Preprint of an article that will be published in the Papers of the Sixty-Second Annual Meeting of the Seminar on the Acquisition of Latin American Library Materials (SALALM)

Maximizing the Use and Exposure of Archival Data through Linked Open Data

Mairelys Lemus-Rojas (IUPUI University Library)

Timothy A. Thompson (Yale University Library)

Introduction

Archival institutions collect, describe, preserve, and provide access to primary and secondary materials. Collecting materials does not come without its challenges, since archivists have the difficult task of determining what is "worth" archiving. This decision is informed, in part, by an institution's interest in strengthening a particular area of its collection. Another important aspect of the archival process is to facilitate the discovery of and access to these unique materials. One way in which this is achieved is by creating finding aids for individual collections to provide contextual information and a detailed account of the collection's content. Finding aids are encoded using the EAD (Encoded Archival Description) XML schema. In an effort to provide the archival community with a mechanism for describing archival records, EAD version 1.0 was released in 1998 after a five-year development period (Pitti 2006, 10). By using the EAD XML schema we are ensuring that the information can be read by both humans and machines, which facilitates discoverability.

Traditionally, information about our materials has been shared in our library or archives website. However, despite the community's effort to share information about its resources by way of finding aids, finding relevant information can still be challenging. An understanding and knowledge of repositories and the collections they hold is necessary in order to locate relevant resources (Larson, Pitti and Turner 2014, 1). In this chapter, we will highlight the importance of collaboratively creating and curating data in an effort to provide more visibility and access to our

archival data in projects such as the Social Networks and Archival Context (SNAC)[1] cooperative, Wikipedia, and Wikidata.

SNAC

The archival community has been in need of a central place where archival authority data could be contributed and shared. SNAC was intended to address this issue. SNAC is an international cooperative that emerged as a way of providing the archival community with a central repository for maintaining authority data using the EAC-CPF (Encoded Archival Context—Corporate Bodies, Persons, and Families) standard. EAC-CPF is an XML schema that was designed to encode standardized information about the people and organizations associated with archival collections and their social context and networks. SNAC acts as a hub where information from different sources can be merged and presented to users in a cohesive way, facilitating the discovery of and access to archival collections. It also aggregates and displays images representing collection creators. These images are pulled from Wikimedia Commons, a sister project to Wikipedia, often referred to as Commons.

The U.S. National Archives and Records Administration (NARA) serves as the host for the SNAC cooperative, and the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia served as the infrastructure host during the first 2-year pilot project phase. In 2017, SNAC reached the end of its first pilot phase, where the focus was on issues of administration and governance infrastructure: "editorial policy and standards; technology requirements; communication; and training" (Pitti and Simmons 2017). Seventeen institutions from across the United States were invited to participate in this first pilot phase. A SNAC prototype is available for any user to access and determine the viability of the interface.

[1] SNAC Prototype page: http://snac-web.iath.virginia.edu/

In SNAC, a record for a creator (corporate body, person or family that authored/compiled the works that are part of their archival collection) is defined as an "identity constellation." If one performs a basic search in the prototype interface for Gabriela Mistral[2] (see Figure 1), one is able to see how information from different data sources has been aggregated and presented to the user. The identity constellation includes personal information such as birth/death dates, nationality, language, and a link to an extensive list of alternative names. There are also biographical notes credited to the institution that created the record. In this example, one can see that the University of Texas at Austin, the University of Texas Libraries, and the University of California, Los Angeles, are the institutions responsible for these biographical notes. In addition, there is an area for "Links to Collections" where we can find a list with links to other "Archival Collections" holding materials related to the Chilean writer. There is also a list under "Related Resources" with links to WorldCat records of works by or about her, as well as a list of "Related External Links" which can include the Virtual International Authority File, Wikipedia, WorldCat Identities, LC Name Authority File or Wikidata. Another area named "Related names in SNAC" allows us to record information about the people, families, and organizations that Mistral has been either associated or corresponded with. For instance, looking at these lists, we learn that she corresponded with Victoria Ocampo, and a link pointing to the Argentine writer's constellation is provided. These types of associations are what make SNAC such a powerful resource for the archival community. Among other fields, the identity constellations can include subjects, occupations, and places. All constellations are assigned a unique identifier.

