The Social Networks and Archival Context Project

3. NARRATIVE

3.1. SIGNIFICANCE

3.1.1. INTRODUCTION

The *Social Networks and Archival Context Project* (SNAC) will address the ongoing challenge of transforming description of and improving access to primary humanities resources through the use of advanced technologies. The project will test the feasibility of using existing archival descriptions in new ways, in order to enhance access and understanding of cultural resources in archives, libraries, and museums.

Archivists have a long history of describing the people who—acting individually, in families, or in formally organized groups—create and collect primary sources. They research and describe the artists, political leaders, scientists, government agencies, soldiers, universities, businesses, families, and others who create and are represented in the items that are now part of our shared cultural legacy. However, because archivists have traditionally described records and their creators together, this information is tied to specific resources and institutions. Currently there is no system in place that aggregates and interrelates those descriptions.

Leveraging the new standard Encoded Archival Context–Corporate Bodies, Persons, and Families (EAC-CPF), the SNAC Project will use digital technology to "unlock" descriptions of people from descriptions of their records and link them together in exciting new ways. First, it will create an efficient open-source tool that allows archivists to separate the process of describing people from that of records, meaning that it will pave the way for improving the quality of description and the quantity of resources described. And it will create an integrated portal to creator descriptions—linked to resource descriptions in archives, libraries and museums, online biographical and historical databases, and other diverse resources—thereby providing more effective access and robust historical context to a broad array of humanities materials.

At the core of the project are the following goals:

- Support scholars and other users in discovering and identifying persons, families, and organizations by making all of the names used by and for them searchable.
- Enhance access to primary resources by linking descriptions of people to a wide variety of resources by and about them.
- Provide access to the social and profession networks within which people live and work by systematically documenting their relationships with one another.
- Provide archives with an efficient and reliable means to exploit and transform archival description, thereby improving description of archival records and re-using metadata across repositories.

The prototype access system will demonstrate that descriptions of persons, families, and organizations can be used as access points to archive, library, and museum resources. SNAC will thus benefit the humanities community most broadly—scholars, educators, students, and anyone interested in the record of our past.

The Institute for Advanced Technology in the Humanities at the University of Virginia will be the lead institution for the SNAC Project, and will partner with two University of California Institutions: the California Digital Library and the School of Information, University of California, Berkeley. The project will derive EAC-CPF records from existing archival finding aids from the Library of Congress, the Online Archive of California, the Northwest Digital Archive, and Virginia Heritage; and also from name authority files supplied by the Library of Congress, Getty Vocabulary Program, and OCLC Research.

3.1.2. BACKGROUND

Centrality of context in archival description

At the heart of access to archival records is the finding aid,[1] a description of an individual archival collection. The finding aid enables users to discover, locate, identify, and understand records. Usually, it is a hierarchal description of records that have a common origin or provenance. It begins with a description of all of the records in the collection, and then provides a description and analysis of the collection's components (generally categorized and grouped 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 and detail of analysis is determined by both intellectual and economic criteria, but the finding aid for a large collection can be hundreds of pages long.

In addition to the description of the records comprising a collection, archivists also describe in the finding aid 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 when and where the creator was active. Providing a description of the context in which the records originated helps users to understand and interpret them. Without such context, many, if not most, records would be unintelligible. Therefore, in addition to record descriptions, finding aids typically provide the name of the collection's creator and biographical/historical information. Archivists consider the provision of contextual creator information essential for the documentation and use of records.[2] See Appendix A, "Example of creator and biographical/historical data."

In the process of describing records, archivists also situate them more broadly in the creator's historical and social contexts. These contexts are reflected in the records that the creator created or accumulated. 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. Finding aids thus contain the names of people, corporate bodies, and families that are connected in some manner to the creator, which makes them an excellent documentary source of information on the professional and social networks within which record creators were active. See Appendix B, "Example of controlled access and series description."

Separating creator and record descriptions

With the release of Encoded Archival Description (EAD) in 1998, many archival repositories around the world began to convert print finding aids (see Appendix C for an example) into machine-readable form. Based on the International Council on Archive's (ICA) *General International Standard Archival Description (ISAD(G)),* EAD has become a standard for 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 using digital tools, the possibility of using more effective and efficient methods has become evident. In 1996, two years before the release of EAD, ICA released the first version of *International Standard Archival Authority Records–Corporate*

---

[1] Archival records are the evidence of people, acting individually, in families, or in formally organized and named groups. Records are the byproducts of people living their lives or carrying out official duties or responsibilities and may include items such as personal letters and notes from meetings.

[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.

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, based on archival processing efficiencies, the intellectual quality and depth of resource description, and enhanced access to primary humanities resources for all users.

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 determines the preferred name for that entity, 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). The user may know the entity primarily or only by an alternative name, and alternative name entries lead the user to the preferred 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.[3] In many modern government bureaucracies, responsibilities and functions are regularly reallocated among different departments and agencies. For example, the function of "immigration control" can pass from one agency to another, as in a recent major U.S. government reorganization, leading to records that are distributed among several creators. This distribution complicates both managing and using such records. Instead, the series system advocates describing records linked by a common function into a series, thereby maintaining 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.[4] In these cases, repeating biographical or historical information in two or more finding aids is an unnecessary and avoidable duplication of effort; a single creator description could be shared. This 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. For example, the Name Authority Cooperative Program 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; it is labor-intensive, and sharing the creation, maintenance, and use of authority data improves catalogers' productivity.

Archival cooperative authority control has a similar economic rationale. Archival records are created in social contexts, which are in turn represented in archival descriptions. For example, the creator of one collection may be a correspondent in another and perhaps a research collaborator in a third. The same person, corporate body, or family names can appear in different archival descriptions in multiple roles.

---

[3] See Peter Scott, "The Record Group Concept: A Case for Abandonment" *American Archivist* 29:493-504 (October 1966), and more recently, Cunningham.

[4] 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.

Sharing descriptions of creators in dispersed collections saves time and labor, and distributing the work 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 offering various levels of online 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 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, but 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 involves intellectual, technical, and political challenges that will require a significant investment of time and effort.

In the meantime, descriptions of persons, corporate bodies, and families can be leveraged to provide a means of locating a particular entity, via links to existing resource descriptions and dynamically generated searches of archive, library, and museum resource description systems. An archival authority record can provide not just contextual information for understanding records, but also access to and context for understanding art, buildings, novels, scientific reports, poetry, business records, music, and all that constitutes the human record and 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 the identification of related entities. However, archival authority records go on to describe the context in which archival resources were created, in the form of biographical/historical data about the creator, such as when and where the creator existed, significant activities and functions performed by the creator, and other significant dates, places, and events (see Appendix J for sample library and archival authority records). Context description also frequently includes chronological lists detailing significant dates, places, and events, or prose discourse (see Appendix D for examples of a prose biography and a chronological list). This biographical/historical detail extends the utility of archival authority records beyond just providing context for archival records, 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, and to provide historical context for understanding not only archival records but other cultural resources as well.

