Social Networks and Archival Context Project

Archival Authority Control

Daniel V. Pitti
University of Virginia

Summary

Social Networks and Archival Context (SNAC) is a four-year research and demonstration project funded by the U.S. National Endowment for the Humanities and the Andrew W. Mellon Foundation. The objective of the project is to explore the feasibility of extracting the names of corporate bodies, persons, and families and related historical data found in archival record descriptions, assemble this data into archival authority descriptions, and use these authority descriptions to provide researchers union or integrated access to dispersed archival holdings *and* the socio-historical contexts of the records.

Introduction

Social Networks and Archival Context (SNAC) is a research and demonstration project which has as its objective serving both the professionals who are responsible for the care of archival records, and the diverse research communities that use these records.[1] Currently, the names, social-professional relations, and biographical-historical description that constitute archival context lie buried in finding aids and other archival descriptions that are themselves available only in dispersed systems. SNAC's core activities are to extract those names, social-professional and resource relations, and biographical-historical data from dispersed archival descriptions; reassemble the data in standard archival authority descriptions; and use these authority descriptions to build a prototype publicly accessible system that will provide integrated or union access to distributed primary resources and, at the same time, access to biographical-historical information that will enable users to identify and learn about persons, families, and organizations, their histories, and the social networks in which they lived and worked.

Participants, Funding, and Timeline

SNAC is a collaboration of the Institute for Advanced Technology in the Humanities, University of Virginia (lead); the School of Information, University of California, Berkeley; and the California Digital Library, University of California. The National Endowment for the Humanities (U.S.) funded the first two-year phase of the project, May 2010 to April 2012. The Andrew W. Mellon Foundation is funding the second two-year phase of the project, May 2012 to April 2014.

Centrality of context in archival description

At the heart of access to archival records is the finding aid. A finding aid is a written description of an individual archival collection and it enables users to discover, locate, identify, and understand records. It is usually a hierarchal description of records that have a common origin or provenance. This group of records is known as a *fonds* or collection. The finding aid begins with a description of all of the records in the collection, and then provides a description and analysis of the components in the records (generally categorized by the activity or function they document, such as correspondence or minutes of meetings). Components of the components are then similarly described, until the hierarchy terminates at the description of a file or item. The depth of analysis is determined by both intellectual and economic criteria, but for a large collection it can be hundreds of pages long.

It also includes a description of the provenance or context within which the records were created. Establishing and describing the provenance of a collection involves identifying the creator by name, describing the creator's essential functions, activities, and characteristics, and the dates and places the creator was active. Thus, in addition to description of records, finding aids typically provide the name of the creator of the records and biographical/historical information. Providing a description of the context for the origination of the records enables users to understand and interpret them: without such context, many (perhaps most) records would be difficult, if not impossible, to understand and interpret. Archivists consider the provision of contextual creator information essential in the documentation and use of records.[2]

In the process of describing records, archivists also situate them more broadly in the historical and social context within which the creator existed. The specific contexts with which record creators exist are reflected in the records created or accumulated by them. Archivists document this broader context in the finding aid either through formal references to other corporate bodies, persons, and families or, less formally, in the description of the records themselves. Letters and other communications are particularly valuable as evidence of the social contexts within which creators lived and worked. In addition to describing the record creators, finding aids frequently contain the names of people, corporate bodies, and families that are connected in some manner to the creator. Finding aids are thus an excellent documentary source of information on the professional and social networks within which record creators were active.

Separating creator and record descriptions

Following the release of Encoded Archival Description (EAD) in 1998, many archival repositories around the world began to convert traditional print finding aids into machine-readable form. Based on the International Council on Archive's (ICA) *General International Standard Archival Description (ISAD(G)),* EAD is a standard for the computer representation and communication of archival description.

