Prospectus

Social Networks and Archival Context (SNAC) OpenRefine Plugin
(Technical Topic)

The Interplay Between Lurkers and Online Communities
(STS Topic)

By

Grace Wu

October 30, 2019

Technical Project Team Members:

Charles Chang, Sandra Gould, Mark Jeong, John Perez, Victor Shen, Peter Tran, Jessica Xu

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: _____

Approved: _____ Date _____
Rider Foley, Department of Engineering and Society

Approved: _____ Date _____
Ahmed Ibrahim, Department of Computer Science

Introduction

Social Networks and Archival Context (SNAC) is a free, online resource that helps users discover information about people, families, and organizations that are documented in historical resources (primary source documents) and their connections to one another. SNAC, like the more famous website Wikipedia, is an international cooperative that includes (but is not limited to) archives, libraries, and museums. SNAC seeks to build a collection of reliable descriptions of people, families, and organizations that link to and provide a contextual understanding of historical records (About SNAC, 2010). Presently, there are approximately 136 active users that contribute to the SNAC database (Jeong, 2019). While this is an excellent start, the number of users is affected by the current unoptimized method of data insertion. The present method of inserting data into SNAC is manually inserting data into form fields. This method does not indicate whether the records that are being inserted into SNAC already exist in SNAC (and need to be updated in SNAC) or if the records are completely new (and need to be inserted into SNAC).

To address this, the development team is creating an OpenRefine plugin to help compare and share data between SNAC and other archival organizations. OpenRefine is a Java-based tool that allows a user to upload data and then analyze, clean, reconcile, and augment it through an online interface. The technical topic section will address in detail how the plugin will be developed and the methods that will be employed in order to successfully create the product by the end of the year.

While the chosen technical project is not directly related to the STS topic, they are still connected. The technical project addresses how a tool (specifically, a plugin) will be built for an online cooperative, while the STS section of the prospectus analyzes how software developers

can design user interfaces that encourage more participation in online communities. A significant reason why users are unable to participate in online communities is due to the usability of the software (Preece, 2001). Thus, by considering factors like usability, the development team can design a user interface for the plugin that encourages more user participation and makes it easier for users to actively import and reconcile data into SNAC.

SNAC 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 (About SNAC, 2010). 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 (About SNAC, 2010, para. 1). In order to create these contextual connections, SNAC sources its information from many different libraries and archival institutions. SNAC cooperates with over 4,000 institutions to gather and reconcile data (About SNAC, 2010). 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 (About SNAC, 2010). 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 flow and components needed for the project are illustrated in Figure 1.

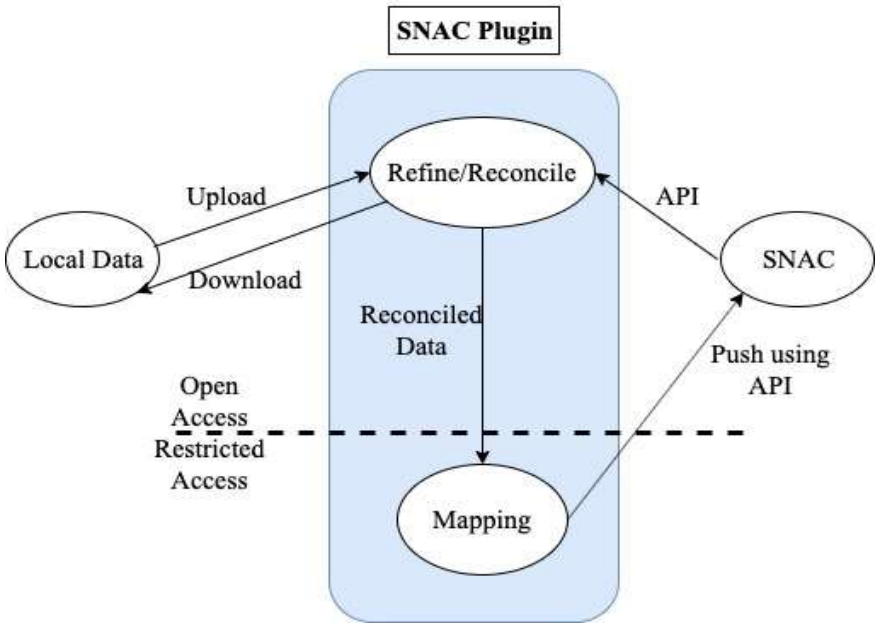

Figure 1. SNAC Plugin Model: An overview of the design of the plugin, depicting the different processes and functions that will be made available by the plugin (Xu, 2019).

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, 2010, 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.