Social/Historical Context. Also similar to library practice, archival authority control identifies other entities related to the people and groups described. But archival practice has the potential to cover a much broader range of relations. As noted above, persons, corporate bodies, and families all create and accumulate records in a social context. People live and work with other people, both as individuals and as members of families and organizations. These social and professional 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 and professional relations. In describing correspondence and other records in the finding aid, archivists often document the relations reflected in them both formally, listing those judged to be most significant via controlled entries, and informally, in the description of the correspondence itself by using the names of correspondence in file or item titles (see Appendix B for examples).

Information documenting these social and professional relations can be found in existing finding aids, but as of right now 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. However, archival authority control records provide a potential 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.

Thus, there are many compelling benefits to archival authority control, for both archivists and users of primary humanities resources. These include: an economic incentive for institutions, which could more efficiently and effectively describe complex and dispersed records; enhanced access and understanding through the use of 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

Archivists will soon have an international communication standard for realizing the objectives and benefits of archival authority records with the release of the Encoded Archival Context–Corporate Bodies, Persons, and Families standard (EAC-CPF) in the fall of 2009. EAC-CPF is based on the second edition of ICA's *International Standard Archival Authority Records–Corporate Bodies, Persons, and Families (ISAAR(CPF))*, 2004.[5] The Society of American Archivist's (SAA) Encoded Archival Context Working Group (EACWG) is responsible for the development and maintenance of EAC-CPF. The members of the EACWG are representative of the international archival community, with two members also serving on complementary ICA standards committees (see Appendix H for a list of EACWG members).

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.[6] With the upcoming release, the archival community is poised to take the next major step in transforming and enhancing archival description—separating the description of creation context and corporate body, person, and family names from the description of records. A major challenge at this point is extracting the creation context and related names from EAD-encoded findings aids and migrating them into EAC-CPF-encoded records, where they can be independently maintained and used to the benefit of the humanities community. SNAC will create open-source software to address this challenge, and will create an open-source prototype access system demonstrating the effectiveness of using archival authority control to serve a wide variety of access and research objectives.

3.1.3. THE SNAC PROJECT

Using authority records to transform archival research

The SNAC Project will develop open-source software that will facilitate efficiently and accurately deriving EAC-CPF records from existing archival finding aids, and then enhancing them by matching them against records in the Library of Congress Name Authority File (LCNAF) and the Getty Vocabulary Program Union Lists of Artist's Names (ULAN). The software will be developed and tested using EAD-encoded finding aids from the Library of Congress (LoC) and three consortia, the Online Archive of California (OAC), the Northwest Digital Archive (NWDA), and Virginia Heritage (VH). These three consortia and the Library of Congress collectively have over 28,000 finding aids contributed by nearly 300 repositories located in Alaska, California, Idaho, Montana, Oregon, Virginia, Washington (State), and Washington, D.C. See Appendix E for a brief description of the finding aids in the Library of Congress and the three consortia.

---

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

[6] 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).

How it will work

The extraction and migration of the records will take place in three steps. First, EAC-CPF creator records will be derived from two descriptive components in EAD-encoded finding aids: 1) the name given as the creator of the described archival collection and 2) biographical/historical information describing the creator. Additional EAC-CPF records will be derived from the finding aids for related entities from two major descriptive components: 1) controlled entries and 2) identifiable names in the titles of correspondence description. Extraction and migration of creator names, biographical/historical data, and names from controlled entries are relatively straightforward processes, given the precise encoding of the EAD finding aids. Identifying and extracting names from the titles of correspondence description presents challenges because names and other data are intermixed. There have been significant developments in techniques for identifying names in "free text," and these techniques will be used for this extraction.

The second step will be to match the derived EAC-CPF records against one another to identify and merge duplicates. Because names for entities may not match exactly or the same name string may be used for more than one entity, contextual information from the finding aids will be used to evaluate the probability that closely and exactly matching strings designate the same entity.[7] For matches that have a high degree of probability, the EAC-CPF records will be merged, retaining variations in the name entries where these occur and retaining links to the finding aids from which the name or name variant was derived.

The third and final step will be to match the resulting set of EAC-CPF records against authority records in the Library of Congress Name Authority File (LCNAF) and the Getty Vocabulary Program Union List of Artists' Names (ULAN). The technique of using contextual information described above will be used to enhance the accuracy of the matching. Matched LCNAF or ULAN records will be merged with the finding aid-derived EAC-CPF records, with authoritative or preferred forms of names recorded for both authority record sources (a union set of alternative names will also be incorporated into the EAC-CPF records). Finally, the biographical/historical data commonly found in ULAN records will be merged with the context data in the EAC-CPF records. Records that might be for the same entity but have insufficient available contextual information for a confident match will be flagged for human review.[8]

The resulting set of interrelated EAC-CPF records will represent the creators and related entities extracted from EAD-encoded finding aids, with a subset of records enhanced with entries from matching LCNAF records and with entries and biographical/historical data from ULAN records. The EAC-CPF records will thus represent a large set of archival authority records, related to one another and to the archival records descriptions from which they were derived. This record set will then be used to build a prototype corporate body, person, and family name and biographical/historical access system.

Prototype access system

The prototype name and biographical/historical access system will provide a variety of indexed access to record creators and related entities. The primary form of access will be name searches. Name access will be based on preferred (or authorized), alternative, and related names. The use of alternative names will provide superior access to that currently available in finding aids, where only the preferred is used. Users will be able

---

[7] Using contextual information in determining that two or more records represent the same entity has been successful in matching and merging authority records in an international context. See Rick Bennett, Christina Hengel-Dittrich, Edward T. O'Neill, and Barbara B. Tillett *VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority File*: http://www.ifla.org/IV/ifla72/papers/123-Bennett-en.pdf

[8] It will be outside the scope of the project to attempt to merge records manually. Data will be collected on the percentage of records requiring human review to help in evaluating the overall success of the methods employed and in evaluating human factors in processing economies.

to qualify (or restrict) name searches by type of named entity (personal, corporate body, or family), making result lists more precise and focused. Users will also be able to search the biographical/historical data. All searching will offer keyword, phrase, and proximity searching, as well as full Boolean operations.

While authority file interfaces for library authority control is reasonably well understood, archival authority records provide more detailed description and possibly extensive entries to related persons, corporate bodies, and families and related resources. Therefore, defining the public interface to the access system will present new opportunities and challenges. The prototype access system will strive to present clear and easily understood displays of names, biographical/historical data, and related entities and resources.

Each EAC-CPF record will also contain entries and links to socially or professionally related entities described in other EAC-CPF records. The links will be reciprocal. For example, there will be a link from Robert Oppenheimer to Earnest Orlando Lawrence, and vice versa. The prototype system will display the related entity links as a list of names, as well as experiments with social graphs and visualizations that allow users to explore and navigate the social and professional relations among the described entities.