Both ISAD(G) and EAD reflect the traditional archival descriptive practice of combining record and creator description in a single apparatus. However, as archival and library practices have begun to take advantage of digital tools, it has become clear that this intermixing of record and context description is intellectually and functionally restrictive. In 1996, two years before the release of EAD, ICA released the first version of *International Standard Archival Authority Records–Corporate Bodies, Persons, and Families (ISAAR (CPF)).* One of its principal objectives was enabling the separation of record and creator description. Each type of description, though interrelated, would be created and maintained separately: the final archival description would combine the two at the time of use to form a complete finding aid.

There are several interrelated intellectual and practical rationales for this approach that are based on archival processing efficiencies, the intellectual quality and depth of resource description, and enhanced access to primary humanities resources for all users.[3]

Authority Control. An important benefit of describing creators in a separate dedicated apparatus is the imposition of authority control on the forms of the names used to represent named entities. Archival authority control has the same function as library authority control. For a single person, corporate body, or family, it provides preferred entries, typically based on the name most commonly used (e.g., Bill Clinton) and notes other entries based on known alternative names (e.g., William J. Clinton and William Jefferson Clinton). Providing alternative name entries assists in leading users to the preferred name when the user knows the entity primarily or only by an alternative name.

Flexible Description. While repositories commonly use a single finding aid to describe all records created or accumulated by the same creator (that is, all records with a common provenance), many repositories are shifting to the "series system," first advocated and used in the Australian National Archive and increasingly used around the world.[4] In many modern government bureaucracies, responsibilities and functions are frequently reassigned over time to different departments and agencies. For example, the function of "immigration control" might pass from one agency to another, as it recently did in a major government reorganization in the United States. This distribution complicates both managing and using such records. The series system advocates instead describing "function-based" series, which maintains the integrity of each series. It also maintains provenance, by linking the series description to the various agencies that have been successively responsible for carrying out the function. The separation of record and creator descriptions is practically and economically essential in the series system.

Similarly, dispersed collections benefit from associating one creator description with two or more record descriptions. Dispersed collections are relatively common for prominent individuals and families when disposition of records involve financial interests. Walt Whitman's papers, for example, are distributed among more than seventy repositories.[5] Repeating biographical or historical information in two or more finding aids is an unnecessary and avoidable duplication of effort if a single creator description can be shared among multiple record descriptions. Such sharing would be possible in a cooperative authority control environment.

Cooperative Authority Control. Separate creation and maintenance of bibliographic description and authority control records has long been the practice of the library community, and makes it possible to share the creation, maintenance, and use of authority data among libraries. For example, the Name Authority Cooperative Program (NACO) is devoted to the international collaborative creation and maintenance of the Library of Congress Name Authority Files (LCNAF), which now contains over five million records for persons and corporate bodies. While there are intellectual rationales for cooperative authority control, the primary incentive is economic. Authority control is labor-intensive, and sharing the work improves catalogers' productivity.

Archival cooperative authority control has a similar strong economic rationale. Records are created in social contexts, which are represented in archival descriptions. For example, the creator of one collection may be the correspondent in another and perhaps a research collaborator in still another. The same person, corporate body, or family names may thus appear in different archival descriptions in different roles. Sharing descriptions of creators in dispersed collections saves time and labor, and distributing the work of creating and maintaining authority records can provide significant economic benefits for the archival community.

Integrated Access to Cultural Heritage. Over the course of the last decade, archives, museums, monuments, and historical sites have increasingly joined libraries in providing intellectual Internet access to their holdings. During this same period, cultural heritage institutions of all kinds have begun providing digital surrogates and representations of traditional media and other cultural objects, as well as original digital resources. While these activities are providing unprecedented and dramatically improved access to our cultural heritage, further improvement is impeded by differences in both descriptive practices and the many sites and systems making holdings available.

There are increasing efforts to integrate access to cultural heritage. To date, most of these efforts have focused on reconciling or ameliorating differences in the descriptive practices of the archive, library, museum, and other cultural heritage communities, in order to build a single catalog of resources. As important as such efforts are, reconciling resource description across these communities presents intellectual, technical, and political challenges that will require significant investments of time and effort to address.