Lurkers and Online Communities

I will focus on the interplay between user participation and online communities. Online communities, defined as "any virtual social space where people come together to get and give information or support, to learn or to find company", serve a wide demographic of users and serve a variety of purposes (Preece, 2001, p. 348). These communities can manifest in many forms including listservs, forums, instant messaging websites, and social networks. Online communities are formed and shaped by their particular interests which, in turn, affects how the user interface and the website itself are built to fulfill that purpose (Preece, Nonnecke, Andrew, 2004).

Oftentimes, the majority of users do not participate in online discussions. In a study done by Van Mierlo (2014), four digital health social networks (DHSN) were monitored and 578,349 DHSN posts were observed for up to eleven years. Statistics revealed that less than 25% of the

actors in each DHSN authored one or more posts. Users were sorted into three categories depending on how much they posted: 90% were “lurkers”, 9% were “contributors”, and the top 1% were “superusers”. This study offers evidence of the 90-9-1 principle within online communities and shows that there is a large demographic of users (lurkers) that software developers can encourage to participate in online communities through better designed user interfaces. Nonnecke and Preece (2001) define lurkers as “one of the ‘silent majority’ in an electronic forum; one who posts occasionally or not at all but is known to read the group’s postings regularly” (p. 294).

There are two main reasons why people lurk: lack of usability or personal reasons. In a survey done by Preece and Andrews (2004), 7.8% of users stated the lack of software usability as a reason for lurking. In order to encourage user participation, Preece (2001) offers a framework for sociability and usability in the context of an online community, where usability is defined as “how intuitive and easy it is for individuals to learn and interact with a product” and sociability is defined as “developing software, policies, and practices to support social interaction online” (p. 349). Table 1 gives a more comprehensive analysis of this framework and it can be used to measure the success of a community. With this information, software developers can better utilize features such as avatars, signatures, rankings, and point systems which enhances a person’s identity and encourages participation in a quasi-anonymous online community (Liao, 2007).

Framework	Design	How are they different?
Sociability	*Purpose*. A community's shared focus on an interest, need, information, service, or support, that provides a reason for individual members to belong to the community. *People*. The people who interact with each other in the community and who have individual, social and organization needs. Some of these people may take different roles in the community, such as leaders, protagonists, comedians, moderators, etc. *Policies*. The language and protocols that guide people's interactions and contribute to the development of folklore and rituals that bring a sense of history and accepted social norms. More formal policies may also be needed, such as registration policies, and codes of behavior for moderators. Informal and formal policies provide community governance.	Sociability focuses on human-human interaction supported by technology.
Usability	*Dialogue and social support.* The prompts and feedback that support interaction, the ease with which commands can be executed, the ease with which avatars can be moved, spatial relationships in the environment, etc. *Information design.* How easy to read, how understandable and how aesthetically pleasing information associated with the community is, etc. *Navigation*. The ease with which users can move around and find what they want in the community and associated websites. Many online community users have suffered from inconsistencies of data transfer and differences in interaction style and the website housing the community. *Access*. Requirements to download and run online community software must be clear. In addition, if high bandwidth and state of the art technology is needed to run the community there should be a low bandwidth text only versions and clear instructions about how to obtain it.	Usability focuses on how members of a community interact with each other via the supporting technology (human-computer interaction).

Table 1. Framework using sociability and usability to measure the success of an online community (Adapted from Preece, 2001).

Nonnecke and Preece (2001) offer several reasons for why people lurk that reflect users' different personalities: they want to learn more about the group first, they feel that there is no need to post, or they are still building an identity. Understanding a user's behavior and cognition will allow software developers to decide whether to include or exclude certain features when developing an application. It is equally important for software developers to consider the culture and values of their users. In a study by Malinen (2015), the author compares cultural differences between the US, Netherlands, and South Korea. South Koreans tend to value collective activities, which means they are more likely to participate in online communities of interest and organizations within their local communities. Therefore, if the community did not pertain to their own personal interests and was in a low-context culture, South Koreans were not as likely to participate in the community as compared to the users in the US and the Netherlands. Software developers need to consider their user base and ensure that the user interface fits the users' cultural values.

While some would argue that lurking is bad for an online community, Malinen (2015) argues that lurking is defined as a transformation from newcomer to regular member; it is a passive but non-negative way to enjoy an online community and can eventually allow users to feel like they belong in the group, as "both forms of participation, reading and posting, have a positive influence on the development of a sense of community, and spending time in the online community and reading messages may actually lead to closer attachment to the group" (Malinen, 2015, p. 232).