The prototype access system interface will present two types of links to related cultural heritage resources created by or about the named entity. The first type of link will be to the finding aids out of which the EAC-CPF records were derived. Thus authority records will be linked to either descriptions of records created by the named entity or to records with which they are associated. When the nature or role of the association is known, it will be displayed. For example, if the named entity is represented in an archival collection as a correspondent, this information will be provided so that the user has as much information as possible when interpreting and navigating links between named entities.

The second type of link to resources will be in the form of dynamic searches. The user will have the option of executing searches in a selection of both cultural heritage access systems and Internet resources in order to locate resources created by or about the named entity. The searches will be configured based on the type of access system. Searches of library systems that are based on the *Anglo-American Cataloging Rules, second edition* (AACR2) will use the AACR2 authorized form of the name. Searches of Wikipedia and other Internet resources will use "query expansion," a technique that uses a Boolean "or" operation to search using all forms of the name, authorized and alternative. This technique will increase recall when querying access systems or resources when the form of the name is either not controlled or the form of control is unknown. At a minimum, the prototype will link to the bibliographic catalogs of the participating institutions, as well as to WorldCat, Wikipedia, Flickr, and DBpedia.

Two research scenarios

The prototype access system will enhance the archival research process, enabling scholars to move from individuals to records and back again. For example, an historian researching the life of the physicist Robert Oppenheimer searches the prototype access system and locates his authority record. From the "Biographical Note" in the record, she learns that from 1929 to 1947, he taught physics at the University of California, Berkeley, and the California Institute of Technology, Pasadena. In the "Related resources" section of the record, she finds a link to *J. Robert Oppenheimer: A Register of His Papers* in the Library of Congress. Following this link leads to the Library of Congress finding aid for his papers. In the "Related to" section of the record, she finds an alphabetical list of people and organizations with whom Oppenheimer corresponded, with each linked to a creator authority record. She follows links to Raymond T. Birge, Felix Bloch, Richard Feynman, Earnest Orlando Lawrence, Julian Seymour Schwinger, and Leo Szilard and finds records describing each. In the "Related resources" section of each record, she finds a link to an Online Archive of California finding aid to papers held in an archive in California, and in the "Related to" section of each description, a link to Robert Oppenheimer.

The prototype will also enable researchers to discover a range of materials both created by and related to individuals, including secondary sources. For example, a college student in California reading a history of

science text comes across a reference to the physicist Richard Feynman. Searching the SNAC access system, he finds a detailed record. A biographical note provides a brief overview of Feynman's life and a set of links for additional biographical information, searches of participating library catalogs, and other Internet resources. Intrigued, he selects the option of searching the California Digital Library Melvyl Catalog, where he finds eighteen books by or about Feynman. Among these books, he finds one that further interests him, *Genius: the Life and Science of Richard Feynman* by James Gleick.

3.1.4. RESEARCH AND DEVELOPMENT BENEFITS

The SNAC Project will have significant research and development benefits. The systematic extraction and gathering of professional and social relations data will demonstrate how archival record description can be exploited to broaden and understand the context within which record creators lived and worked. Archival authority control can become an important new research tool for biographers and historians. The project will also demonstrate the utility and feasibility of authority control to the archival profession community, and provide it with open-source software to begin using an important new international archival communication standard to enhance the effectiveness and improve the economy of professional practice.

Preliminary testing and exploration of the structured data to be used in this project gives us confidence that we can successfully extract names and biographical/historical data from existing EAD-encoded finding aids and create preliminary EAC-CPF records. There are, however, research challenges, particularly with respect to identifying names that are not specifically designated as names, and matching, disambiguating, merging, interrelating, and normalizing data from diverse sources using contextual information. There has not been research across a large number of EAD-encoded finding aids from a wide range of diverse resources that has specifically focused on name use and encoding practices. The SNAC project will thus be able to provide a detailed evaluation of the relative effectiveness of methods of name extraction and matching when applied to structured archival description, as well as data documenting archival name practices and correlation with records in the LCNAF and ULAN.

The SNAC Project will collect the following quantitative data:[9]

- EAD tag use in encoding the creator name; biographical/historical description; and controlled access of related corporate body, person, and family entities
- Percentage of EAD-encoded finding aids using reliably identifiable corporate body, person, and family names in correspondence file description (as opposed to chronological or alphabetic description)
- Average number of unique names per each EAD-encoded finding that can extracted into EAC-CPF records
- Percentage of reliable matches between EAD-derived EAC-CPF creator records and LCNAF and ULAN records
- Percentage of reliable matches between EAD-derived EAC-CPF related entity records and LCNAF and ULAN records from both EAD controlled name entries and correspondence description

Such data will provide the archival community with information it will need to make strategic decisions about the use of technology and standards to economically transform and enhance archival description and access.

[9] While it will be possible to reliably gather certain types of quantitative data, other data will require human review of random samples of data. For example, matching and merging is based on the probability that two occurrences of the same or similar name strings from the same or different use contexts designate the same or different entities. Judgments of probability, whether made by humans or computer algorithms, are based on the quantity and quality of available evidence, and are subject to error.

This process requires good understanding of the technology and technological expertise needed, as well as the degree to which human review and editing is necessary to achieve an acceptable level of authority record quality and reliability.

Additionally, SNAC findings will provide information that will assist archival and information technology scholars in identifying new areas of research. The SNAC Project will produce a large test bed of EAC-CPF records that can be used by researchers to study issues of user interface design, integrated access and information retrieval, data extraction and transformation, and matching and disambiguation in the aggregation of authority control records from diverse sources, among other possibilities. The project will make its EAC-CPF data freely available. In particular the data will be made available to OCLC for two of its ongoing research projects, WorldCat Identities[10] and the Virtual International Authority File (VIAF).[11]

The Project will also carefully document challenges and issues encountered in the use of EAD and EAC-CPF. With respect to EAD, such challenges may be encoding practices or related to the semantics and structure of the EAD Schema. Encoding practice issues will be reported to the participating consortia and the Library of Congress for consideration in the ongoing revision of encoding guidelines. Semantic and structural issues associated with the EAD and EAC-CPF Schema will be reported to the EAD and EAC Working Groups for consideration in ongoing revision of the two standards. Particular attention will be given to recommendations for economically feasible enhancements to encoding practice that can enhance the quality and quantity of derived authority data, and revision of the standards that will increase their utility in realizing archival description and access objectives.

3.2. Background Of Applicant

Founded in 1992, the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia is one of the world's leaders in transforming humanities research through the application of computing and network technologies. The Institute hosts over forty projects and reports to the Provost's Office. The Institute is funded by the University of Virginia, with additional funding from grants and gifts. Since the Institute was founded, it has been awarded over $12 million in grants.

