Social Networks and Archival Context Project
(Technical Paper)

How Can We Prepare for the Future of Artificial Intelligence?
(STS Paper)

A Thesis Prospectus Submitted to the
Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia
In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

John Perez
Fall, 2019

Technical Project Team Members
Charles Chang
Sandra Gould
Mark Jeong
Victor Shen
Peter Tran
Grace Wu
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

Signature John Perez Date 12/4/19
John Perez

Approved Ahmed Ibrahim Date 11/25/2019
Ahmed Ibrahim, Department of Computer Science

Approved _____ Date 12/4/19
Thomas Seabrook, Department of Engineering and Society

Introduction

First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

Isaac Asimov introduced his famous Three Laws of Robotics in one of the short stories from his collection "I, Robot" (1950) in which a robot finds it impossible to obey the laws simultaneously, resulting in a loop of repetitive behavior. With this publication Asimov inspired engineers to ponder how the human mind might work. Sixty years later, we have a humanoid robot that is capable of seeing, making facial expressions, and communicating through AI – she is also named Sophia and has been granted citizenship in Saudi Arabia (Gallo, 2019). How is artificial intelligence being regulated to prevent misuse? As robots become smarter with advances in technology, will they make our lives easier or more difficult? Since any technology is prone to misuse, the future of artificial intelligence is on us as a society (Marr, 2019). The technical portion of this paper will address my team's efforts in creating an extension for the Social Networks and Archival Context Project (SNAC). SNAC is an open-source tool that helps users discover biographical and historical information about persons, families, and organizations that are documented in primary sources and their connections to one another ("Social Networks and Archival Context", n.d.). Users can locate primary sources at over 4000 cultural heritage institutions like archives, libraries, and museums all over the world (Jeong, 2019). SNAC is a

brownfield project, meaning that it was in production prior to my team working on it. My team is implementing new functionality and maintaining an existing product.

Technical Topic

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 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

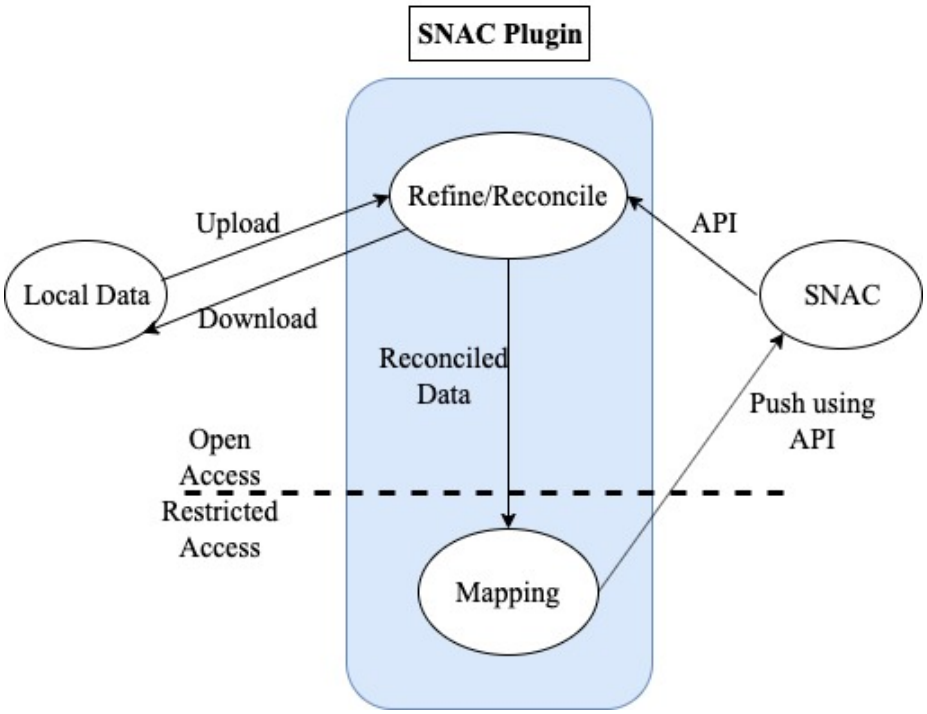

Figure 1: SNAC Plugin Model: An overview of the design of the plugin, depicting the different processes and functions that will be made available by the plugin (Xu, 2019).

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.