[2] Gabriela Mistral identity constellation in SNAC: http://n2t.net/ark:/99166/w6mm72zn

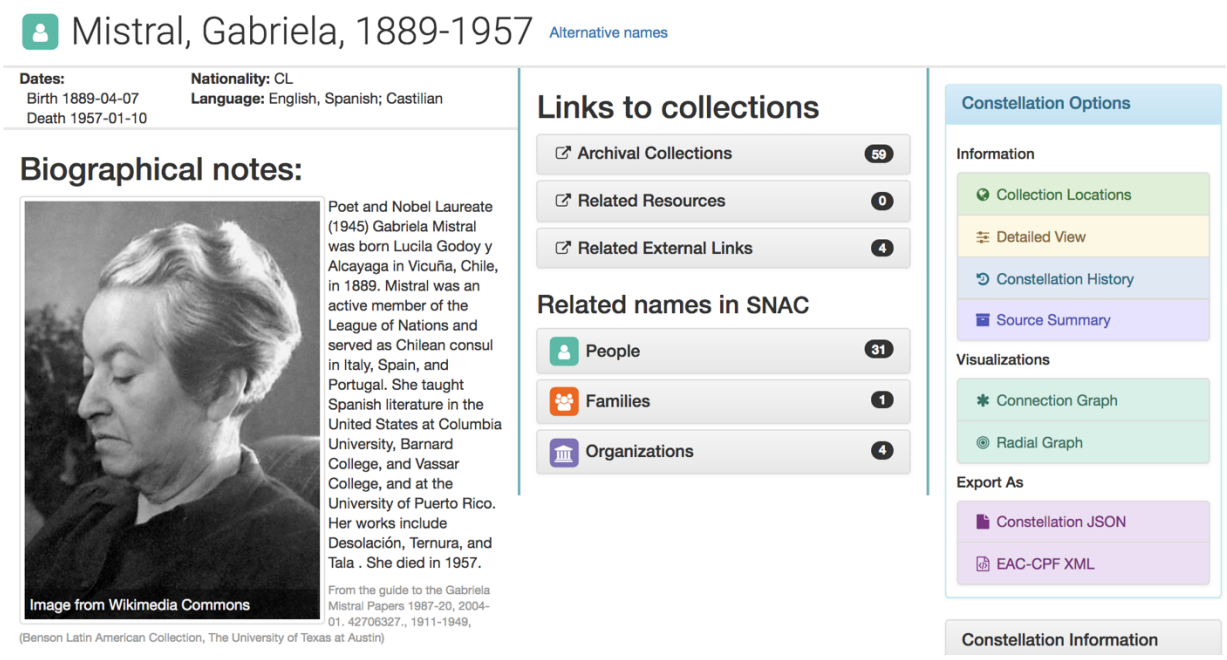

Figure 1. Screenshot of Gabriela Mistral identity constellation in the SNAC interface

A discovery interface like SNAC can be very beneficial for the Latin American archival community. Publishing metadata about collections on a single platform, regardless of where they are located geographically, greatly increases the chances for discoverability and access. It would also serve as an opportunity to analyze the data and gain a better understanding of the collections and the work that could potentially be done to enrich its quality. Working together toward a common goal—establishing the identities of persons, families, and corporate bodies—seems like a more efficient and logical way of utilizing institutional resources. In the library community, we have experienced the benefits that come with being part of a cooperative in which we create and have access to authority and bibliographic records. Being able to also collaboratively work in providing description and access to archival collections would be of benefit not only to participating institutions, but also to researchers. Such collaboration would greatly reduce duplication of effort because everyone is able to contribute, enhance, and access archival

collections in one central location. The discovery of archival collections can also be improved by sharing archival data with open knowledge projects like Wikipedia and Wikidata.

Wikipedia

In the archival community, there has been a slow but steady interest in participating in projects like Wikipedia. Wikipedia, the online encyclopedia that anyone can edit, is one of the projects under the umbrella of the Wikimedia foundation, a nonprofit organization. The encyclopedia has been in existence for 17 years and is currently ranked as the fifth most visited site on the web ("List of most popular websites" 2018). There are many efforts underway at various institutions around the world to increase participation in this project. One way in which they are approaching this is by coordinating edit-a-thons for groups of people to increase coverage in Wikipedia for a particular topic. Others have taken the time to add links to existing Wikipedia articles pointing to their institution's archival collections.