Lurkers and software developers in an online community can be considered through Habermas' theory of communicative action (Bohman and Rehg, 2014). In the context of online communities, when two or more users engage in a mutual deliberation and argumentation with

the goal of reaching a common understanding of a topic, discourse occurs, eventually resulting in a consensus (Habermas, 1984). A consensus will occur when a user successfully convinces other members of the community of an idea and they take up an affirmative position towards the claim. Once a consensus is reached, the conversation in an online community will significantly decrease in terms of number of related posts and quality of meaningful discussion. Otherwise, discourse arises, where "claims…are tested for their rational justifiability as true, correct, or authentic" (Bohman and Rehg, 2014, para. 22).

This cyclical process will occur many times throughout the lifespan of the online community. As long as there is recurring mutual deliberation in an online community, then users will be more incentivized to participate in the conversation, provided that software developers have created a user interface that promotes usability and sociability. Instead of measuring the "success" of an online community through the number of active users, it would be more meaningful to holistically examine the lifecycle of an online community through Habermas' theory of communicative action, analyzing the meaningfulness of posts and discussion.

Research question and methods

My research question is: Why do a small percentage of users contribute to discussions within online communities and how can software developers design user interfaces that encourage more participation? Answering this question will help build upon current Human-Computer Interaction (HCI) research and bridge the disconnect between software developers and the users of an online community. In order to further investigate and analyze the data, three different methods will be employed and interpreted in the context of the research question: interviews, surveys, and literature reviews.

Jennifer Preece is a subject matter expert in online communities and has written several papers on how usability and sociability factors into online communities. Since several of Preece's paper are repeatedly referenced throughout this prospectus, I plan to contact her via email and inquire about methods to gather data on lurkers and how characteristics like gender, age, nationality, etc. affect online forums. This would aid in how I could conduct online surveys and how to interpret such data. Similarly, I plan to contact and interview Yair Amichai-Hamburger and Brandie Nonnecke about lurkers to gain a wider perspective and inform my survey.

Another primary source of information that can be used to analyze the research question is through surveys. These surveys will be mainly used to gather information about lurkers in various online communities. I plan to conduct surveys primarily on reddit, Facebook, and Instagram, all of which are prominent online communities. In the survey, I plan to ask these questions:

- Why do you use this platform?
- What do you like about this platform?
- Has there ever been a time where you wanted to do something but the platform prevented you from doing it?

Since lurkers by nature do not post in online communities, it will be challenging to gauge what percent of the online community is actually participating. In order to incentivize participation, I will offer a gift card or a monetary raffle that will encourage users to participate. Surveying across platforms will show differences in how users use various online communities for their own purposes and ways in which platforms can become more usable. In doing so, I will

also analyze differences in updates of each platform and how effective each update was in increasing usability and sociability.

Finally, existing literature reviews will provide a plethora of information and studies that have already been conducted in the field. One literature review that I plan to extensively use is "Understanding user participation in online communities: A systematic literature review of empirical studies" by Malinen (2015). In this literature review, the author uses comparative case studies to evidence their claims. These studies have significant outcomes that will be used to shape my own interviews and surveys and help support or disprove my own findings.

Conclusion

My technical topic focuses on designing an OpenRefine plugin that will help users compare and reconcile data between SNAC and other archival organizations. The goal for this project is to create a deliverable that will be used by institutions like archives, libraries, and museums to effortlessly upload and insert large amounts of data into SNAC. Figure 3 is a Gantt chart that contains a potential timeline of features that the development team plans to implement in the following year. Ideally, the plugin for SNAC will be integrated into the current build and be used by active SNAC users. The creation of this plugin will open up opportunities for the open-source community to build off what the development team has created and make it more convenient for users to upload valuable information to SNAC.

My STS topic addresses the interplay between software developers and lurkers in online communities using a framework for usability and sociability and the Habermas Communicative Action framework. These frameworks will help understand why users lurk and how software developers can create user interfaces that encourage more participation. The timeline for my STS research will be contacting Jenny Preece, Yair Amichai-Hamburger, and Brandie Nonnecke in

mid-November as well as reading more literature reviews. Once I have obtained data on how to survey lurkers and how characteristics like gender, race, etc. affect online communities, I will use this information to formulate my online surveys. I plan to implement my surveys in late January through late March and gather surveys. While the research done in this topic is fascinating and is applicable in human-computer interaction, I do not have current plans to get the paper published.

