Social Networks and Archival Context: Cooperative Program Planning

A Proposal to The Andrew W. Mellon Foundation

PROPOSAL SUMMARY

The Institute for Advanced Technology in the Humanities (University of Virginia) in collaboration with the U.S. National Archives and Records Administration, the School of Information (University of California, Berkeley), and the California Digital Library (University of California) proposes to transform *Social Networks and Archival Context* (SNAC)[1], a research and demonstration project, into a sustainable archival cooperative program maintained by the archive, library, and scholarly communities. Funded by grants from the National Endowment for the Humanities (2010-2012) (phase one) and the Andrew W. Mellon Foundation (2012-2014) (phase two), SNAC has demonstrated that scholarly historical research can be dramatically transformed by a research tool that both integrates access to the dispersed resources that document the lives, work, and events surrounding historical persons, and provides unprecedented access to the biographical-historical contexts of the people documented in the resources, including the social-professional-intellectual networks within which the people lived and worked. In October of 2011, the Institute for Museum and Library Services awarded the University of Virginia a two-year grant to begin the process of building professional and research community support for and the preliminary planning of a cooperative program that would build on the research and data foundation established by the SNAC project. A series of meetings and discussions established the support of key institutions and individuals, and identified the principal issues that need to be addressed in establishing a Cooperative. While significant progress has been made in building support and preliminary planning, additional planning and foundational work need to be completed before the Cooperative can be formally established. We are thus requesting $344,214.56 funding to complete the planning and foundational work over twelve months; in particular to perform user requirement studies and detailed technical infrastructure planning, to establish the initial administrative and governance structure, to determine requirements and develop models for sustainability, to establish a legal foundation and intellectual property policies, and to recruit core initial members. The comprehensive plan developed would serve as the foundation for launching a pilot phase of the Cooperative that would commence at the end of the planning process.

PROPOSAL NARRATIVE

Archivists, librarians, and scholars involved in SNAC and cooperative planning activities have responded enthusiastically to its promise and the premise of a Cooperative founded upon it, seeing an opportunity to dramatically improve the economy and nature of research into the lives of, work by, and events surrounding historic persons. The research tool and the resource being developed enables two novel forms of access: *integrated access* to distributed archival and published resources by and about persons, families, and organizations; and *access to biographical-historical information about the people including the social, professional, and intellectual networks* within which people lived and worked. The SNAC prototype research tool and resource enables users to search for persons, organizations, and families by name; read biographies or histories for the entities found; find people based on occupation, or topics with which they are associated; link to finding aids and other archival record descriptions for primary resources related to the entity; link to bibliographic descriptions for published works by the entity; and link to persons, organizations, and families with whom the entity corresponded or is associated (that is, other persons, organizations, and families). Thus the prototype is both an access tool and a biographical-historical resource, both a means and an end. Currently the SNAC phase one prototype is available to users; the phase two prototype is currently being revised based on user study findings, and the number of descriptive records expanded from 128,787 to nearly 3 million or possibly more.

Ed Ayers, President of the University of Virginia and a Civil War historian, captures this enthusiasm succinctly:

> SNAC promises to change the way history is imagined and written! For all that the digital revolution has

[1] See project site http://socialarchive.iath.virginia.edu/ and project prototype http://socialarchive.iath.virginia.edu/xtf/search

> revolutionized, the heart of research lies within the primary record embedded in archives large and small. The pioneering work of SNAC will unlock that record, revealing connections and patterns invisible to us now.

Alan Liu, Professor of English, University of California, Santa Barbara and Director of Research oriented Social Environment (RoSE) summarizes the promise of SNAC:

> SNAC employs state-of-the-art computational techniques to do three things very well: 1) unlock information originally recorded for specific purposes in library and other archival finding aids to make them usable in new contexts; 2) connect widely-distributed information of this sort from around the world; and 3) marry the "library" or "archive" model of knowledge to a whole other model of social networks that both humanizes our understanding of the way knowledge emerges from communities of knowledge creators and seekers, and speaks powerfully to today's "social network" generation.

For archivists and librarians, a SNAC-based Cooperative offers not only the opportunity of leveraging their own descriptions to dramatically enhance access to and understanding of original historical resources, but also of improving the economy and quality of description by sharing the products of their work. Michael Rush, Processing Archivist, Beinecke Rare Book and Manuscript Library, Yale University, and Co-chair of the Society of American Archivists' Technical Subcommittee-Encoded Archival Description describes these dual benefits:

> From the perspective of a practicing archivist, a cooperative to assemble descriptions of people and groups with links to the archival records that document their activities will be a boon to the efforts of archivists everywhere. It will connect my collections to related ones elsewhere, and it will help connect end-users with collections everywhere. I anticipate that a central resource for the description of the people and groups documented in archives will result in a measurable economy of scale, promoting less duplication of effort both by archivists engaged in description and users engaged in research.[2]

The SNAC research tool is based on extracting data describing the creators and others documented in historical resources from the descriptive guides to the resources, and assembling these into descriptions of people who are in turn linked to one another (via social, professional, and intellectual networks) and to the resources that document their lives and work. While SNAC has convincingly demonstrated the potential professional and scholarly benefits of the novel research tool, it is also clear that *computer-based techniques alone* cannot fully realize the transformative potential that this research tool offers researchers. Archivists and librarians did not create the descriptive data in the guides (or finding aids) with SNAC's use of the data in mind. The quality of the data is uneven and many of the relations of individuals, families, organizations and associated historical documents are overlooked. The quality and thoroughness of the research tool demonstrated in SNAC, while impressive and compelling, reflects the varied quality of the data sources The desire to realize the research tool's full transformative potential is the primary rationale for establishing a cooperative to build, maintain, and enhance the data, thereby improving its quality, extending the social and document interrelations they document, and integrating and interrelating additional description of people.

During the first and second phases (see History of SNAC, below), with the goal of transforming SNAC from a research and demonstration project into a sustainable cooperative program, the support of principal institutions and individuals was secured, legal and intellectual property issues needing to be addressed, and areas requiring additional detailed planning, in particular technical infrastructure planning, were identified. Of particular significance, the U.S. National Archives and Records Administration (NARA) committed in January 2013 to hosting and administering the Cooperative, including its community-based governance. The technological infrastructure of the Cooperative will be hosted and developed outside of (though in consultation with) NARA. While national archives frequently provide leadership and services to archives within their various nations, historically NARA has not played this role. But, under the leadership of David Ferriero, Archivist of the United States, NARA is actively pursuing a leadership and service role in the national and international archival communities. The archival community has long desired NARA to play a leading role, and thus welcomes this change. As the "nation's records keeper," NARA is preeminently positioned to host the Cooperative.

---

[2] Please see Appendix 2 for statements from the SNAC Advisory Board.

For this planning stage, Laura Campbell, retired Chief Information Officer at the Library of Congress, has committed to providing expertise and assistance in establishing a federally hosted cooperative program. This is particularly important because NARA lacks experience and expertise in this area, and Campbell will provide training and guidance to NARA staff. NARA will need assistance specifically in developing a charter document that will place the Cooperative within its legal mandate, in navigating the intellectual property and data ownership issues associated with the cooperatively created and maintained data, and in developing a sustainable business and operational model that will work within the context of a federal agency.