The Institute's projects are primarily faculty-driven, though many projects involve digital library research and development in collaboration with the Library and other institutions around the world. With an abiding commitment to long-term preservation and access to humanities collections and research, the Institute is dedicated to the promotion and use of international archive, museum, library, and humanities standards and to the use of open source software. Daniel Pitti, associate director, has extensive experience in the collaborative development of international archive and library standards, including the two principal archival standards to be used in the current project. Pitti also has extensive experience assisting in the establishing of archival access consortia, both in North America and in Europe.

The Institute has twelve staff (three vacancies). Eight of the staff are programmers with a range of different expertise: database technologies (SQL, PHP, Ruby on Rails); markup technologies (Extensible Markup Language (XML) and related languages); network and systems administration; Web standards and design; Geographic Information Systems (GIS); and graphics, audio-video, pictorial, and 3-D representation. In addition to programming staff, the Institute's associate directors have extensive experience in collaborative project design and implementation.

The Institute has four servers, one Dell 2950 (24GB RAM) and three Sun Fire X4450 (one 32 GB RAM and two 48 GB RAM) and a total of 8.5 TB of storage. The Operating System (OS) on all servers is CentOS 5 (64 bit), an Enterprise class Linux OS. The University of Virginia's Information Technology and

[10] WorldCat Identities: http://www.oclc.org/nextspace/006/research.htm

[11] VIAF: http://www.oclc.org/research/projects/viaf/

Communications (ITC) provides 24x7x365 maintenance of the servers as well as backup and security services. The Institute has a full-time system administrator who provides additional support for the servers and for the maintenance and publishing software. The four servers host all of the Institute's publicly accessible projects using dedicated virtual servers that ensure that each project has dedicated processing resources sufficient to provide efficient and reliable processing and response. SNAC will be hosted in a virtual server environment on one of the Institute's four servers.

3.3. HISTORY, SCOPE, AND DURATION

The Project Director is a member of the SAA EACWG and is the chief technical architect of EAC-CPF. With Richard Szary (University of North Carolina–Chapel Hill) and Wendy Duff (University of Toronto), he initiated an effort to develop a communication standard for archival authority control based on ICA's ISAAR in 1998. A preliminary meeting was held in New Haven in 1998, funded by the Digital Library Federation. With subsequent funding from the Gladys Krieble Delmas Foundation, an internationally representative group of archivists met in Toronto in 2001 where EAC was initially designed. Subsequent meetings in Charlottesville (2002) and Stockholm (2003) led to the release of EAC *alpha* and *beta* versions. In 2006 SAA created and charged the EACWG with evaluating and revising EAC and releasing it as a standard. With funding from the Delmas Foundation, the EACWG met in Bologna in 2008, thoroughly revised the intellectual structure of EAC, and renamed it EAC-CPF to reflect that creator description is one component of contextual description (the other being activity and function description). It will be released in fall 2009.

During testing of the new standard, 25 sample EAD-encoded finding aids randomly selected from the OAC were used as a source for extracting EAC-CPF records. Testing demonstrated that EAC-CPF records for creators and related corporate bodies, persons, and families referenced in controlled entries could easily be extracted from various components of the finding aids using Extensible Stylesheet Language–Transformation (XSLT). Testing also revealed that additional information on the creator, such as birth and death places, and other entities not referenced in controlled name entries could also be extracted, though extraction would depend on sophisticated matching algorithms not easily and efficiently implemented in XSLT.

The 25 creators represented 18 persons, 5 corporate bodies, and 2 families. The names of the creators were searched against the Library of Congress Name Authority File (LCNAF). Of the 25 creators, 16 records matched authority records in the LCNAF. Of the 16 records, 14 were persons and 2 were corporate bodies. Eleven LCNAF records had alternative names. The high percentage of matching records and alternative names in matching records indicated that matching against the LCNAF will enhance the overall quality of EAC-CPF records derived from EAD-encoded finding aids, through normalization of authorized forms and addition of alternative names. Given Library of Congress archival description practices, it is anticipated that all or almost all creator entries in Library of Congress finding aids will match LCNAF files. Controlled headings and names used in correspondence description were not systematically examined.

The Project Director contacted Ray Larson at the School of Information, University of California, Berkeley (UCB/SI). Larson is currently working on an IMLS-funded project entitled *Bringing Lives to Light: Biography In Context*, that involves, among other activities, matching and extracting personal names and biographical information from digital text resources.[12] Based on Larson's successful experience with the IMLS-funded project, the Project Director concluded that that XSLT, enhanced with additional full-text matching and extraction algorithms, could accurately match and extract corporate body, person, and family names, and related biographical and historical data from EAD-encoded finding aids, into EAC-CPF descriptions.

Further discussions with Larson focused on the feasibility of matching EAC-CPF records with one another and with LCNAF and ULAN in order to merge and enhance records describing the same named entity. Both

[12] *Bringing Lives to Light: Biography In Context*: http://ecai.org/imls2006

agreed that the methods for matching records being used in the Virtual International Authority File (VIAF) project could be adapted for use in the SNAC Project.[13]

In addition to Larson, the Project Director also contacted Adrian Turner, Rosalie Lack, and Brian Tingle at the California Digital Library (CDL). CDL agreed to participate and to take the lead in contacting the Northwest Digital Archive (NWDA) to secure its support and access to its EAD-encoded finding aids. The Project Director agreed to take responsibility for contacting the Library of Congress to secure support and access to its collection of EAD-encoded finding aids and the LCNAF, the Getty Vocabulary Program to secure access to ULAN, and OCLC Research to secure access to the "enhanced" LCNAF records used in the VIAF project.

Both SI/UCB and CDL will be acting as subcontractors for the Project. See Appendix I for background information on the project partners and Appendices N and O for detailed statements of work.

The Project will last two years and have two major components. The first component will involve acquisition of the data, extraction of names and biographical/historical data and migration to EAC-CPF records, and the matching, disambiguation, and merging of EAC-CPF-derived records with LCNAF and ULAN records. The second component of the Project will be devoted to the development of the prototype access system based on the derived and enhanced EAC-CPF records created in the first phase. While the fully developed access system will be dependent on the completion of work in the first phase, it is anticipated that a sufficient quantity of representative examples of EAC-CPF records will be available early in the Project to enable development work on the access system to begin early in the project.

Project work will continue beyond the proposed two-year period. The software and prototype access system will be maintained on IATH's servers and will remain available to the public via the project's web site. The derived and enhanced EAC-CPF descriptions will also be maintained at IATH. The findings from this project will be disseminated to the archival and library communities and the standards organizations which work with these communities. We anticipate taking the project results forward into a second phase and plan to submit further grant applications to support future development of this research.