Rather than attempting to reconcile descriptive practices, by linking resource descriptions (irrespective of type) to the descriptions of persons, corporate bodies, and families, the authority records can serve to integrate access to resources by and about the described entity. Further, an archival authority record can provide not only contextual information for understanding records, but also access to and context for understanding art, buildings, novels, scientific reports, poetry, business records, music, and everything else that constitutes the human record, our cultural heritage.

Biographical/Historical Resource. Like library authority control, archival authority control involves the selection and formulation of preferred and alternative name entries, and identification of related entities. However, archival authority records go on to describe the context in which archival resources were created, which is essential for understanding and interpretation. Context description is provided in the form of biographical/historical data about the creator such as dates and places of existence, significant activities and functions performed by the entity, and other significant dates, places, and events. Context description also frequently includes chronological lists detailing significant dates, places, and events, or prose biographies or histories.

This biographical/historical detail extends the utility of archival authority records beyond providing context for archival records, as central as this is to archival description. It can be used as an independent resource that can assist users in identifying and learning about the described entity.

Social/Historical/Intellectual Context. Similar to library practice, archival authority control identifies other entities related to the person, corporate body, or family described in a record. Archival practice, though, has the potential to cover a significantly much broader range of types of relations. Persons, corporate bodies, and families all create and

> accumulate records in a social context. People live and work with other people, as individuals and as members of families and organizations. People learned and are influenced by the work. These social, professional, and intellectual relations are reflected in records and consequently in the descriptions of the records. Letters and other communications among individuals, families, and organizations, in particular, are important evidence of social relations. In describing correspondence and other records in the finding aid, archivists often reference, by name, the related individuals, corporate bodies, and families documented in the records, listing those judged to be most significant via controlled entries, and informally, in the description of the correspondence itself in file or item titles.
>
> While information documenting these social and professional relations is currently available in finding aids, it is isolated. Users must painstakingly analyze and piece together the relations by manually compiling lists of names from one finding aid and then searching and analyzing other finding aids and catalogs. Archival authority control records provide a means to systematically gather and document these social and professional relations in links that interrelate descriptions of people, organizations, and families. This documentation can provide convenient access to the broad social-historical contexts within which corporate bodies, persons, and families were active, and convenient, navigable access to related or complementary resources.

There are thus many compelling benefits to archival authority control for both archivists and users of primary humanities resources. There are economic benefits for archivists in cooperative authority control and more efficient and effective description of complex and dispersed records. Both archivists and users benefit from the enhanced access and understanding provided by alternative names; integrated access to the entire spectrum of cultural heritage resources; and biographical/historical information about persons, families, and organizations, including the broader social-historical context within which they were active.

EAC-CPF: archival authority records

With the release of EAC-CPF in March 2010, archivists now have an international communication standard for realizing the objectives and benefits of archival authority records. EAC-CPF is based on the second edition of ICA's *International Standard Archival Authority Records–Corporate Bodies, Persons, and Families (ISAAR(CPF))*, 2004.[6] The Society of American Archivist's (SAA) Technical Subcommittee-Encoded Archival Context (TS-EAC) is responsible for the development and maintenance of EAC-CPF. The members of the TS-EAC are representative of the international archival community, with three members also serving on complementary ICA standards committees.

While there has been significant experimental use of EAC *alpha (2001)* and *beta* (2004) in projects, the archival community has been waiting for an official stable version of the standard before embarking on programmatic use.[7] With the release of EAC-CPF, as the standard was renamed, the archival community is now poised to take the next major step in transforming and enhancing archival description—separating creation context description and control of all corporate body, person, and family names from record description. A major challenge at this point is extracting the creation context and related names from EAD-encoded findings aids and assembling the extracted data into EAC-CPF-encoded records where they can be independently maintained and used to the benefit both the professional archival community and the users of archival records.