Several other tasks will be carried out during the grant period. A team led by Campbell will document the requirements for administration and governance, including core staffing, as well the methods by which Cooperative members will be involved in determining policy. Working with the PI, the development team will also recruit participants for the pilot phase of the Cooperative. Rights and permissions secured from data providers for the SNAC research and demonstration phase will be renewed and extended (as needed) to cover use of the data in the Cooperative. An advisory board representing the primary anticipated users of the Cooperative data will be recruited. Rachael Hu, a User Experience Design Manager at the California Digital Library, will conduct two types of user studies. The first will be targeted at researchers, such as historians, academic faculty, instructors, and genealogists, and scholars, and thus will focus on the public interface requirements. This study will focus on researchers and scholars, in particular historians, documentary editors, and genealogists. The second study will target librarians and archivists who will contribute data to and assist in the maintenance of cooperative data, and thus will focus on processing workflows and maintenance interface. The results from the user study will be integrated into user-centered requirements and specifications for the maintenance system and the public research tool. In collaboration with the User Experience Design Manager, a team led by IATH's SNAC Data Extraction Programmer will develop detailed technical architecture specifications for the two major components of the Cooperative: maintenance system and public research tool.

Immediately following this planning grant, based on foundational documents and technical specifications developed, the Cooperative will be implemented in a pilot phase. The objective for the pilot phase implementation will be an established Cooperative based on a sustainable business and operations model.

History of SNAC

SNAC began in 2010 with a National Endowment for the Humanities (NEH) award of $348,221 for the two-year initial phase (phase one) of SNAC (May 2010-April 2012). The University of Virginia contributed an additional $70,026 in cost-share. The NEH and University of Virginia funds supported acquiring and processing data, and the development of a public prototype (see a detailed description of work below). As part of this initial phase, the Andrew W. Mellon Foundation awarded the University of California, Berkeley $20,154 in November 2011 for SNAC-related work. Because of overly optimistic assumptions in the NEH work plan, the full cost of the School of Information, University of California, Berkeley's (SI/UCB) ongoing development of match/merge algorithms and programs and the use of them in processing was funded only through the first eighteen months of the first phase. The Mellon funds enabled SI/UCB to employ a graduate student to assist in the development and application of match/merge processing programs for the last six months of the first phase.

In 2011, the Institute for Museum and Library Services awarded the University of Virginia $193,794 for developing a plan for establishing a sustainable archival description cooperative using the technological infrastructure and data assembled in SNAC as a foundation, with the objective of maintaining and building on the data into the future. Of these funds, $26,191 went to Simmons College and the Society of American Archivists to conduct regional workshops for training archivists and librarians in the use of Encoded Archival Context–Corporate Bodies, Persons, and Families (EAC-CPF), the standard used in the description of organizations, persons, and families in SNAC. The remaining funds were directed to securing professional community support for and to initiate the planning of the Cooperative.

In 2012, the Andrew W. Mellon Foundation awarded the University of Virginia $575,000 for an additional two years of SNAC research and development (phase two).The total funding for SNAC research and development, cooperative planning, and related activities to date has been $1,207,195.

Both phases of SNAC share two primary objectives.

- Extracting or migrating biographical and historical data from existing archival descriptions and assembling these data into standardized descriptions that identify and document organizations, persons, and families based on EAC-CPF, and then matching the descriptions against one another and against data in the Virtual International Authority File, combining records that identify the same entity, to produce a set of unique EAC-CPF records. In the first phase, the source archival descriptions were 30,000 Encoded Archival Description (EAD) finding aids. In the second phase, the number of finding aids was increased to 150,000 (and probably substantially more), and 2.2 million MARC-encoded archival descriptions were included. Further, approximately 380,000 original archival records from the British Library, U.S. National Archives and Records Administration, New York State Archives, Archives nationales (France), and the Smithsonian Institution were included, with the objective of migrating the descriptions from the diverse formats into EAC-CPF.

- Developing a prototype access system, based on Extensible Text Framework (XTF), open source software developed at the California Digital Library (CDL). The prototype access system has three major functional components: 1) display of the EAC-CPF records; 2) sophisticated searching and exploration of the EAC-CPF records; and 3) and exposing the data to enable third-parties to access and use it in other applications.

The first set of activities involved two separate steps in the processing. The first step, extracting or migrating data into EAC-CPF identity records is performed at IATH. The second step, matching and combining the EAC-CPF records with one another and with VIAF records, is being done at SI/UCB. The second set of activities, developing a prototype access and historical resource, is still being performed at CDL.

SNAC: First Phase 2010-2012

The following is a summary of the major accomplishments in the first two-year phase:

- EAD-encoded finding aids were acquired from the Library of Congress (LoC) (1,159), Online Archive of California (OAC) (~15,400), Northwest Digital Archive (NWDA) (5,160), and Virginia Heritage (VH) (8,390). A total of approximately 30,500 finding aids were collected and made available on the project server. (IATH)
- VIAF authority files were acquired and indexed using Cheshire, a probabilistic XML-based indexing system being utilized in the match/merge processing. 5,091,703 VIAF records were indexed and made available for efficient match processing. A sample set of EAC-CPF records were used to successfully test matching and record combining using each both the EAC-CPF and VIAF authority files.
- Successfully generated 173,297 EAC-CPF records from 30,500 finding aids.
- Successfully matched and combined the derived EAC-CPF authority records against one another for all personal, corporate, and family names and against the personal names in VIAF. 128,297 EAC-CPF records resulted from matching.
- A prototype access system was developed, with a first public release in December 2010 (http://socialarchive.iath.virginia.edu/xtf/search). Development employed use cases and identifying baseline functionality. A major focus was developing a graph database for storing and accessing social network relations in support of visualizing relations among organizations, persons, and families. Graph data was migrated into Resource Description Format (RDF), and exposed using a SPARQL endpoint. A "context widget" was developed that enabled users, when browsing linked finding aids in their source contexts to view what SNAC identity records matched named entities in the finding aid, and to access the EAC-CPF records for additional information.

During this phase, several technologies were developed.

- IATH developed a complex XSLT program for deriving data from EAD-encoded finding aids.
- SI/UCB developed code written in the programming languages C and Python for using Cheshire as an

efficient index of the authority files, and to employ its sophisticated Information Retrieval algorithms it matching person, organization, and family names. Also utilized in the match process was the open source PostgreSQL, a robust, efficient object-relational database.

- CDL developed the prototype public system by writing extensions in Java and XSLT to the open source XML publishing platform, XTF, also developed by CDL, and has utilized and developed graph data, retrieval, and rendering programs based on additional open source software: SPARQL, TinkerPop (Stack and Neo4J) and the JavaScript InfoVis Toolkit (JIT) for visualization.

SNAC: Second Phase 2012-2014

In the first phase of SNAC, the project focused on extracting and assembling the authority descriptions from 30,500 finding aids and augmenting the derived authority records with additional data from library and authority records. In the second phase, the number of finding aids was increased to more than 150,000 and augmented by 2.2 million MARC collection-level archival descriptions contributed by OCLC WorldCat. In addition NARA, the Smithsonian Institution, the British Library (BL), Archives nationales (France), and Bibliothèque nationale de France (BnF) contributed over 380,000 original archival identity records in a variety of formats. Twenty-five million VIAF authority files contributed by OCLC Research are being used to facilitate the match processing, and also for data to enhance matching EAC-CPF records, in particular alternative names, languages used, life dates, links to published works by the named entity, and "sameAs" links to Library of Congress Name Authority Records, WorldCat Identity records, VIAF records, and Wikipedia entries.

The following is a summary of the major accomplishments and work underway in the second phase of SNAC.