3.4. METHODOLOGY AND STANDARDS

Core standards

The descriptions of the creators will be based on Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF), an international standard for authority control of corporate body, person, and family name entries *and* biographical or historical description expressed as an Extensible Markup Language (XML) schema.[14] (See Appendix K for an example of an encoded EAC-CPF instance.) The Society of American Archivist (SAA) Encoded Archival Context Working Group (EACWG) is responsible for ongoing intellectual and technical maintenance of the standard. The Bundesarchiv (Germany) and Staatsbibliothek zu Berlin host the XML-based schema, Tag Library, and additional supporting documentation.[15] The EACWG group has representatives from Australia, Canada, Great Britain, Greece, France, Germany, Italy, Sweden, and the United States (see Appendix H for a list of members). EAC-CPF is a communication standard for *International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (ISAAR(CPF))*, an ICA standard.[16] EAC-CPF is designed to support the following ISAAR(CPF) descriptive components:

---

[13] See Bennett *et al.*

[14] XML: http://www.w3.org/TR/xml/

[15] EAC-CPF: http://eac.staatsbibliothek-berlin.de/

[16] ISAAR(CPF): http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf

- Identity. Addresses authority control and includes one or more alphanumeric identifiers, and one or more controlled headings for both authorized (or preferred) and alternative names.
- Description, including:
  - Places and dates, such as birth, death, jurisdiction, etc.
  - Legal Status (corporate bodies)
  - Functions, activities, and occupations
  - Mandates (corporate bodies)
  - Organizational structure or genealogy
  - Prose or chronological list (date, place, event) biographical or historical description
- Entries for related corporate bodies, persons, and families.
- Entries for related archive, library, museum or Internet resources by or about the described entity.
- Entries for function authority control records.
- Control. Information used in managing and maintaining the record, including detailed information about data sources, such as finding aids and LCNAF and ULAN records.

EAC-CPF also supports the merging and aggregation of authority control records and collaboration between multiple institutions using different cataloging rules or languages. Merging authority records and descriptive rule diversity will be demonstrated in this project; cooperative authority control will not.

The primary source of name and biographical and historical information for corporate bodies, persons, and families will be derived from finding aids represented in Encoded Archival Description (EAD), an international communication standard for archival description expressed as an XML schema. First released in 1998 and revised in 2002, EAD is used internationally by hundreds of government archives; research archives; businesses; museums; and international, state, and regional consortia. EAD has been translated into French, Spanish, Dutch, Greek, German, Italian, and Chinese. EAD is based on *General International Standard Archival Description (ISAD(G))*, an ICA standard.[17] The EAD Working Group is responsible for ongoing intellectual and technical maintenance of the standard. The Library of Congress hosts the XML-based schema, Tag Library, and additional supporting documentation.[18] In addition to serving as a source of authority control and biographical and historical information, links to EAD finding aids will also be used as one demonstration of the utility of creator description for access to cultural heritage resources.

The Resource Description Framework (RDF) will be used for interrelating EAC-CPF entity descriptions and EAC-CPF with EAD-encoded findings aids. Like XML, RDF is a World Wide Web Consortium (W3C) standard that is broadly supported with a variety of open-source tools. Use of RDF will facilitate efficient representation and processing of linking (or relational) data and facilitate linking to DBPedia, a "community effort to extract structured information from Wikipedia" that contains over 200,000 persons and 20,000 companies.[19]

The Metadata Authority Description Schema (MADS)[20] and MARC (Machine-Readable Cataloging) Authority,[21] both maintained by the Library of Congress, support authority control of names. EAC-CPF goes

[17] ISAD(G): http://www.ica.org/sites/default/files/isad_g_2e.pdf

[18] EAD: http://www.loc.gov/ead/

[19] DBPedia: http://dbpedia.org/About

[20] MADS: http://www.loc.gov/standards/mads/mads-doc.html

[21] MARC 21 Authority: http://www.loc.gov/marc/authority/

well beyond both of these standards by supporting detailed biographical and historical information, as well as detailed information on related named entities and resources. These features are essential in meeting the descriptive needs of archivists as detailed in ISAAR(CPF). They are also essential in meeting project objectives.

Authority control information in EAC-CPF instances derived from EAD finding aids will be normalized and augmented with alternative names derived from MARC 21 Authority records in the LCNAF. The Library of Congress has granted research access to the 5 million records in the LCNAF. MARC Authority records will thus function as sources of information but will otherwise not have a role in the project. EAC-CPF will also be matched against Getty Vocabulary Program ULAN records. The Getty Vocabulary Program has granted research use of ULAN. Similar to EAC-CPF, ULAN records address both authority control and additional biographical/historical data.[22] (See Appendix L for a sample LCNAF record and sample ULAN record.) Both controlled headings and biographical/historical data in ULAN records matching EAC-CPF will be incorporated into the EAC-CPF record. Given that the ULAN creator domain is limited to artists, it is anticipated that matching frequency will be low. Nevertheless, matching records will offer the opportunity to explore merging museum authority records with archival authority records.

Neither the LCNAF nor the ULAN specifically addresses families. The Library of Congress treats families and family names in the Library of Congress Subject Headings (LCSH). The rules for creating family name entries for librarian differ substantially from the practice of archivists. The differences in practice make the matching EAC-CPF family names against LCSH family names of only limited interest.

The prototype access system developed in the project will be based on Extensible Text Framework (XTF), an open-source platform for publishing XML-encoded documents and data developed by the CDL.[23] XTF incorporates other open-source software, in particular Saxon, an XSLT processor, and Lucene, a robust, high performance text search engine developed by the Apache Project.

In addition to XML, the related technologies Extensible Stylesheet Language-Transformation (XSLT) and XML Path Language (XPATH) will be employed in the project in a number of capacities.[24] XSLT/XPATH will be used in the extraction of data from EAD-encoded finding aids and its transformation into EAC-CPF instances. XSLT will also be used for extracting both MARC Authority and ULAN data and merging it with derived EAC-CPF records. Finally, XSLT will be used in the configuration of the prototype access system and in rendering the EAC-CPF records.

