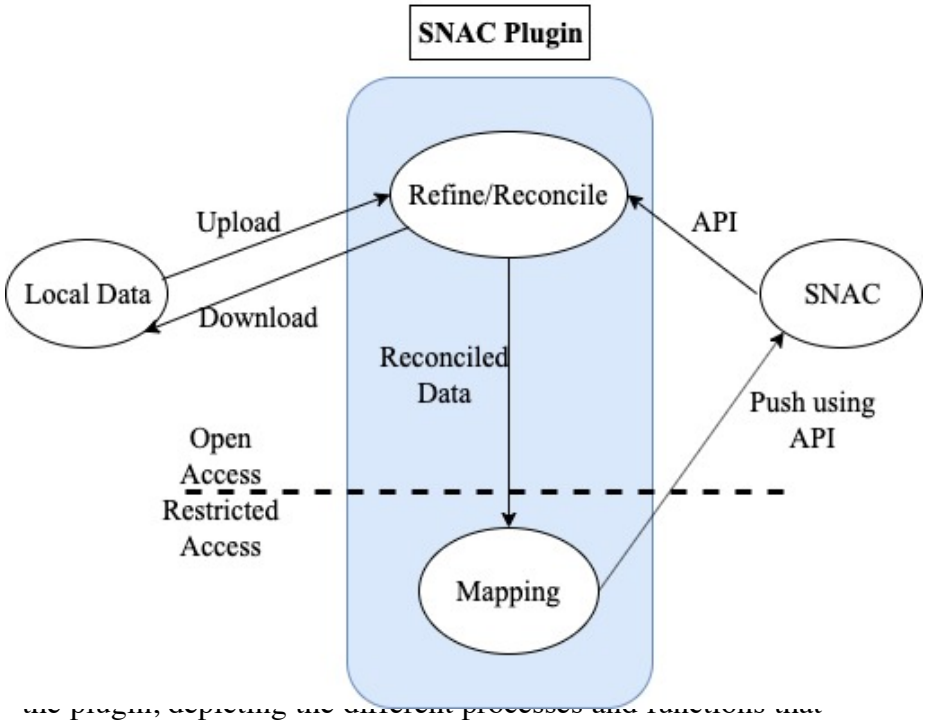
Social Networks and Archival Context OpenRefine Plugin

Social Networks and Archival Context (SNAC) is a free, online resource that allows users to discover information about the people and organizations that are documented in primary source documents and the connections between them (Social Networks and Archival Context [SNAC], n.d.). SNAC is used to locate archived collections as well as related resources held around the world. As an international cooperative, SNAC works to "build a corpus of reliable descriptions" of people and artifacts that link to and "provide a contextual understanding" of historical records (SNAC, n.d., para. 1). In order to create these contextual connections, SNAC sources its information from many different libraries and archival institutions. SNAC cooperates with over 4000 institutions to gather and reconcile data (SNAC, n.d.). 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 to clean up the data manually or with simple tools (Ham, 2013). The reconciliation of this data is vital to the functionality of an archival organization such as SNAC because it is crucial for efficient and accurate querying (Park, 2008).

The technical project seeks to develop a standalone plugin for Social Networks and Archival Context (SNAC) using OpenRefine. OpenRefine is an open source software that is community-maintained designed specifically for data normalization, transformation, and cleaning (Hill, 2016). It allows users to import and normalize data with a series of pre-existing default user interfaces after connecting to a target resource. OpenRefine provides a "powerful yet user-friendly interface" for experimenting with and querying data (Hill, 2016, p. 228).

With over 700 edits occurring to its data schema in week, Social Networks and Archival Context (SNAC) is no small data archive (SNAC, n.d.). The current workflow for refining and updating data in SNAC is quite difficult and inaccessible to inexperienced users. It involves users hitting SNAC's APIs for refining data on their server from the user's local machine. The technical project aims to greatly simplify this process by creating a streamlined plugin that will have all the functionalities needed to refine and upload data in one location. The logical



the plugin, depicting the different processes and functions that will be made available by the plugin (Xu, 2019).

flow and components needed for the project are illustrated in Figure 1. The plugin will serve as a connection between the user's local data and SNAC's server. It will allow users to import external data in the form of comma-separated values (CSV) files and make use of APIs provided by SNAC to reconcile and refine that data with SNAC's unique JavaScript Object Notation (JSON) data structure. The plugin will have two main user groups: privileged and unprivileged users. Both types of users will be able to use the plugin to format any data ported in using SNAC's organizational schema. Only privileged users will be able to then push the formatted data into SNAC's own database utilizing the APIs provided by SNAC. The technical project will provide an easy way to reconcile outside data with SNAC's existing data in addition with an improved user interfaced for an enhanced user experience.