- 2.2 million WorldCat MARC archival descriptions were acquired from OCLC. Authority files were acquired from the following: Smithsonian Institution Archives (SIA) Joseph Henry correspondents (33,417); New York State Archives (NYSA) agency histories (258); SIA agency histories (243); SIA expeditions and participants (940); and British Library (BL) name authority files associated with the manuscript collections (297,731). 145,771 EAD-encoded finding aids from 42 consortia or individual repositories.[3] Finding aids need to be collected from an additional 14 repositories. The NARA authority files are also outstanding.
- Twenty-five million VIAF "cluster records" were acquired and indexed and made available for match and merge processing in Cheshire.
- 4.5 million EAC-CPF records were derived from the 2.2 million WorldCat archival descriptions; 297,731 from the BL authority files; 1,250 from the SIA expedition data; 32,988 from SIA Henry data; 2,083 from the SIA agency histories; and 541 from the NYSA agency histories. The total number of EAC-CPF records created to date is 4,834,593.

We successfully matched and combined the EAC-CPF records derived from the WorldCat MARC data against one another and against the 25 million VIAF cluster records. The final results from this processing yielded 1,972,224 unique EAC-CPF records (1,289,557 persons; 518,550 organizations; and 518,550 families). Processing of additional EAC-CPF records is ongoing.

Software was developed to extract or migrate data into EAC-CPF format for the following: WorldCat MARC archival descriptions; SIA Joseph Henry correspondents; New York State Archives agency histories; SIA agency histories; SIA expeditions and participants; and British Library name authority files. The software represents a single suite and is currently capable of dealing with MARC bibliographic format (archival description and an idiom used for agency histories), and each of the three data formats specific to the SIA Henry correspondence, SIA expeditions, and the BL authority files. The software has been released as open source on Github, and a web service has also been established where repositories can upload MARC records and in return receive derived EAC-CPF records.[4]

---

[3] Because many of the contributors are contributing more finding aids than were committed at the time of the writing of the second phase proposal, the number of finding aids will far exceed the 136,000 committed. At least 150,000, though perhaps substantially more will be acquired.

[4] The software is available at https://github.com/twl8n/snac_eac_cpf_utils and the web service is available at http://socialarchive.iath.virginia.edu/dev/

An initial end-user requirement study was completed. Ten one-hour user engagement sessions were conducted that included a background interview as well as a usability test on the initial public access prototype design from March of 2013. Users surveyed were derived from the following user types: researcher (genealogist, academic researcher, and reference staff (as researcher surrogate), archival/library staff (processing, description, cataloging, and administrator), and other users with interest in utilizing the network of data amassed within SNAC for digital humanities project interests.

The results of the study revealed user interface "pain point" findings, recommendations for interface fixes and record refinement, as well as user profiles and background stories for all major user types. These user profiles provided the program and development team members with the necessary context to refine program goals, EAC-CPF record processing, and re-working of the interface design for the public access prototype.

Further tasks that are being carried out:

- A thorough revision of the prototype access system, first released in December 2010, is underway, with a major release plan for early 2014. This release is informed by the user requirement study described above, with additional assistance of an experience graphical designer.
- Software for normalizing dates and normalizing geographic names with addition of longitude and latitude data for entries in chronological lists is underway. To date, the date normalization software is complete and tests indicate that the software can accurately normalize over 99% of all dates found! The geographic name normalization is just beginning. The normalization accomplished will enable creating map-timeline graphical displays of peoples' lives in the prototype access system.
- A PhD candidate in the Computer Science Department, under the supervision of Co-PI Martin, is focusing his dissertation research on methods for rendering and navigating the vast amount of historical social network data being compiled in SNAC. The results of this will inform the development of the prototype access system.

There are three primary areas of research and development in this second phase that are ongoing:

- Extraction of data from EAD-encoded finding aids and migration of data from original archival identity descriptions. The EAD extraction software developed in the first phase of SNAC has been thoroughly revised to take into account "lessons learned" and is being finalized to process the 150,000+ finding aids.

- Identity Resolution (identifying when two or more matching or similar name strings represent the same entity) using sophisticated matching algorithms and contextual data. The Cheshire and PostgreSQL match and merge processing has been thoroughly revised to accommodate incrementally processing the records. This extensive and complex revision was necessary, because unlike in the first phase of SNAC, the quantity of records was too large to be processed iteratively as the matching and merging algorithms were developed and refined. This revision also positions the infrastructure to be a viable candidate for serving as the foundation for the maintenance system for the cooperative. Quality review of the match and merge processing was completed, and results incorporated into refinement of the processing.

- Development of techniques for extracting, recording, indexing, searching, and displaying social-professional graphs or networks, in particular techniques for addressing the graphical display of connectors or nodes with large numbers of links to other nodes.

Cooperative Program Planning (2011-2013)

With funding from IMLS, building community support and planning for an archival description cooperative founded on and intended to support the ongoing curation, enhancement, and professional and research use of methods and data assembled in SNAC. Building community support and planning began late in 2011. Since then, a series of meetings have been convened, with each meeting after the first slowly but steadily formulating an increasingly detailed and coherent vision of the Cooperative.

The following is a summary of the meetings and outcomes to date:

2012

In May, eighty-seven archivists, librarians, curators, scholars, and representatives of funding agencies and foundations met for two days at the National Archives and Records Administration (NARA) in Washington, D.C. The meeting was largely educational. While those attending were enthused with the research benefits demonstrated in SNAC, and generally understood the rationale for a cooperative program to sustain and expand the benefits, many had difficulty understanding exactly how a cooperative would technically work. Nevertheless, there was general consensus around that idea that a cooperative would be based on professional editors with computer assistance. There was also concern about the potential benefits and impact participation in a cooperative would have on workloads. Finally, those gathered embraced the idea that "professional" would be broadly conceived, to include librarians and archivists, but also scholars with complementary and overlapping interests.

In October, twenty-seven archivists, librarians, and leaders from the archive, library, and scholarly communities met at NARA in Washington, D.C. While this meeting was intended to make progress on planning, again it became largely educational, explaining the rationale for a cooperative and how it would work, in general. Nevertheless, there were some important developments: Laura Campbell explained how the National Digital Information Infrastructure and Preservation Program (NDIIPP) and the International Internet Preservation Consortium (IIPC), two cooperative programs at the Library of Congress, were established. Key foundational requirements were identified. Don Waters suggested that it would be better to proceed incrementally, beginning with a pilot phase of implementation that focused on basics, and then, building on this, develop a full sustainable business, governance, and technology infrastructure. It was also agreed that it would be more productive to hold a series of meetings with small groups.

2013

In January, Clifford Lynch (Coalition for Networked Information), Don Waters (Andrew W. Mellon Foundation), Anne Van Camp (Smithsonian Institution Archives), Laura Campbell (Library of Congress, retired), David Ferriero (Archivist of the United States), Deborah Wall (Deputy Archivist of the United States), William Bosanko (COO NARA), Michael Wash (CIO NARA), Pamela Wright (Office of Innovation NARA), and John Martinez (Office of Innovation NARA), met at NARA in Washington, D.C. At this meeting, NARA *committed* to hosting the administration and governance of the cooperative, with the technological infrastructure being developed in collaboration with though outside NARA. Following the meeting, Ferriero appointed Martinez to collaborate with Pitti in moving the cooperative forward.

In June, Lynch, Van Camp, Campbell, and Laine Farley (CDL) met with Pamela Wright, Martinez, and Jerry Simmons at NARA. This meeting focused on a range of administrative and governance issues, with some issues that needed to be addressed: confirming NARA as administrative/governance host with outside organization serving as technical host; a formal charter (mandate) for NARA to host the cooperative; intellectual property and data ownership policies, inaugural partners and types of partners; and others.