Related Standards and Projects

The most recent version of the *Text Encoding Initiative Guidelines* P5 (TEI P5) extended the schema's semantics and structure to address personal and organizational names in more detail, and now accommodates biographical, prosopographical, and historical data in elective degrees of structure. However, the TEI P5 approach to the description of corporate bodies, persons, and families does not address the issue of authority control as such. While TEI P5 supports highly structured descriptions, it does so using general descriptive components and thus does not specifically address the components of description articulated by ISAAR(CPF). Finally, TEI P5 does not specifically and formally address access to related resources, and interrelating entity description with other entity descriptions.[25]

---

[22] ULAN: http://www.getty.edu/research/conducting_research/vocabularies/ulan/about.html

[23] XTF: http://www.cdlib.org/inside/projects/xtf/

[24] XSLT: http://www.w3.org/TR/xslt20/ and XPATH: http://www.w3.org/TR/xpath20/

[25] See TEI P5, "Names, Dates, People, and Places," http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html

There are several important projects that are or will be using EAC-CPF. The most important project is *People Australia*, a project sponsored by the Australian National Library. *People Australia* is creating an online resource discovery service that will allow users to access information about significant Australian people and organizations and related biographical/historical information.[26] The SNAC Project Director is in regular contact with the chief technical architect of the project, Basil Dewhurst, who also is a member of the EACWG. Two additional members of the EACWG, Karin Brendenburg, at the Riksarkivet (National Archives of Sweden), and Anila Angjeli, at the Bibliothèque nationale de France, have also reported to the Project Director on the use of EAC *beta* nationally in both countries and the plans to migrate to the use of EAC-CPF. Finally, the Archive Portal of Europe (APEnet), a project funded by the European Union, also plans to use EAC-CPF.[27] The SNAC Project Director is providing technical consulting to APEnet and is in regular contact with Angelika Menne-Haritz, Vizepräsidentin Bundesarchiv (Federal Archive of Germany) and Peder Andrén (Riksarkivet), both playing leadership roles in the development of APEnet. The SNAC project will monitor and communicate with all of these projects through established contacts.

Also worthy of mention is People of the Founding Era (PFE).[28] This is an ongoing project of Document Compass, an organization that provides services to the documentary editing community. Using biographical data from several documentary editing projects focused on the U.S. founding era, PFE intends to develop a resource and tool to facilitate prosopographical research. While the project is in its early stages, the Project Director has discussed the possibility of creating an EAC-CPF "export function" that would allow use of the biographical data in a broader socio-historical context. The SNAC Project Director will continue to discuss opportunities for collaboration with PFE.

Methods

Based on preliminary testing of a set of EAD-encoded finding aids, the extraction of name and biographical/historical data for the creators of finding aids will not present significant technological challenges. (See Appendix F for examples of EAD encoding for creator name and biographical/historical data). Some biographical/historical data, such as occupation, place of birth or death, and life dates, may require identifying key data components in discursive text if each datum is not specifically encoded (or tagged). The identification of occupations will be enhanced by compiling a list from existing finding aids that formally provide occupation terms and then using this list to enhance identification of occupation terms given in discursive description.[29] The identification of U.S. geographic names will use a technique that can be described as semantic structure analysis that identifies proper noun strings and compares them with gazetteers. The matching of geographic names will be enhanced using the Geographic Names Information System (GNIS)[30] maintained by the U.S. Board on Geographic Names, and identification of geographic names outside of the United States will be enhanced by GeoNames,[31] an international community-maintained geographic information data aggregating service.

Extracting the names of related named entities from encoded controlled lists will not present major processing challenges because names are specifically tagged. (See Appendix G for examples of EAD encoding

[26] *People Australia*: http://www.nla.gov.au/initiatives/peopleaustralia/

[27] APEnet: http://www.apenet.eu/

[28] PFE: http://www.documentscompass.org/DCMellonPressRelease.html

[29] While the Library of Congress Subject Headings (LCSH) provides many occupation terms, the terms are not distinguishable from other topical terms. The U.S. Department of Labor also maintains a list of occupational titles, but it is generally more specific than the terms used in LCSH (e.g., it does not include "journalist" but LCSH does).

[30] GNIS: http://geonames.usgs.gov/domestic/download_data.htm

[31] GeoNames: http://www.geonames.org/

for controlled lists and correspondence description). Many controlled name entries also contain trailing terms that specify the role the named entity plays in the described records, such as "correspondent." This information will be removed from the name entry but recorded in the EAC-CPF record. Matching of role terms will be enhanced through the use of the Library of Congress's *MARC Code Lists for Relators, Sources, Description Conventions.*[32]

A third category of name extraction involves isolating and identifying names in the description of letters and correspondence. The names used in providing descriptive titles for files or items may or may not be specifically tagged as such, and so identifying names will require employing the semantic structure analysis technique described above but comparing the proper noun strings with name dictionaries or authority files to identify corporate body, person, and family names. Both LCNAF and ULAN will be used to enhance the identification of person, and corporate body names.[33] Identifying family names should present less of a challenge, because it is widespread archival descriptive practice to use the term "family" following the family name. If this approach is insufficient, then the proper noun can be verified as a name by matching it against LCNAF and ULAN.

Matching EAC-CPF records with one another and with LCNAF and ULAN in order to merge and enhance records describing the same named entity presents numerous processing challenges. Names with three or more matching components in the same order and names with one or two components but qualified with birth or death dates can be reliably matched. Other names require using clues from the contexts in which the name appears to increase the reliability of matches. For example, the abstract in the finding aid to the George Oppen Papers (Mandeville Special Collections Library, University of California, San Diego) describes Oppen as an "objectivist poet and winner of the Pulitzer Prize for Poetry in 1969." The source information in the LCNAF record for Oppen contains the description "objectivist poet." By including the occupation description found in the finding aid and the LCNAF record in the matching process, the reliability of matches can be greatly increased.

For the Virtual International Authority File (VIAF) OCLC Research enhances LCNAF records before attempting to match them against authority records from other sources, such as the Australian National Library. Enhancing the authority records involves adding information from associated bibliographic records that provides additional clues for determining matches.[34] OCLC Research has granted access to the "enhanced" LCNAF for use in this project, which will increase the quantity of reliable matches.

At each step, random samples of the results of the processing will be evaluated for completeness and accuracy. Based on these evaluations, processing algorithms will be adjusted and the processing repeated until an acceptable level of completeness and accuracy is achieved.[35] Particularly critical is the matching error rate when comparing authority records with the same or similar names with one another. This will take place at two steps in the processing: 1) when matching derived EAC-CPF records with one another; and 2) when matching the resulting set of unique EAC-CPF records with enhanced LCNAF records and ULAN records.

Also at each step in the processing, data will be collected that will be of immediate benefit to the project and of long-term benefit to the archival community for evaluating practice and planning and to utilities providing services to the archival community. Some data will be collected algorithmically, such as the precise number of finding aids processed, tag usage in EAD-encoded finding aids, and the number of names extracted from

---

[32] *MARC Code Lists for Relators, Sources, Description Conventions*: http://www.loc.gov/marc/relators/

[33] It is important to note that this use of authority files is limited to identifying proper nouns that are names, in order to derive preliminary EAC-CPF records which will subsequently matched against the LCNAF and ULAN to find records that describe the same named entity.

[34] See Bennett *et al.*

[35] "Acceptable level of accuracy" will be based on benchmarks established in the VIAF project.

finding aids before removing duplicates and merging. Other data will require qualitative evaluation based on statistically valid random samples of authority records produced.