Research Activities

In the first phase of SNAC, the project focused on extracting and assembling EAC-CPF authority records from approximately 30,500 finding aids and augmenting the derived authority records with additional data from library and museum authority records. In the second phase, the number of finding aids has been increased to more than 150,000, and has been augmented by 2,022,450 MARC collection-level archival descriptions contributed by OCLC WorldCat. The WorldCat collection-level descriptions provide only a brief, top-level description of a collection, and not the detailed hierarchical description found in finding aids. The use of MARC collection-level description has been almost exclusively restricted to U.S. repositories. Though not as extensive as finding aids, collection-level descriptions nevertheless typically contain the name of the creator of the collection and frequently include a brief biographical-historical description of the creator, occupation, and the names of other named entities with whom the creator is most prominently related. Because creating collection-level descriptions was common in the U.S., the more than two million MARC descriptions provide comprehensive national coverage of (minimally or fully) processed archival holdings. In addition, the National Archives and Records Administration (NARA), Smithsonian Institution, British Library (BL), and the New York State Archives will contribute over 375,000 original archival authority records in a variety of formats that will be converted into EAC-CPF. The Archives nationales (France), and Bibliothèque nationale de France (BnF) will contribute a small number of EAC-CPF records that will support experimenting with methods for working with a multilingual environment. The derived and original archival authority records will be augmented with additional data from library and museum authority records: 16 million Virtual International Authority File (VIAF) cluster records; and 120,000 Union List of Artist Names (ULAN) records.

SNAC processes the source data and creates the archival authority records that are the content of the prototype public historical resource and access system in three steps, each step being the responsibility of one of the three project partners. IATH is responsible for acquiring and managing all of the data from the contributing institutions. IATH is also responsible for extracting archival authority data from the contributed EAD-encoded finding aids and MARC cataloging records and assembling them into Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF) records, as well as converting original archival authority records into EAC-CPF. It is estimated that between 1.5 and 4 million EAC-CPF descriptions will be produced in the processing of the more than 2.5 million finding aids, MARC records, and original authority records. IATH also transforms contributed archival authority descriptions that are received in an alternative format. Once the finding aids have been processed, the result is a set of EAC-CPF records. Each contains a single identified name, along with identification of the source finding aid or catalog record, and, in the case of creators, any biographical information, dates of existence, language or languages used, links to related people, etc. that are found in the source. Since EAC-CPF descriptions are derived independently from each EAD or MARC description, there may be multiple EAC-CPF instances representing the same entity. A key challenge, then, is to identify multiple EAC-CPF descriptions that represent the same entity and combine them into a single description.

The second step in the processing is the responsibility of the University of California, Berkeley (SI/UCB). This involves two activities. First, the EAC-CPF instances created or acquired in the first step are matched against one another. Records identified as matching are combined into a single record, which retains links to the EAD or MARC descriptions and to other EAC-CPF entities. This accumulating of links provides integrated access to the primary resources and continues the process of interconnecting people to build the social-professional networks. Next, the resulting EAC-CPF instances are matched against library and museum authority records in VIAF, ULAN, and LCNAF. Alternative names used by or for the entity and additional *non-duplicating* descriptive data (sex, country or countries of affiliation, and languages used) are added to the EAC-CPF instances. Additional biographical or historical description is added from matching ULAN records. The resulting set of EAC-CPF records is the foundation for the next step in the processing, the prototype public historical resource and access system.

The California Digital Library (University of California) (CDL) is responsible for the third and final step in the processing. Using XTF (Extensible Text Framework), an open-source XML publishing system, CDL is developing a sophisticated public research tool that at once serves as a historical resource and provides integrated access to the distributed archival resources whose descriptions provide the primary data for the project. The archival authority descriptions are indexed, to provide searchable access to the individual records. The searching is faceted, enabling users to qualify searches by occupations and subject headings used in describing records created by them. Individual records provide information (when available) on dates of existence, sex, occupations, languages used, subjects reflected in related primary resources, affiliated country or countries, and biographical-historical description (prose or a chronological list of major life events). Lists of links to all related primary and secondary resources are provided, as are all links to related persons, corporate bodies, and families. The latter social-professional relations may also be explored using a graph. (See Appendix Two for an example of a social-network graph.) Additional features, in particular a timeline-map display of biographical-historical information, and searching of the social graphs will be provided. During the development process, CDL will work to establish use cases that depict how the system will be used, as well as conduct face-to-face scholar and educator user testing to evaluate the usability and usefulness of functions and features.