In July, Pitti, Martinez, Lynch, Waters, Campbell, Farley, Ray Larson, Jerry Simmons, Brian Tingle, Terry Catapano, Carol Lagundo (NARA), and Mai Bui (NARA) met at the Coalition for Networked Information (CNI), Washington, D.C., to discuss the cooperative technical infrastructure. Discussion focused on maintenance-editing interface, with a specific consideration of the feasibility of using the SNAC technical infrastructure developed at SI/UC Berkeley as the foundation of this essential component. While the group was able to get a better understanding of the Berkeley infrastructure, it became clear that it would require more research, including in particular how the infrastructure could be revised to support professional editing procedures in addition to the batch processing already in place. The group also agreed that more planning was necessary before embarking on an implementation phase.

In October, Pitti, Larson, Lynch, Tingle, Yiming Liu (SI/UC Berkeley), Tom Laudeman (IATH) Mark Matienzo (Yale/ArchivesSpace), Trevor Thornton (NYPL), and William Stingone (NYPL) met at CNI, Washington, D.C.,

to focus again on the technical infrastructure. Quyen Nguyen, John Martinez, Jerry Simmons, Mai Bui, and Carol Lagundo from NARA were scheduled to participate but were unable to attend the meeting because of the U.S. government closure. The group again focused on Berkeley infrastructure, with Larson and Liu providing a detailed description of the underlying technology. The group was able to isolate a key piece of the configuration, "identity clusters," the constellation of data that represents the convergence of matching records and the union of the unique data therefrom that is serialized in the EAC-CPF records that serve as the data in the public interface. The consensus was that the infrastructure would be well suited to serve as the foundation for the maintenance system, but that the configuration of the "identity clusters" would need to be considerably expanded to accommodate real time updating of records in addition to the batch loading of records, as is now accommodated in SNAC. It was also agreed that the underlying platforms PostgreSQL and Cheshire were the best options, though the latter should be further evaluated, in particular with respect to its documentation, to ensure that it could be efficiently and effectively maintained outside of its development environment.

Merging and Redesign of SNAC and Cooperative Web Sites

From the beginning of the cooperative planning process, it was conceived that a Cooperative would be based on the data and technological results of the SNAC research and development. But because the two activities were funded from different sources, two web sites — one for the planning, one for the research and development — have been maintained. This separation of the two has made it difficult for the community to understand the complementary nature of the two activities, and so it was decided to bring the two activities together under the SNAC "identity," as it is the more widely recognized of the two. With the assistance of a professional designer, the two sites are now are currently being merged: Research and Development, and Cooperative Program Planning. The redesign will also be applied to the prototype access system.

Rationale for Cooperative Program Planning

Scholars interested in the lives of significant historical persons, their work, and the events in which they participated use as primary evidence the historical records that document their lives. These records are held in archives and manuscript libraries, large and small, around the world. The SNAC project is addressing the longstanding research challenge of discovering, locating, and using these *distributed* historical records and, at the same time, building a new resource that provides access to the socio-historical-intellectual contexts within which the records were created.

Though the international archive community has made great strides since the advent of a communication standard for archival description (Encoded Archival Description (EAD)) and the subsequent emergence of a limited number of sites that give access to state or regional holdings, scholars must still search scores of different archival access systems to find all of the records relevant to their research. This means that they need to know in advance where records are *likely* to be found. The process involves painstakingly accumulating and following clues, hunches, and leads. It is a frustrating and time-consuming process. Employing such methods, it may easily take a scholar years of persistent, focused work to locate resources relevant to her research, with a high probability that many clues will be overlooked, "buried" as they are in complex and detailed archival descriptions.

As a fundamental part of their work, archivists and librarians that process and describe primary resources document the creators and others with whom the creators lived and worked. Traditionally this description has been intertwined with the description of the resources, and as such lies largely hidden and disconnected from the vast social-document network within which the lives and resources that document them existed. In addition to rendering the vast network of interconnections obscure and disconnected, traditional practices are inefficient and costly, as professionals work in isolation from one another to establish the identities of the same persons, organizations, and families, duplicating this labor-intensive work. Since the 1960s, archivists have argued for *separating* and *interrelating* the descriptions of people documented in resources from the description of the resources themselves in order to make the description more flexible and useful, and to economize by sharing one another's work.

The SNAC project from its inception has had a complex set of interrelated objectives. The research and development activities were intended to demonstrate that the description of people embedded in archival description could be successfully separated from the description of the resources, interrelated with one another, and pooled together to produce a research tool that dramatically transforms research based on primary resources,

offering the efficiency of integrated access to the resources distributed throughout the world and new insights into the social-historical contexts of the lives and events documented. The SNAC project has been well received by archivists, librarians, and scholars. Public presentations on and demonstrations of the SNAC, and presentations on the proposed Cooperative in a variety of professional venues have been well attended and have generated enthusiastic discussion of the potential benefits both for professional archivists and librarians, but also for researchers and academic educators served by them. The SNAC Advisory Board, comprising both professional archivists and scholars, was enthused about the potential of SNAC to have a major impact on the processing and description of archives, and a transformative impact on historical research (See Appendix 2 for board members statements). SNAC has been well received both in the United States, and internationally, with invited presentations given in Paris, Strasbourg, Stockholm, Berlin, and Brisbane. The British Library, Bibliothèque nationale de France, Archives nationales (France), Staatsbibliothek zu Berlin, and APEx (Archives Portal Europe)[5] have all expressed interest in participating in SNAC. Based on the successful realization of research and development and the enthusiastic reception by both the end-user community and the professional archive and library communities, SNAC now has as its objective the establishment of a sustainable Cooperative of archivists, librarians, and scholars that will collectively maintain, improve, and extend the vast amount of data extracted and assembled in the SNAC project (nearly three million and possibly more individual descriptions of persons, organizations, and families). The Cooperative will benefit scholarly users of the primary resources, offering markedly improved efficiencies in locating resources and at the same time rendering explicit the social-document network that currently lies hidden. At the same time, the Cooperative will create new efficiencies in the description of the resources through the sharing of data and new opportunities to enhance access to and understanding of primary resources.

As compelling as the professional and scholarly research benefits are, it is abundantly clear that the transformative potential the resource tool offers researchers and professionals cannot be achieved using only computer-based techniques. As noted earlier, the quality of the data in existing archival description is uneven, and many potential interrelations go undiscovered. In order to fully realize the benefits for both end-users and the professionals that serve those users, the data needs to be cooperatively built, maintained, enhanced, and shared, to improve its accuracy and reliability, extend and develop the social and document interrelations, and integrate and interrelate additional descriptions. A Cooperative will allow archivists, librarians, scholars, and eventually interested communities of end-users both to curate and use the data.

Related Projects

The nearest analog to the proposed cooperative is the Name Authority Cooperative (NACO), administered by the Library of Congress. NACO is an international library name authority file. While NACO is very similar to what is proposed, library name authority work differs fundamentally from archival identity description. Library name authority identifies various names used by a person or organization, establishes a single name to be used in cataloging, and makes references from alternative names. Archival identity description involves much more than authority control as such. For archivists, identity description plays a major role in establishing the historical context in which the primary resources were created and used. Thus archivists, in addition to the control of the various names used by and for an entity, provide biographical-historical information (occupations, functions and activities performed, prose or chronological list histories, and more) describing the entity. Further, archivists also interrelate in detail the named entities, providing a macro-context for the resources. This detailed description is considered fundamental to archival description as it provides the historical context for understanding the origins of the primary resources, and reliably interpreting them. A representative of the Library of Congress NACO program has been involved in the preliminary discussions concerning the Cooperative, and a core objective of the Cooperative is to work in collaboration with the NACO program.