The development will conduct biweekly customer meetings with the client in order to gather system requirements and get feedback about ongoing work. The minimum requirements for the plugin to be completed by the end of this semester include:

- Allowing users to import CSV data into the plugin
- Connect the data fields with different SNAC IDs
- Search for constellations in SNAC and match them to the imported data
- Allow a human editor to choose from several options to match for when the plugin is unsure
- Reconcile the imported changes based on the connection and matches
- Download the data that is now reconciled with SNAC's structure
- Users with privileges will be able to publish the data to SNAC

Desired requirements include:

- Users will be able to reconcile more complex data items like relationships and geolocations
- Users will be able to edit already existing resources and constellations

So far, no optional requirements have been specified by the client.

The technical project will be developed over the course of the two-semester capstone series led by Professor Ahmed Ibrahim from the Computer Science department, and will result in a technical report. To create this plugin, OpenRefine will be used, as it is a powerful tool for working with disorganized data that can "[transform] it from one format into another; extending it with web services and external data" (OpenRefine, n.d., para. 1). A similar project exists already for WikiData, but the technical project will create a new implementation specifically for Social Networks and Archival Context (SNAC). The plugin will hopefully provide a faster and more intuitive way for SNAC users to reconciliate and update data.

Libraries and the Promotion of Collections

How are library systems encouraging the use of collections? With misinformation more accessible than ever, libraries know the importance that the public be aware of and able to access collections housed at libraries. However, according to the Horrigan (2015), overall library use declined from 2012 to 2015.

Many institutions digitize their resources for easy public access, but in so doing they must comply with copyright law. Astle and Muir (2002) studied the risk of copyright infringement in digitization. An Independent Study Group sponsored by the United States Copyright Office and the Library of Congress recommended exceptions to the Copyright Act for digital archives (Rasenberger & Weston, 2008). Nelson (2009) found that most archives store

only miniscule fragments of their collections online. Most institutions either do not put their data in “formats that make them easily searchable and retrievable” or they “still don’t archive their data.”

The Social Networks and Archival Context (SNAC) digitizes collections for public access. SNAC is a “free, online resource that helps users discover biographical and historical information about persons, families, and organizations that created or are documented in historical resources.” SNAC aims to remedy the challenge of “discovering, locating, and using distributed historical records.” They do so by consolidating multiple archives into one location and distinguishing descriptions of people, families, and organizations from descriptions of historical resources. This is to “convey honest, transparent, and helpful context to archival researchers, not exhaustive or laudatory chronicles of our subjects” (SNAC, n.d.).

Publishing companies have recently been restricting access to ebooks, a common electronic medium in libraries. For instance, Macmillan Publishers is implementing a new ebook lending model. They give only one copy of an ebook to each library system upon the ebook’s publication. Libraries cannot purchase additional copies until eight weeks after publication. The model is a response to falling revenue from ebook reads: “well under two dollars and dropping, a small fraction of the revenue” that Macmillan shares with Macmillan authors, illustrators, and agents on a retail read, as 45% of Macmillan’s ebook reads are free from libraries (Sargent, 2019).

The American Library Association (ALA) advocates for more open access to library resources and services to everyone, regardless of age or the material’s content. ALA disapproves of governing bodies requiring “internet filters or other technological measures that block access to constitutionally protected information” (ALA, 2019).