STS Topic

This paper will use the Social Construction of Technology (SCOT) to analyze the relationship between society and artificial intelligence. I chose this STS theory because artificial intelligence in theory long-precedes artificial intelligence in practice. As a result of this ordering, the technology as we know it has been mostly a result of social construction. Engineers were influenced by social factors like Asimov's Three Laws when creating artificial intelligence. The advantage of this framework is the ability to analyze and "distinguish cause and effect," unlike something like Actor Network Theory (Klein & Kleinman 2002). By effectively eliminating the variable of technology's influence on society, we can clearly see how society has influenced technology. The problem with SCOT is that different groups in different societies construct different problems, leading to different designs (Russell 1986). Realistically, any corporation could be a stakeholder because the proliferation of artificial intelligence could affect every corner of industry. From the stock exchange to biomedical sciences to agriculture to art to the justice system, if artificial intelligence has not already found a place in these fields, it surely will in the future. The potential pervasiveness of artificial intelligence can be worrying. However, any technology is susceptible to misuse. Since artificial intelligence is socially constructed, it is our responsibility as a society to determine what research within the field should be prioritized, how government policies need to change to account for artificial intelligence, and what strategies the artificial intelligence community needs to pursue in order to maximize benefits and minimize societal risk (Comets Labs Research Team, 2016). Relevant artifacts include training models for machine learning algorithms, public policies regulating artificial intelligence, and limitations of existing computer hardware. The most important of these artifacts is arguably training models for

machine learning algorithms because these are the most susceptible to misuse. It is vital that data scientists recognize their moral responsibility to train machine learning models with unbiased data. Only then can we digress from innate human biases rather than perpetuate them. This research is important because many people wrongfully fear artificial intelligence as a result of social media. The truth is that artificial intelligence is an incredibly powerful tool and because it is still a relatively new technology, there is minimal infrastructure to handle it. We, as a society, must push for regulation and policies to accommodate artificial intelligence so that we are prepared to harness its power for good while minimizing damages and misuse.

I will research what society might look like with the proliferation of artificial intelligence and how we can prepare to optimize the technology. The two parts of this investigation go hand-in-hand and are equally important. As artificial intelligence become more commonplace, our day-to-day lives will certainly change as the job market transforms and public policies adapt to accommodate the new technology. The changes can be positive or negative depending on how we as a society prepare for them. For this reason, I will investigate how we can prepare our society for such a transformation in a way that maximizes utility and minimizes societal risk. To investigate how emerging artificial intelligence might affect our society, I will research how specific industries anticipate to change as a result of artificial intelligence and how artificial intelligence has shaped the industry in the past. Some fields to research include ethics, politics, war, and economics. Since artificial intelligence has already permeated many of these fields, I can look into how the market has adapted to the technology. As for the future of each field, expert opinions would be the most credible and accurate predictions. With these approaches for research, I can reasonably predict what society will be like with more applications of artificial

intelligence. By looking at the ethical and political perspective, I can determine what preparations should be made in anticipation of artificial intelligence becoming more common.

Conclusion

Overall, the technical work that I will address is an extension to an existing product that networks historical entities together. This extension will automate the user's task of inserting data into the massive database containing over 3 million entries. Within the social analysis, I will research what effects artificial intelligence might have on society as it becomes more widespread, and how we can prepare to handle those changes in a way that maximizes the utility of the technology and minimizes societal risk. By performing the technical side, I will obtain experience developing software in the workplace while working on a product that will improve the lives of hundreds of users by automating processes (Jeong 2019). My research on the social aspects of artificial intelligence will provide insight as to how we can prepare for such a revolutionary technology. It will also improve our understanding of artificial intelligence and what transformations it might make in society as it becomes more widespread.

References

Asimov, I. (1950). *I, Robot*. Street & Smith.

Comet Labs Research Team. (2016, January 19). Algorithms Among Us - Societal Impacts of Machine Learning (NIPS 2015, Day 4). Retrieved from https://medium.com/@comet_labs/algorithms-among-us-societal-impacts-of-machine-learning-nips-2015-day-4-3cbd9a82cb38.

GitHub features: the right tools for the job. (n.d.). Retrieved from https://github.com/features.

Glass, J. (n.d.).

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.

Jira: Issue & Project Tracking Software. (n.d.). Retrieved from https://www.atlassian.com/software/jira.

Klein, H. K., & Kleinman, D. L. (2002). The Social Construction of Technology: Structural Considerations. Retrieved from https://pdfs.semanticscholar.org/4c17/4dc2d6a6a4f6b8a75df2bc255b287ba2acaa.pdf.

Marr, B. (2019, March 14). Is Artificial Intelligence Dangerous? 6 AI Risks Everyone Should Know About. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/11/19/is-artificial-intelligence-dangerous-6-ai-risks-everyone-should-know-about/#74010deb2404.

OpenRefine. (n.d.). *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.

Reynoso, R. (2019, March 1). A Complete History of Artificial Intelligence. Retrieved from https://learn.g2.com/history-of-artificial-intelligence.

Russell, S. (1986). The Social Construction of Artefacts: a Response to Pinch and Bijker Retrieved from https://journals.sagepub.com/doi/10.1177/0306312786016002008.

Social Networks and Archival Context. (n.d.). Retrieved from https://snaccooperative.org/.