There is also an increasing number of scholarly prosopographical projects that are focused on particular geographic areas and times. The People for the Founding Era (PFE) is of particular interest. PFE is a project documenting the people of the founding era of the United States and is based on several documentary editing projects. Susan Perdue, PI for PFE, is on the advisory board of SNAC, and there have been preliminary discussions concerning involving the documentary editing community in the Cooperative.

[5] http://www.apex-project.eu/

There are several library and one museum authority control programs that involve the aggregation of authority data. The Virtual International Authority File (VIAF), administered by OCLC, aggregates library name authority records from throughout the world. The aggregated library authority records, however, are not maintained centrally, but merely gathered together. The Getty Vocabulary Program maintains the Union List of Artist Names (ULAN). The museum name authority records are similar to the archival identity descriptions in that they contain detailed descriptions of the named entities, though not in the detail provided by archivists. ULAN is maintained by the Getty staff rather than cooperatively by the community.

It is worth noting that VIAF data contributed by OCLC plays an important role in SNAC, as vast amounts of VIAF aggregated name control data (names and alternative names) are used in the match-merge processing. Alternative names, life or existence dates, affiliated countries, and languages used are added to the SNAC data from the matching VIAF records. Further, "sameAs" links in the EAC-CPF records are made to the VIAF record, Wikipedia, WorldCat Identities, and LCNAF records. Because NACO and ULAN both contribute to VIAF, data from these two sources also is interrelated to the SNAC data.

Project Description

The objective of the planning project is to develop a comprehensive plan for establishing a sustainable Archival Cooperative Program hosted by the U.S. National Archives and Records Administration that will aggregate, interrelate, and maintain archival descriptive information in collaboration archival repositories of all kinds within and outside of the government. Establishing a sustainable cooperative program hosted by a U.S. federal agency is complex and involves an array of legal, business, administrative, governance, and technological challenges. The planning will therefore focus on a pilot implementation by a small number of inaugural members. The pilot implementation will provide a firm foundation and will enable the participants to fully work through challenges and establish a sustainable business, administrative, governance, and technological infrastructure. The pilot implementation of the Cooperative will begin April 1, 2015, following the end of the planning grant period.

John Martinez (NARA), Daniel Pitti (UVa), Laine Farley (CDL), and Laura Campbell (LoC, retired) will coordinate and provide overall leadership of the planning activities, and in this capacity will constitute the Planning Coordination Team. Two teams will focus on two principal areas: administration and governance, and technological infrastructure. Campbell will lead the administration and governance team and Tom Laudeman (IATH) will lead the technical architecture team. In close collaboration with the two teams, Rachael Hu (CDL) will lead the user-centered requirements gathering process which will include conducting a user study with two user segments as well as creating requirements and specifications that result from the study. The studies will focus on two types of users: 1) archivists, librarians, and scholars who will actively engaged in creation and maintenance of identity descriptions; and 2) researchers and academic educators who will constitute the primary, but by no means only, users of the public access system.

Administration and Governance

The Administration and Governance Team will develop foundational documents that will serve as a starting point for launching the Cooperative. The AGT will focus primarily but not exclusively on the following activities and outcomes:

- Developing a charter that will establish the Cooperative within the legislated mandate of NARA. This activity will be conducted in consultation with the General Counsel of NARA.
- Developing Intellectual Property (IP) and data ownership policies. This activity will be conducted in consultation with the General Counsel of NARA. Marybeth Peters, retired U.S. Registrar of Copyright, will serve as a consult to this team, investigating the IP issues associated with the data that will constitute the primary asset of the Cooperative, both with respect to the data sources as well as uses of the data by others. The policy objective is that the maintained data will be owned by the Cooperative and made freely available without use restrictions and without substantive risk to the Cooperative.
- Developing a pilot administrative structure with responsibilities shared by NARA and UVa, with the long-term goal of the responsibility to be fully assumed by NARA after the pilot phase.

- Developing a pilot governance structure, drawing from inaugural members of the Cooperative and representatives of communities from which members will be recruited in the future, for example, the documentary editing community. In coordination with other foundational documents, draft by-laws covering including officers; committees for policy, steering, operations and communications; selection or election of representatives; reporting requirements; levels of membership (if appropriate); rights and obligations of members; and meeting structures.
- Develop a mandate for an advisory board and recruit inaugural members drawn from the research, genealogy, and education communities, as well as representatives from initiatives and programs with the potential for data sharing partnerships, such as the Virtual International Authority File (VIAF), the Digital Public Library of America (DPLA), Library of Congress Name Authorities Cooperative (NACO), Archives Portal Europe (APEx), and Europeana.
- Develop Cooperative promotional documents that describe the benefits of participation for professional archivists and librarians, and scholars; and develop promotional documents that describe the benefits the Cooperative for researchers, educators, students, and the interested public.
- In collaboration with the Society of American Archivists' Education Program, develop a plan for training professionals that will maintain the Cooperative data.

In addition to the work described above, the AGT will also begin the process of developing a long-term sustainability model that best suits the goals and characteristics of the Cooperative and its participants. The team will develop a framework of what specific questions and answers need to be addressed in four critical dimensions of sustaining a cooperative organization. This includes looking at the financial environment, organization and operational needs, ongoing technology requirements, and culture of the participants in the proposed Cooperative. Among other possible areas to be investigated will be the following:

- Economics – what are the financial requirements of supporting a long-term cooperative resource, the potential sources and uses of funds, and successful models for membership contributions?
- Organization and Operation – what are the organizational options, and the administrative and operating costs, legal considerations and impacts?
- Technology Requirements – what are on-going technical requirements, and the costs of attracting and retaining appropriate expertise needed for updating and maintaining the technology overtime?
- Culture – what environment will be effective in attracting and motivating collaborators and users, what environment will enable or allow for creativity and innovation, how broad does the participant experience need to be terms of such activities as education, forums for synergistic work or encouraging and promoting new and related research that benefits the cooperative?

Answers to these and other questions will help focus the business and operational assumptions that will be tested in the pilot and help shape sustainability recommendations resulting from the pilot phase.

In order to facilitate Laura Campbell working closely with NARA staff, NARA will provide her office space. This will enable her to spend sustained time sharing her expertise in establishing cooperative initiatives within the context of a U.S. federal agency.

Technical Architecture

The Technical Architecture Team will focus primarily though not exclusively on the following activities and outcomes:

- Continue the evaluation of the matching and merging platform developed by Berkeley and develop detailed specifications for revising the architecture to enable it to accommodate both batch loading of EAC-CPF records and record-by-record editing by archivists, librarians, and scholars. While the technologists involved in the development of SNAC in consultation with others have determined that the PostgreSQL and Cheshire components of the Berkeley-developed system are viable for serving as the foundation for the Cooperative maintenance platform, the PostgreSQL component will need to be revised to accommodate record-by-record editing. The Cheshire Information Retrieval (IR) plays a critical role in Identity Resolution,

that is, in determining whether two records with the same or similar names represent the same or different persons. Cheshire enables the creation of multiple specialized indexes of over twenty-five million VIAF authority records that contain both authorized and alternative names for persons, organizations, and families. The multiple indexes permit fast retrieval of candidate matching records based on different specialized search techniques, for example, exact string, Boolean, probabilistically ranked keyword, and NGRAMs. Both the processing speed and the specialized search capabilities are essential in match processing. While Cheshire is a robust and extremely powerful IR system, developing specifications for improving the efficiency of the NGRAM processing (a new feature added for SNAC), further code documentation, and perhaps some revision, to ensure that it can be efficiently and effectively maintained outside of the research environment within which it was developed. Included in the design specifications will be the development of an API (Application Programming Interface) that will support a standalone client editor, as well as clients embedded in other applications, such as ArchivesSpace.