Under the Library Services and Technology Act , libraries in the United States receive most of their federal funds from the Institute of Museum and Library Services (IMLS) (ALA, 2019). An IMLS goal is to "invest in tools, technology, and training that enable people of all backgrounds and abilities to discover and use museum and library collections and resources" (IMLS, 2018). From 2011 to 2016, approximately 36 percent of IMLS project grants were intended to improve information access (IMLS, 2017). Through its grants to libraries, the Andrew W. Mellon Foundation aims to "strengthen, promote, and defend the centrality of the humanities and the arts to human flourishing and to the well-being of diverse, fair, and democratic societies." It funded the Council on Library and Information Resources (CLIR) for the Digital Library of the Middle East (Andrew W. Mellon Foundation, 2019). CLIR (2019b) itself "aspires to transform the information landscape to support the advancement of knowledge." With the help of the Mellon Foundation, CLIR's Digitizing Hidden Special Collections and Archives program supports institutions "digitizing rare and unique content" (CLIR, 2019a).

Indigenous peoples, however, take issue with the public digitization of their resources. The Tulalip Tribes of Washington state that "there is no public domain in traditional knowledge," especially concerning secret and sacred knowledge systems. For them, treating such knowledge systems as belonging in a free-for-all public domain "would be detrimental to the cultural survival of Indigenous peoples" (Karlsson, 2019).

References

Andrew W. Mellon Foundation. (2019). Report 2018. https://mellon.org/media/filer_public/72/e7/72e72b50-3397-4fc2-9aa4-29eeb3a77b12/2018.pdf

ALA (2019). American Library Association. Advocacy and Public Policy. http://www.ala.org/advocacy/advocacy-public-policy

Astle, P.J., & Muir, A. (2002, June 1). Digitization and preservation in public libraries and archives. *Journal of Librarianship and Information Science*, 34, 67-79. https://doi.org/10.1177/096100060203400202

CLIR (2019a). Council on Library and Information Resources. About. https://www.clir.org/about/

CLIR (2019b). Council on Library and Information Resources. Digitizing Hidden Special Collections and Archives. https://www.clir.org/hiddencollections/

Ham, K. (2013). Free, Open-source Tool for Cleaning and Transforming Data. *Journal of the Medical Library Association*. https://www.ncbi.nlm.nih.gov/

Hill, K. M. (2016). In Search of Useful Collection Metadata: Using OpenRefine to Create Accurate, Complete, and Clean Title-Level Collection Information. *Serials Review*, 42(3), 222-228. Retrieved from https://www.sciencedirect.com/journal/serials-review

Horrigan, J. (2015, Sept 15). Libraries at the Crossroads. *Pew Research Center.* https://www.pewresearch.org/internet/2015/09/15/libraries-at-the-crossroads/

IMLS (2017). Institute of Museum and Library Services. (June 30). Institute of Museum and Library Services Funding Report by State FY 2011-2016. https://www.imls.gov/sites/default/files/publications/documents/imlsfundingreportallstates.pdf

IMLS (2018). Institute of Museum and Library Services. (Jan). Transforming Communities. https://www.imls.gov/sites/default/files/publications/documents/imls-strategic-plan-2018-2022.pdf

Kastellec, M. (2012). Practical Limits to the Scope of Digital Preservation. *Information Technology & Libraries*, *31*(2), 63–71. https://doi.org/10.6017/ital.v31i2.2167

Karlsson, Á. G. (2019). Copyrighting the Copies. *Library Quarterly*, *89*(4), 333–347. https://doi.org/10.1086/704961

Nelson, B. (2009, Sept 09). Data sharing: Empty archives. *Nature*, *461*, 160-163. https://www.nature.com/articles/461160a

OpenRefine. (n.d.). Introduction to OpenRefine. http://openrefine.org/

Park, J-R. (2008). Metadata Quality in Repositories: a Survey of the Current State of the Art. *Cataloging & Classification Quarterly*, 47(3), 213-228. doi:10.1080/01639370902737240

Rasenberger, M. & Weston, C. (2008, Mar). The Section 108 Study Group Report. http://www.section108.gov/docs/Sec108StudyGroupReport.pdf

Sargent, J. (2019, July 25). Confidential – Embargoed until Thursday, July 25 at 10:15 AM EST. https://www.publishersweekly.com/binary-data/ARTICLE_ATTACHMENT/file/000/004/4222-1.pdf

SNAC. (n.d.). Social Networks and Archival Context. About SNAC. https://portal.snaccooperative.org/about

Voutssas, J. (2012). Long-term digital information preservation: challenges in Latin America. *Aslib Proceedings*, *64*(1), 83–96. https://doi.org/10.1108/00012531211196729

Xu, Jessica (2019). Figure 1: SNAC Plugin Model.