The project will produce two significant products: *open-source software* that will facilitate efficiently and accurately deriving authority control records from existing archival finding aids and Library of Congress and Getty Vocabulary Program name authority records, and an *open-source prototype archival access system* based on the derived and enhanced records. The success of the project will be based on two methods. The open-source software will be evaluated based on the quantitative data and qualitative assessment performed by project staff. In particular, the open-source software will need to produce a high percentage of reliable authority relative to records that require human review and revision. Both the findings of the internal evaluation and the effectiveness of the open-source prototype archival access system will be evaluated by representatives from the collaborating and cooperating institutions in a two-day meeting held three months before the end of the project. Holding the meeting three months before the end of the project will leave sufficient time to incorporate suggested revisions in the prototype. At the end of the project, the representatives of the collaborating and cooperating institutions will be asked for a final written evaluation of the prototype system.

All data used in the project are either in the public domain or permission for use has been granted for research purposes. See letters granting permission in Appendix M.

3.5. WORK PLAN

The proposed work plan involves two sets of related activities:

1. Develop software to derive names and biographical/historical data from EAD-encoded finding aids, migrate that data into EAC-CPF records, and match and merge data from LCNAF and ULAN authority records to produce the final EAC-CPF records.

2. Development of the prototype access system, based on XTF. The prototype access system will essentially have three components: 1) display of the EAC-CPF records; 2) browse and search of the EAC-CPF records; and 3) create a proof of concept API so that other interested institutions can embed the prototype's functionality into their site.

While the final development of the access system will be dependent on the completion of the first set of activities, work on both can proceed in parallel. In evaluating the feasibility of the proposed project, 25 EAC-CPF records were created. These records and others generated early in the process will provide a sufficient number of records to begin development of the access system.

Pre-grant

- Project Director (PD) will continue to work with samples of EAD-encoded finding aids from the participating consortia and the Library of Congress, generating additional EAC-CPF records. This work will ensure that the project begins with a detailed profile of the range of descriptive and encoding practice represented in the nearly 30,000 finding aids (see Appendix E for a complete list) to be used in the project. It will also ensure that the project can begin developing the prototype access system at the start of the grant.

- PD will contact participating consortia and repositories to gather URI protocols for accessing individual finding aids, and for performing name searches in the participating archival access systems and a subset of affiliated online library catalogs, to include the Library of Congress catalog and WorldCat.

- PD will compile URI protocols for name searches in Wikipedia, Flickr, and DBpedia. URI protocols for accessing individual finding aids will be essential in deriving EAC-CPF records linked to the finding aids from which they were derived. The URI for protocols will be essential for building dynamic searching options into the prototype access system interface.

- IATH system administrator will set up virtual server at IATH to function as the central repository, software development environment, and prototype access system. System administrator will also set up a dedicated Subversion repository to provide version control of all programs and scripts developed by the project and for all documents and static Web site pages. Set up will involve creating accounts for staff working on the project at the University of Virginia, UCB/SI, and CDL.

May 2010 - October 2010

- A project Web site will be established describing the project and making the project proposal and timeline available. Project and Web site will be announced to the various online discussion groups, including the EAD List, relevant SAA groups (such as the EAD Roundtable, Metadata and Digital Objects Roundtable, etc.), relevant H-Net groups, and the Archives and Archivists, TEI, and DigLib lists.

- PD, Co-PD, Tingle, Turner, and Larson will meet for two days at either CDL or SI/UCB to review the work plan, paying particular attention to detailed description and sequencing of the necessary steps, the specific roles of the project staff in each step, and the technology and techniques essential to each step.

- CDL staff (Tingle and Turner) will begin exploration to identify the features and functionality for the access system. CDL staff will first collect and create use cases. These cases will then inform the creation of personas (archetypal users that represent the needs of the larger user community). A persona essentially describes users' motivations, expectations, and goals with regards to their online behavior. The goal of these two methods is to gain a clear understanding of user needs, in order to identify and prioritize features and functions.

- Once user data has been gathered, the CDL staff, in close consultation with IATH staff, will use the findings to inform the design and creation of the EAC-CPF records display. A follow-up lightweight assessment will then be conducted.

- IATH system administrator will acquire copies of OAC, NWDA, VH, and LoC EAD-encoded finding aids, and copies of the VIAF enhanced LCNAF and ULAN.

- IATH programmer will develop XSLT programs to derive preliminary EAC-CPF records from EAD-encoded finding aids from tagged name entries for creator and controlled entries. Derived creator records will also have derived biographical/historical description. The process will be iterative, with PD, Turner, and programmer evaluating the results and refining processing until the results are satisfactory. Larson and UCB/SI programmer will write programs to further enhance the biographical/historical description by identifying occupation, and birth and death dates and places. Larson and UCB/SI programmer will develop program to derive additional EAC-CPF records from names used in the description of correspondence.

- Beginning in the fall of 2010, using a complete set of derived EAC-CPF records, Larson and UCB/SI programmer will begin developing programs to identify and merge EAC-CPF records describing the same entity, to produce a unique set of EAC-CPF records. This process will be iterative, with PD, Turner, and Larson evaluating the results and refining processing until the results are satisfactory. Processing must distinguish between different names for the same entity and the same name for different entities. If necessary, preceding processing steps will be further refined and enhanced to improve the quality and quantity of matching.

- PD will compile quantitative and qualitative data documenting the results of the processing to date.

November 2010 - April 2011

- Larson and UCB/SI programmer will complete development of programs to identify and merge EAC-CPF records. Subsequently, Larson and UCB/SI programmer will begin developing programs to match the unique set of EAC-CPF records created in earlier processing against the VIAF-enhanced LCNAF and ULAN records. The new programs will incorporate algorithms used in earlier matching and merging processing, as well as new processes and algorithms for exploiting the data in the LCNAF and ULAN records. LCNAF headings will be used to verify or normalize the heading in the EAC-CPF record, and both the LCNAF and ULAN records will be used as a source of alternative names. ULAN records also contain data on nationality, occupation, and birth and date dates and places. This additional data will be used to verify or augment the EAC-CPF description. This process will be iterative, with PD, Turner, and Larson evaluating the results and refining processing until the results are satisfactory.

- In consultation with IATH staff, CDL staff will create the search/browse functionality for the EAC-CPF records. Another round of end user assessment will be conducted.

- PD will compile quantitative and qualitative data documenting the results of the processing to date.

May 2011 - October 2011

- CDL and IATH designers and programmers will continue design and development of the access system.

- In the summer of 2011, Larson and UCB/SI programmer will produce final set of derived and enhanced EAC-CPF records. Final set of documents will include RDF triples used to relate EAC-CPF records to finding aid and EAC-CPF records to one another. As the final processing is refined, sample records will be ingested into the access

- CDL and IATH designers and programmers will ingest the final set of EAC-CPF records into the access system. Any indexing and rendering issues presented by the full set will be identified and resolved.

- CDL staff will work on the creation of an API component that will allow other institutions to embed the access system functionality into their locally-hosted sites.