- Development of specifications for a standalone client editor that will serve as the interface for professionals performing record-by-record maintenance. The results of the professional user requirement study will inform the development of the specifications.
- Development of specifications for revising ArchivesSpace in order to identity description within the archival management tool can both draw from and contribute to Cooperative data.[6] The results of the professional user requirement study will inform the development of the specifications.
- Development of specifications for an administrative client for the maintenance system, including management of users, authentication, and permissions.
- Development of specifications for the public access system. The end-user requirement study will inform the development of the specifications.
- Evaluate each of the three principal technology components developed in SNAC research and development (Extraction, Match/Merge, and Public access and resource system), and determine the most efficient computational methods for managing the now manually assisted data flow between the components.

User Requirement Studies

The user study work in phase two of SNAC (described above) was targeted primarily at public access interface users to determine initial problem areas with the internal conceptualization of the prototype. The user study work in the in the planning phase will be used to validate the changes made to the public interface based on the findings made in the phase one study, and suggest additional changes that may still be needed. It will also expand the diversity in the initial researcher user base recruited, and thus provide an opportunity to reveal new requirements. It will also focus on uncovering the description and processing workflow practices of archivists. The user study would form the basis for requirements and specifications in the creation and design of the archival maintenance interface, which is the crucial data contribution point for the ability to amass the records necessary to form the foundation for the Cooperative.

Two user requirement studies will be conducted in the planning phase. The first will focus on researchers and academic educators and thus address public interface requirements. Three types of end-users will be targeted in this study: historians, academic educators, and genealogists. The second study will target librarians and archivists that will edit (create, revise, merge, split, and interlink identity description records). This study will focus on archival processing workflows and descriptive practices, to ensure that the all editing functions are necessary and configured in a manner that addresses the well established steps involved in establishing and describing organizations, persons, and families. The end-user studies will be conducted at a number of venues, such as professional conferences, at the National Archives and Records Administration, and at the headquarters of a major genealogy society. Potential venues include: American Association of State and Local History (St. Paul, MN); Association of Documentary Editing (Louisville, KY); American Historical Association (New York City); and the New England Historic Genealogical Society (Boston). The professional editor user studies will be conducted *in situ*, in libraries and archives located in the same cities as the conferences listed above, as well as in conjunction with the annual conference of the

[6] ArchivesSpace is an archival management system that is increasingly used in the archival community. Among the management tasks supported in ArchivesSpace is archival description, including context and identity description that will benefit from a shared cooperative maintained resource.

Society of American Archivists (Washington, D.C.). The results of the user study will inform creation of user-centered requirements and specifications for the maintenance system as well as the public access system. These requirements will be developed in close collaboration with the technical team as they work through technical architecture specifications for the same system components.

The methodology for the two user studies consists of three phases: planning/recruitment; field research; and analysis/reporting. The planning phase will involve refinement of the methodology plan, drafting of research instruments, application for human subjects approval, and recruitment and planning for site and conference visits.

The second phase of the user studies will consist of field research focused on understanding user practices surrounding public interface research and archival processing workflows, descriptive practices, and record creation. For this phase, we have selected a structured qualitative method – qualitative so that we have the amplitude of a personal interaction to gather a holistic view of how archivists describe and create special collection records within the full context of their processing workflow and structured so that we can extrapolate results, findings and requirements from the captured data. The qualitative method will also provide the same flexibility in data gathering with users of the public interface for EAC records. The anticipated data gathering methods will include structured interviews and observations of archivists and researchers that will provide repeatable and reliable inquiry methods. The anticipated data capture method within these interviews will include observation and note taking. During the planning phase for this project, we will experiment with methods of data capture including video or still screen capture of the user's physical or virtual working environment. In-person engagement between interviewers and interviewees, especially for the archival maintenance and data entry interface, is a critical aspect of this project, allowing interviewers to follow up more immediately and capture more explicitly the nuances of the interviewees' practices that we would be unable to capture in a virtual user engagement. We anticipate that these user engagements will last between one or two hours per user.

Concluding this phase of work, we will analyze findings. In our final phase, a final report and requirements document will be developed that will summarize findings, structure them into groupings of behavior, create requirements and initial wireframes for future development of the archival maintenance interface as well as further refinement of the public access interface, and reflect on the user study process, undertaken.

Collaborators

IATH will be the lead institution in the project,[7] and will collaborate with the National Archives and Records Administration (NARA); School of Information, University of California, Berkeley (SI/UCB); and the California Digital Library, University of California (CDL). LYRASIS (representing ArchivesSpace) and the New York Public Library will also be participating. Laura Campbell, retired Chief Information Officer at the Library of Congress, will also play a pivotal role.

As the lead institution, IATH will be responsible for overall management of the project. In addition, IATH will coordinate the organization of all meetings with the Planning Coordination Team (PCT). IATH will be responsible for arranging accommodations, meeting rooms, and processing travel reimbursements. IATH will also, in consultation with the PCT, oversee the writing of all planning documents and provide editing assistance in the drafting of planning documents.

NARA, as the host for the Cooperative, will participate substantively in the planning. NARA will contribute expertise in program operations, archival processing, technology (in particular knowledge of the technological infrastructure of NARA), and legal counsel for establish a charter for hosting the Cooperative, and development of the Intellectual Property and data ownership policies.[8] A NARA representative, John Martinez, will serve as one of the Planning Coordination Team, and NARA will participate on both the Administration/Governance Team and the Technological Architecture Team.

[7] See Appendix 3 for backgrounds of the collaborating institutions.
[8] While NARA will not host the Cooperative technological platform, NARA processing and public access systems will need to interoperate.

SI/UCB will provide expertise on the SNAC matching and merging processing platform that is the chief candidate for serving as the foundation for the maintenance-editing component of technological infrastructure of the cooperative. SI/UCB will also train CDL staff in installing and maintaining the two principal components of the matching and merging infrastructure: PostgreSQL and Cheshire, in order to ensure that maintenance responsibility for the platform can be successful transferred. It is assumed that "rehearsing" will reveal any major issues that need to be addressed before transitioning to an implementation phase. SI/UCB will participate in all meetings of the Technological Architecture Team.

CDL will be responsible for several activities. Laine Farley, Director of CDL will serve as a member of the Planning Coordination Team. CDL will also conduct two user requirement studies, one focused on professional (librarian, archivists, and scholars) editors, and the other on end-users of the public access system. This work will contribute directly to the development of the technical infrastructure detailed specifications. CDL, as the leading candidate for hosting the cooperative technical infrastructure, will also participate in rehearsing the transfer of the Berkeley matching and merging platform to CDL, bringing it together with the public access system. CDL will participate on both the Administration/Governance Team and the Technological Architecture Team.

ArchivesSpace will also collaborate in the planning, with the Chief Technical Architect participating particularly in developing specifications for incorporating an interface within ArchivesSpace for interfacing with the Cooperative maintenance system in order that the benefits of shared identity description can be incorporated into ArchivesSpace. The Curator of Manuscripts at the New York Public Library, will contribute his extensive understanding of all facets of archival processing, and an administrator's perspective on the benefits a cooperative will need to offer within research archive processing units.