In addition to the broad description of the processing steps above, there are also a number of other processing activities involved. Though the most prominent names of people documented in archival records will be found in the description of the records, explicitly tagged by archivists as names, there are many names that occur, in particular in the description of correspondence, that are not explicitly identified as names. In the extraction processing, National Language Process (NLP)/Name Entity Recognition (NER) techniques will be employed to identify the names of correspondents. Given the qualitative diversity of the source data, many found names are not "well-formed" (for example, personal names may be in direct natural language order, rather than the inverted order used in resource description). Researching and developing techniques for improving the quality of names found is an important focus. Many of the dates are given in natural language forms, and thus techniques for normalizing dates will be employed. Further, geographic names found in biographical-historical chronologies will be identified using NER techniques, normalized, and coordinate data added. The normalization of dates and geographic names will support developing timeline-map displays of lives.

Searching and displaying the social-professional networks and organizational hierarchies assembled in SNAC will also present the project with an important area of research. Effectively searching social-professional networks will enable researchers to identify relations and influences, even generational influences that might otherwise be overlooked in the current research environment. Displaying networked information, particularly when the graph data is dense, as it is anticipated to be for certain individuals, presents design challenges. The objective will be to design displays that enable users to "browse" social networks, as well as the networks of resources interrelated to the people. The inclusion of agency histories (primary NARA, Smithsonian Archives, and New York State Archives) will present the project with yet another important area of graphical research, namely the graphical display of organization hierarchies as well as sequential graphical display of when two or more entities merge or one entity splits into two or more entities.

Phase I Results Overview

In the first phase of the SNAC project, the extraction and assembling of authority records produced the following results in processing approximately 30,500 finding aids contributed by the Library of Congress and three archival access consortia.

- Library of Congress: 43,702 authority records derived from 1,159 finding aids
- Online Archive of California: 91,811 authority records derived from approximately 15,400 finding aids
- Northwest Digital Archive: 22,609 authority records derived from 5,160 finding aids
- Virginia Heritage: 15,175 authority records derived from 8,390 finding aids
- Total authority records derived: 173,297

The 173,297 authority records were then matched against one another and matches merged (or combined), and the resulting set was then matched against the Virtual International Authority File (VIAF). This processing resulted in 128,783 “unique” names.

The prototype public system developed in the first phase of the project has demonstrated that the archival authority descriptions that have been derived from EAD-encoded and MARC record descriptions can effectively be used to provide integrated access to the dispersed archival records (as well as other cultural heritage resources) and, at the same time, serve as socio-historical resources where users can learn about individuals, families, and corporate entities.

Each description of a corporate body, person, or family serves as node in a social-document network, providing links to the descriptions of records created by or referencing the described entity. In addition each node is also linked to resources in WorldCat, most of which are published works by the entity. Links are also provided to corporate bodies, persons, and families that are associated with the described entity, with many of the linked entities identified as correspondents. Access to these social, professional, or intellectual networks is provided through lists, a dynamic radial graph, and a SPARQL end-point that will enable sophisticated querying of the relations.[8]