Contributing archival data to Wikipedia is fundamental in providing access to underrepresented subjects in a platform that is more widely accessed. One tool that can facilitate this task is RAMP (Remixing Archival Metadata Project).[3] This web-based editing tool was created at the University of Miami Libraries (UML) and allows users to repurpose metadata and share it on the English language version of Wikipedia. Essentially, the biographical/historical note from an EAD finding aid will become the body of the Wikipedia article. Structured metadata is added to what Wikipedia identifies as an "infobox"[4] based on information in the finding aid. Additionally, links to external sources, including the archival collection's website, are also added to the Wikipedia article. A pilot project conducted in 2014 using the Cuban

[3] RAMP prototype page: https://tools.wmflabs.org/ramp/

[4] An infobox contains structured metadata about the subject being described in a Wikipedia article. It is usually located on the right-hand side of an article for languages that read from left to right.

theater collections at UML helped increase traffic from the English language version of Wikipedia to the finding aid pages for these collections. Although increasing the visibility of and traffic to archival collections can be seen as a motivating factor at first, the real accomplishment lies in repurposing and giving archival metadata a new meaning. Additionally, it is important to understand that our work does not end once we contribute our archival data to a project like Wikipedia. Data needs to be updated and maintained. Fortunately, there is a way of minimizing some of this manual labor by also contributing archival data to Wikidata.

Wikidata

Wikidata, the newest project of the Wikimedia Foundation, is a knowledge base where structured linked data is stored. This data can be used by other Wiki projects (Wikipedia, Wikisource, Wikimedia Commons, and others) interested in having not only structured data, but also the most up-to-date information on their site. Institutions that are already making contributions to Wikipedia might also be interested in participating and contributing to Wikidata. Data contributed to this knowledge base can be used by Wikipedia to generate its infoboxes. For instance, if information about the death date of a subject in the English language version of Wikipedia is updated, we would probably want the information to also be updated in other language Wikipedias (such as the Spanish and Portuguese language versions of Wikipedia). A more concrete example would be if we wanted to update some data points for Gabriela Mistral starting with the article in the Spanish language version of Wikipedia. In this particular case, we would need to make the same changes 85 more times since there are 86 different language versions for her Wikipedia page. Instead of having to manually update this piece of information in other Wikipedias, a more efficient approach is to make the contribution directly to Wikidata.

That way, all other language Wikipedias that are connected to Wikidata would be updated automatically.

Wikidata items have their unique identifier, but they can also include identifiers from external databases or systems, thus providing access to other data sources. All of the data in Wikidata is licensed under a CC0 license, which means that it is free to use. Having free access to the knowledge base dataset has made it possible for developers to create tools and applications for multiple uses. Wikidata's data can be queried through the Wikidata Query Service,[5] where one can ask complex questions and expect to receive all information contained in Wikidata about a particular subject. For instance, we could ask for the number of writers who have won the Nobel Prize for Literature, and then have the results displayed with images, if available. We can narrow down that search by asking for female Latin American writers (Gabriela Mistral shows as the answer for that query). The results are dependent on the information that is available in Wikidata at the time.

In relation to the value of contributing our archival data to other platforms or projects, we can run a query to find out whether there were any items in Wikidata with a SNAC identifier (SNAC Ark ID). According to the results returned on June 15, 2017, there were 128,134 items in Wikidata containing SNAC identifiers. The same query was used a year later and it revealed a small increase in the use of this identifier with 128,633 items. This means that all these items point to the SNAC dataset, providing users with all information that is available in SNAC for a corresponding item, thus enhancing their experience.

[5] Wikidata Query Service: https://query.wikidata.org/

Conclusion

Having our institution's website as the sole point of access to our archival collections can be disadvantageous. By limiting ourselves to sharing information in our local system, we are missing an opportunity to reach out to a wider audience. There are tools and projects available that can facilitate creating, curating, and sharing information on a global scale, and we should take advantage of them. When we contribute our data to other projects, we have the opportunity to take a step back and see how it integrates with other data sources in emerging environments. This also gives us an opportunity to assess the need for any enhancements. Providing access to knowledge is at the core of every library institution, which is why repurposing our archival data through projects like SNAC, Wikipedia, and Wikidata is particularly appropriate. In the end, we all share the same goal of facilitating free open access to reliable information.

References

Larson, Ray R., Daniel Pitti, and Adrian Turner. "SNAC: The Social Networks and Archival Context project – Towards an archival authority cooperative." *IEEE/ACM Joint Conference on Digital Libraries*, 2014. doi:10.1109/jcdl.2014.6970208.

"List of most popular websites." *Wikipedia*, 2018. Accessed June 20, 2018. https://en.wikipedia.org/wiki/List_of_most_popular_websites

Pitti, Daniel V. 2006. "Technology and the transformation of archival description." *Journal of Archival Organization* 3, no. 2-3: 9-22.

Pitti, Daniel V., and Jerry Simmons. "Social Networks and Archival Context: In Transition from Project to Program." Project briefing, Coalition of Network Information Meeting, Albuquerque, NM, April 4, 2017.