Job titles and job descriptions

Principal Investigator – *Daniel Pitti* will serve as the Principal Investigator. He is Associate Director of IATH. Pitti will be responsible for overall management of the Cooperative planning, and will be one of four members of the Planning Coordination Team (PCT). [25%]

Co-principal Investigator – *Laine Farley* will serve as Co-Principal Investigator. She is Director of the California Digital Library, Office of the President, University of California. Farley will serve as a member of the PCT. [5%]

NARA Cooperative Planning Lead – *John Martinez*, Director of the Business Architecture, Standards, and Authorities Division, Office of Innovation, NARA. He oversees the maintenance and development of NARA's lifecycle data standards and authority files and maintains the Business Architecture component of NARA's Enterprise Architecture. During his 13 years at NARA he has worked on: developing NARA policies and procedures in the areas of lifecycle processes, standards, and systems; developing agency business rules for the records lifecycle; and served as liaison to various NARA lifecycle systems and data projects.

Administration/Governance Planning Lead – *Laura Campbell*, retired Chief Information Officer at the Library of Congress. Campbell led the development of two major international cooperative programs, the National Digital Information Infrastructure and Preservation Program (NDIIPP) and the International Internet Preservation Consortium (IIPC). In addition to leading the Administration/Governance Team, Campbell will serve as a member of the PCT. [36 days]

Technical Architecture Design Lead – *Tom Laudeman*, SNAC Data Extraction Programmer, IATH. Tom joined the SNAC project in 2011 and has developed and documented an open source software library of XSLT and Perl scripts for creating EAC-CPF records based on records submitted to SNAC. Tom has more than 15 years of experience as a software engineer and web applications developer and has worked with large scale domain-specific data repositories integrated with data analysis tools and computerized annotation. [30%]

Planning and Documentation Coordinator – *Sarah Wells*, Scholarly and Technical Communications Officer at IATH, will coordinate communication with planning participants and the assembling of planning

documentation, as well as oversee meeting planning. She has been part of the SNAC team since its inception. [15%]

Technical Architecture Team Members –

Ray Larson, Professor, School of Information, University of California, Berkeley. He specializes in information retrieval and database systems and is the principal designer of the Cheshire information retrieval system. His current research focuses on several related areas of information retrieval and digital libraries. He has been a core member of the SNAC project since its inception. [1 month summer salary]

Brian Tingle, Technical Lead, Digital Special Collections, California Digital Library, University of California Office of the President. He has been involved in designing and building web-based access systems for UC Libraries for the last 17 years, including the OAC and Calisphere. He has also been a core member of the SNAC project since its inception. [10%]

Brad Westbrook, ArchivesSpace Program Manager, and formerly the Project Manager of the Archivists' Toolkit. He was the lead author of the technical specifications for both ArchivesSpace and the Archivists' Toolkit. He has been greatly involved in modeling digital object description and rights assertions. Prior to becoming the ArchivesSpace Program Manager in June 2013, Brad was employed in the UC San Diego Library for twenty years, where he had oversight responsibility for the metadata analysis and specification team, was a member of the university's digital curation team, and participated in the CRL TRAC certification of the Chronopolis Preservation System. He is a UC Distinguished Librarian and a SAA Fellow. [10%]

Quyen Nguyen, Systems Architect, Office of Information Services NARA. His duties include enterprise system architecture, software development, requirements analysis, and systems engineering and integration. During his ten years at NARA, he has worked on electronic records preservation systems, NARA's Enterprise Architecture, and on the integration of records lifecycle systems. His current focus includes the incorporation of Cloud infrastructure into NARA's enterprise architecture.

Carol Lagundo, IT Project Manager, Office of Innovation, NARA. Her duties include project management, data standards, software development, and system development. In her 16 years at NARA, she has led projects to build or re-engineer NARA's ARC and OPA description and public access systems.

Administration/Governance Team Members –

William Stingone, Charles J. Liebman Curator of Manuscripts, New York Public Library. He has over twenty years of experience in archival repositories, including extensive experience performing and managing all aspects of archival enterprise: acquisitions/appraisal, collection management, archival description, and reference services. He conceived and led the implementation of archives.nypl.org (an online access system providing researchers with the ability to search across all archival descriptions at NYPL). [10%]

Adrian Turner, Data Consultant, Digital Special Collections, California Digital Library. He 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 and multi-year (2001-2005) Library of Congress-supported California Cultures project. He has also been a core member of the SNAC project since its inception. [10%]

Jerry Simmons, Archives Specialist for Data Standards, NARA. Since January 2000 he has been the Authority Cataloging Team Lead for NARA's Archival Research Catalog (ARC) project and coordinates NARA's NACO and SACO efforts. Prior to his work at NARA, he was the archives cataloger for the United States Holocaust Memorial Museum in Washington, D.C.

User Experience Specialist – *Rachael Hu*, User Experience Design Manager, California Digital Library, University of California Office of the President. She manages and facilitates the discovery and design process for online services and tools produced by CDL and leads the requirements gathering and the design and specification of the user experience for these services. [45%]

Intellectual Property Consultant – *Marybeth Peters*. She is a lawyer and the retired United States Registrar of Copyrights and acting general counsel to the U.S. Copyright Office. She has experience in establishing Intellectual Property rights policies for two international cooperative initiatives at the Library of Congress: NDIIPP and IIPC. Peters will work with the NARA General Counsel Office, Laura Campbell, John Martinez, and the Administrative/Governance Team. [160 hours]

Principal Investigators

Principal Investigator – Daniel Pitti will serve as the Principal Investigator. He is Associate Director of IATH, the chief technical architect of both the EAD and EAC-CPF standards. 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.

Co-principal Investigator – Laine Farley will serves as Co-Principal Investigator. Farley is the Executive Director of the California Digital Library, Office of the President, University of California. She provides leadership to the CDL and determines its strategic direction. She is responsible for the organization's overall management, including planning, policy development, and resource allocation. She has overseen CDL's leadership of or participation in several large collaborative projects: HathiTrust, Western Regional Storage Trust, Public Knowledge Project, and DMPTool, among them.

Length of Project with Timeline

The project will last from April 1, 2014, to March 31, 2015.

There will be three face-to-face meetings for both the Administration/Governance Team (AGT) and the Technical Architecture Team (TAT), in May 2014 (at NYPL), July 2014 (at NARA), and February 2015 (at CDL). The Planning Coordination Team (PCT) and the Planning and Documentation Coordinator (PDC) will attend all meetings. The team meetings will last for 1 ½ days for each team and will be scheduled in tandem in order to minimize travel for the members of the PCT.

User Experience Specialist (UES) will conduct user requirement studies, with the assistance of the CDL Data Consultant, at three conferences and at archives and libraries in the vicinity of the conferences. The three conferences are American Association of State and Local History (St. Paul, MN); Association of Documentary Editing (Louisville, KY); and the American Historical Association (New York City). UES will also arrange to conduct user requirement studies with two to three genealogical researchers at the New England Historic Genealogical Society (Boston).

In addition to the face-to-face meetings, the PCT, AGT, and TAT will hold regular conference calls (every other week), and use email for regular, ongoing communication and coordination.

Pre-grant: January – March 2014

UES, in consultation with others, will prepare formal documentation describing the user studies for Institutional Review Board (IRB) approval. The documentation will be submitted to the University of Virginia IRB, with the objective of having approval in place by the start of the grant period.

Archived mailing lists will be set up for each team. The existing SNAC listserv will be used discussions among the entire planning project staff. A method for sharing and collaborating on planning documents will be set up.