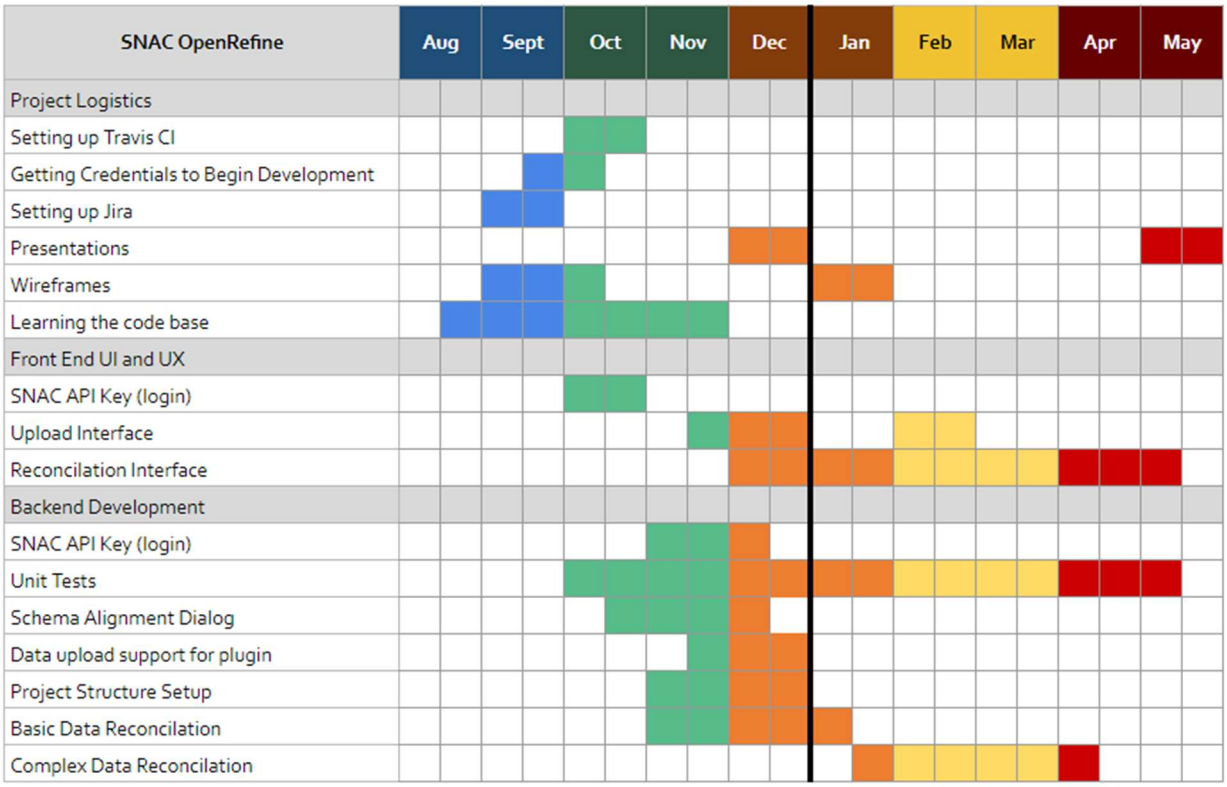

Figure 2. Gantt Chart: timeline for technical project. Bold line indicates separation of semesters.

References

About SNAC. (2010). Retrieved October 28, 2019, from https://portal.snaccooperative.org/about.

Bohman, J. and Rehg, W. (2014). Jürgen Habermas. *The Stanford Encyclopedia of Philosophy (Fall 2014 Edition).* Retrieved from https://plato.stanford.edu/archives/fall2014/entries/habermas/

Habermas, J. (1984). The theory of communicative action, Volume 1. Boston: Beacon.

Ham, K. (2013). Free, Open-source Tool for Cleaning and Transforming Data. *Journal of the Medical Library Association*. Retrieved from 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

Jeong, M (2019, October 28). Personal Interview with J Glass.

Liao, Y. Y. (2007). Promoting online discussion participation by integrating identity-enhancing feature from digital games (Doctoral dissertation, Ohio University), 27-39.

Malinen, S. (2015). Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in Human Behavior*, 46, 228-238.

Nonnecke, B., & Preece, J. (2001). Why lurkers lurk. AMCIS 2001 proceedings, 294.

OpenRefine. (2010). *Introduction to OpenRefine*. Retrieved from 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

Preece, J. (2001). Sociability and usability in online communities: Determining and measuring

success. *Behavior & Information Technology*, 20(5), 347-356.

Preece, J., Nonnecke, B., & Andrews, D. (2004). The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2), 201-223.

Tran, P. (2019). Figure 2: Gantt Chart.

Van Mierlo, T. (2014). The 1% rule in four digital health social networks: an observational study. *Journal of Medical Internet Research*, *16*(2), e33.

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