Conclusion

The SNAC project has successfully demonstrated that the names, descriptions, and interrelations of individuals, corporate entities, and families documented in archival record descriptions can be separated from the record descriptions, assembled into authority descriptions, augmented with additional data from library and museum authority records, and used to create a research tool that integrates access to distributed archival records *and* serves as a socio-historical resource. Despite the promising early results, much research and development remains in order to improve the quality of the processing results. The uneven quality of the archival descriptions presents many challenges. Many of the names found are not well-formed. There are inconsistencies in the order of personal name components. Many names are not correctly assigned a type. Punctuation in both personal and corporate names varies, as does the abbreviation of words in corporate bodies. These and many other issues of quality present both extraction and matching challenges.[9]

The second phase of SNAC will vastly expand the source data as well as the research agenda. The expanded set of source data will vastly increase the quantity of EAC-CPF records created, the geographic scope of the holdings represented, and the density of the social network graphs. Employing Name Entity Recognition techniques, if successful, will also increase the number of records extracted.

The project collaborators will continue to improve the methods and techniques used in extracting, assembling, and matching archival authority records, but it is clear that algorithms alone will not be sufficient to build an acceptably accurate and detailed body of archival authorities data that can used to build a reliable public access and resource. While the immediate objectives of the project are to significantly refine and improve the effectiveness of the methods used in building an innovative research Internet-accessible tool, the long-term objective is to provide both methods and data as a solid foundation for establishing a sustainable national archival authorities program cooperatively governed and maintained by the professional archive and library communities. As envisioned, such a cooperative program would employ computational methods of building and maintaining the data, augmented by archivists, librarians, scholars, and, perhaps, public users that assist by adding additional archival descriptions, enhancing and improving the accuracy and historical detail of existing descriptions, interrelating described entities, and identity resolution and management. [10]

[1] For additional information on SNAC, see http://socialarchive.iath.virginia.edu/.

[2] Adrian Cunningham. "Harnessing the Power of Provenance in Archival Description: An Australian Perspective on the Development of the Second Edition of ISAAR(CPF)" in *Journal of Archival Organization* (New York: Haworth), Vol. 5, Nos. 1/2 2007.

[3] Peter Scott, "The Record Group Concept: A Case for Abandonment" *American Archivist* 29:493-504 (October 1966). In the 1980s, several articles appeared in support of separation: Richard H. Lytle, "Intellectual Access to Archives," *American Archivist* 43 (Winter and Spring 1980); Lytle and David A. Bearman, "The Power of the Principle of Provenance," *Archivaria* 21 (Winter 1985-1986); Max J. Evans, "Authority Control: An Alternative to the Record Group Concept," *American Archivist* 50 (1986): 240-261; Bearman and Richard Szary, "Beyond Authorized Headings, Authorities as Reference Files in a Multi-disciplinary Setting" in "Authority Control Symposium." *Occasional Papers of the Art Library Society of North America*, no. 6 (Tucson: Art Library Society of North America, 1987).

[4] See Scott (1966) and more recently, Cunningham (2007).

[5] Ken Price and Ed Folsom *Re-Scripting Walt Whitman: An Introduction to His Life and Work*: http://whitmanarchive.org/criticism/current/anc.00152.html#app.

[6] *ISAAR(CPF)* 2nd edition: http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf

[7] The most significant project using EAC was Linking and Exploring Authority Files (LEAF), a project funded by the European Union. See Max Kaiser, Hans-Jörg Lieder, Kurt Majcen, and Heribert Vallant, "New Ways of Sharing and Using Authority Information: The LEAF Project" (D-Lib Magazine, November 2003: http://www.dlib.org/dlib/november03/lieder/11lieder.html).

[8] The SNAC prototype public system is available at http://socialarchive.iath.virginia.edu/xtf/search.

[9] It should be noted that this is not a criticism of creators of the archival descriptions. With limited resources, many archives have emphasized making as many holdings as possible accessible, rather than providing "perfect" access to a small subset the holdings. Further, the archivists and librarians never anticipated the descriptions serving the objectives of the SNAC processing.

[10] Developing a "blueprint" for establishing a sustainable national archival authorities cooperative will be the objective a project funded by the Institute of Museum and Library Services that will take place in a parallel to SNAC.