November 2011 - April 2012

- CDL and IATH designers and programmers will continue design and development of the access system, including graphical display of social networks. Final round of end user assessment will be conducted. And final refinements to the access system implemented.

- In March of 2012, representatives from the collaborating and cooperating institutions will meet for two days to evaluate the quantitative and qualitative findings of the project staff, and to evaluate the effectiveness of the prototype archival access system. Participants will be asked for suggestions on the indexing and retrieval, navigation, linking to resources by and about named entities, and the rendering of search results, EAC-CPF records. Additional suggestions for functional or interface improvements will be solicited.

- Based on the suggestions gathered in the meeting, the CDL and IATH designers and programmers will refine and enhance the interface and functionality of the access system. Project and Web site will be announced to the various online discussion groups, including the EAD List, relevant SAA groups (such as the EAD Roundtable, Metadata and Digital Objects Roundtable, etc.), relevant H-Net groups, and Archives and Archivists, TEI, and DigLib lists.

3.6. STAFF

Project Director Daniel Pitti, Associate Director of IATH, is the chief technical architect of both the EAD and EAC-CPF standards and a member of the both the SAA EAD Working Group and the SAA EAC Working Group. In addition to his work with archives and libraries, he has extensive experience in the design and implementation of scholar-driven humanities research projects that employ advanced technologies. As Project Director, he will be responsible for overall supervision and coordination of the project. He will directly supervise two IATH programmers that will work on the project. He will contribute archival, authority control, and technological expertise in the development of the open-source software and prototype access system. He will share responsibility with other project staff for the qualitative evaluation of all critical processing and development steps and will devote 15% of his time to the project.

Co-project Director Worthy Martin, an Associate Professor of Computer Science and an Associate Director of IATH, has been one of the prime information architects on numerous digital humanities projects through IATH. He will provide expert advice on all aspect of the project design and implementation. He will assist the Project Director in supervising IATH programming staff and will devote 5% of his time to the project.

Brian Tingle, Technical Lead, Digital Special Collections, California Digital Library (CDL), has been involved in designing and building web-based access systems for UC Libraries for the last 13 years, including the OAC and Calisphere.[36] Since 2001, he has worked extensively with EAD as the lead technical architect for the OAC. Tingle has experience with all aspects of web development and production including server side technologies that will be used in this project such as XTF, front-end technologies including JavaScript and CSS, user-centered interaction design methodologies, and information architecture. He will be responsible for the production of the access interface to the system and will devote 30% of his time to the project.

Adrian Turner, Data Consultant, Digital Special Collections, CDL, has been a member of the CDL since 2002, and has experience coordinating OAC operations, supporting OAC contributors with EAD encoding, and troubleshooting and quality control checking of encoded content vis-à-vis OAC displays. He has successfully managed CDL activities on past and ongoing collaborative grant projects, such as the multi-year (2000-present) LSTA-funded Local History Digital Resources Program[37] and multi-year (2001-2005) Library of Congress-supported California Cultures project.[38] His previous work experience includes processing archival collections, including the use of EAD and MARC21 standards. He will be responsible for coordinating CDL's work on the project, assisting with quality control checking of records and prototype system, and coordinating assessment and usability testing on the prototype system. He will devote 10% of his time to the project.

Ray Larson, Professor, School of Information, University of California, Berkeley. Larson's current research focuses on several related areas of information retrieval and digital libraries, including how to exploit XML structure and content in heterogeneous collections of XML documents. His work on the IMLS-funded project entitled *Bringing Lives to Light: Biography In Context* is directly related to key components of SNAC.[39] Larson, with a graduate student programmer, will be responsible for extracting names from discursive text; matching, merging, and enhancement of authority records; and mapping data documenting relations among named entities into RDF. He will devote one summer month in the first year of the project (33.3%) and one-half summer month (16.7%) in the second year of the project.

---

[36] Calisphere: http://www.calisphere.universityofcalifornia.edu/

[37] Local History Digital Resource: http://www.cdlib.org/inside/projects/oac/lsta/

[38] California Cultures: http://www.calisphere.universityofcalifornia.edu/calcultures/

[39] *Bringing Lives to Light: Biography In Context*: http://ecai.org/imls2006

Graduate Student Programmer, School of Information, University of California, Berkeley. Under the supervision of Larson, the graduate student programmer will be responsible for extracting names from discursive text; matching, merging, and enhancement of authority records; and mapping data documenting relations among named entities into RDF. The graduate student will devote 47% of one academic year (2010-2011) and 75% of one summer (2011) to the project.

System administrator, IATH, will be responsible for the project server, including all software installation, acquisition and management of data, version control of project-developed software, and assisting all programmers in setting up efficient processing environment. The system administrator will devote 15% of time to the project.

XML Programmer, IATH, will be responsible for development of XSLT/XPath programming for derivation of EAC-CPF records from EAD-encoded finding aids and enhancing the EAC-CPF records using LCNAF and ULAN authority records. The XML programmer will share responsibility for development of XTF-based prototype access system. The XML programmer will devote 35% of time to the project.

Advisory Board

The Advisory Board will provide ongoing advice to the project to ensure that the professional perspectives of archivists, manuscript librarians, and historians inform the projects' activities and products. The Advisory Board will play a significant role in evaluating the two major objectives of the project.

Jodi Allison-Bunnell, Program Manager, Northwest Digital Archives, Orbis Cascade Alliance

Scot French, Director, Virginia Center for Digital History, University of Virginia

Edward Gaynor, Head of Collection Development and Description, Albert and Shirley Small Special Collections Library, University of Virginia

Mary Lacy, Manuscript Librarian, Manuscript Division, Library of Congress

Michelle Light, Archivist and Acting Head of Special Collections and Archives, Libraries, University of California, Irvine

3.7. DISSEMINATION

The open-source software developed in the project to facilitate the creation of EAC-CPF records derived from EAD-encoded finding aids and enhanced with additional data from LCNAF and ULAN records will be made available through SourceForge, with accompanying documentation on its use. The availability of the software will be announced on the EAD Listserv. The editor of the SAA EAD-Help pages will also be requested to post the release of the software in its news section, and to create a link to it in "Tools and Helper Files."

The prototype archival access will be based on XTF, an open-source XML indexing and rendering system developed by the CDL and available through SourceForge. The extension of XTF functionality to work with the EAC-CPF schema and records will be made available in future releases of XTF.

The Project will be announced on the EAD Listserv, relevant H-Net groups, Archives and Archivists list, and the TEI list. The Project will also propose presentations at the 2011 and 2012 SAA Annual Meetings, 2012 Annual Meeting of the American Library Association Authority Control Interest Group (LITA/ALCTS CCS), and 2012 Annual Meeting of the American Historical Association. Publications will be submitted to the *American Archivist*, *D-Lib Magazine*, and *Journal of Archival Organization*.