April – June 2014

PCT with the leaders of the AGT and TAT will review the planning objectives and overall plan of work.

In consultation with NARA's Office of General Counsel, PCT will draft charter document. The charter document will be completed and approved by the end of the first quarter.

PCT will provide Intellectual Property Consultant (IPC) with overview of IP policy objective. IPC will investigate copyright and data ownership issues associated with data contributed by Cooperative members, and data contributed by third parties, such as OCLC. IPC will investigate potential copyright issues associated with biographical and historical data in archival descriptions. IPC will interview a small representative group of SNAC contributors to understand context of rights and ownership issues associated with current data sharing agreements, and analyze other cooperative programs (such as the Program for Cooperative Cataloging (PCC)) with regard to use and reuse policies. Analyze the benefits and risks of different data policy approaches, for example, using common licensing versus developing a policy specific to the Cooperative. IPC will consult, as necessary with the PCT. IPC will prepare a report for the PCT and AGT for the July meeting.

In May, UES will conduct a user requirement study with genealogical researchers at the New England Historic Genealogical Society, Boston. (The study will be arranged with the assistance of David Rencher, Chief Genealogist at the Church of Latter Day Saints and, Ryan Woods, the COO of the Society).

AGT. The focus in the first quarter will be on working with NARA Office of General Counsel on charter document, as well as review of and providing input to the IPC with respect to IP and data ownership. AGT, breaking into small teams, will begin developing 1) administrative plan for pilot phase of Cooperative, with a focus on coordinating and sharing administrative duties among an executive team that includes NARA staff and the PI and Co-PI; 2) governance plan that will focus on the establishing the scope of the duties of a governance committee or committees, and the structure of the committee or committees; and 3) investigating sustainability requirements and review of business and operational models of comparable cooperative initiatives. With respect to governance, AGT will investigate comparable cooperatives such as the PCC, and attempt to adapt proven governance models to the Cooperative.

TAT. The team will focus on the two major technological components that will comprise the full technological infrastructure of the Cooperative: maintenance system (including editing interface), and public interface. The team leader will assign two or three members of the team to focus on investigating and evaluating existing SNAC infrastructure, evaluating existing software components, determining functional requirements, and determining what software development will be necessary to extend the existing functions to encompass functions that will support real-time editing of Cooperative data while also accommodating existing batch import modes. Preliminary work is already completed in this area, but extensive work needs to be performed to determine in detail complex functional requirements, particularly of the PostgreSQL database that serves as the primary means to record data, to track and manage match/merge data, and to stage and finally serialize identity description records that serve as the data in the public access system and resource. The team will also focus on evaluating the existing public access system.

Meeting one: May, NYPL. AGT and TAT will each meet for 1 ½ days. PCT will attend both the AGT and TAT meetings. Meetings will be devoted to discussing preliminary research and plan development underway, and refining the identification of research and development areas, developing strategies and work assignments for ensuring that the necessary administrative and governance planning and foundational documents are developed in a timely manner.

July – September 2014

PCT will provide IPC with response to preliminary report and IPC will draft IP and data ownership policy for Cooperative. IPC will also prepare a draft agreement or agreements for data contributing inaugural members of the Consortium.

Association of Documentary Editing (Louisville, KY) July. UES will conduct user requirement studies with two to three documentary editors. UES will focus on documentary editors as end-users of the Cooperative data, but also explore interest in documentary editors as contributors of data, in that identity description is a fundamental part of their professional activities.

Meeting Two: July, NARA, Washington, D.C. Charter document will have been completed and approved by NARA Office of the General Counsel. IPC will meet with the PCT, AGT, and TAT to present findings of research, explore issues that the various planning staff may not have been considered or need further

consideration. The major research findings of the AGT and TAT will be presented for discussion and refinement. Both the AGT and TAT planning should be well underway, with initial drafts of all major components. Areas needing further research and development will be identified, and work assignments and deadlines made. Preliminary results from user requirement study with genealogists in Boston will be reviewed to inform PCT and both teams. Foundational members of the Cooperative and potential members of the advisory board will have been identified by the AGT in anticipation of initiating the recruitment process, and securing commitments and agreements necessary. All foundational members of the Cooperative will be contributors of data. A subset of the members will also be asked to commit the time of processing archivists and librarians to participate in the testing and use of the editing interface, revising and merging existing and adding new records.

PI, with assistance of others as needed, including IPC, will begin negotiations with OCLC for the right to use data derived from Virtual International Authority File records and WorldCat archival descriptions, with agreements to be finalized by the third quarter of the grant period.

Society of American Archivists (Washington, D.C.) August. UES will conduct user requirement studies with processing archivists and librarians at NARA, the Library of Congress, and the Smithsonian Institution. One to two regular researchers at each of the three federal repositories will be recruited for end-user studies.

American Association of State and Local History (St. Paul, MN). September. UES will conduct user requirement studies with two to three historians. UES will also conduct user requirement studies with two to three processing archivists and librarians at the Minnesota Historical Society, and if needed, at the University of Minnesota.

PI and PDC will begin the process of recruiting the foundational members of the Cooperative, with assistance from other planning staff as necessary.

October – December 2014

Recruitment of foundational members and negotiations for data use rights with OCLC will continue.

Work on all fronts will continue, with ongoing progress evaluation and any adjustments in assignments, schedule, and focus of work will be made. Preliminary findings from the documentary editors, archivists, librarians, and historians user requirement studies will be communicated to the PCT and both teams, and findings will be incorporated into the draft planning documents as appropriate.

In anticipation of submitting a proposal to the Mellon Foundation requesting funding for implementation of a pilot phase, full drafts of administration, governance, and technical specification documents will be completed by October 15. While the draft documents will not be fully completed, the drafts should be sufficiently detailed and thorough in scope to enable work to commence on the development of the proposal, covering each of the three primary areas. IPC will complete IP, ownership, and data contributor rights agreement template. The agreement template will enable the PCT to begin recruiting Cooperative charter members and securing the necessary agreements.

By October 15, AGT and TAT leaders will submit full drafts of all planning documents.

In December, with input from all planning staff, the PCT lead by the PI will develop a detailed proposal to the Mellon Foundation requesting funding for launching a pilot phase of the Cooperative, with an expected launch date of April 1, 2015.

January – March 2015

IPC will be available for any additional IP and ownership issues that may arise as the Cooperative prepares to launch the pilot phase.

UES will complete the final user requirement study, at the American Historical Association (New York City). January. UES will conduct user requirement studies with two to three historians attending the conference, and will also conduct studies with two to three processing archivists and librarians at NYPL.

Meeting Three: CDL, February. The final meeting will focus on 1) review of complete drafts of all planning documents and final report of the UES; 2) identification of any further refinements and revisions necessary, along with assigning work needed; and 3) identifying and carrying final preparations for launching the Cooperative in a pilot phase.

Under supervision of the PI and PCT, and with the assistance of the PDC, all planning and foundational documents will be finalized, in preparation for the launch of the Cooperative. UES will complete final report of the user studies, and relevant findings will be incorporated into the final planning documents.

Expected Outcomes and Benefits

The planning process will produce a comprehensive plan for establishing an archive description and access cooperative hosted by NARA: a legal charter for the cooperative; intellectual property and data ownership policy; intellectual property data contributor agreement templates; detailed administration and governance plan for pilot phase (including strategic plan for developing sustainable business model); a long-term sustainability plan; and detailed technical architecture design specifications (including consideration of both professional-user and end-user requirements). In addition, the project will recruit and secure the commitments and agreements from charter members of the Cooperative, including members that will contribute data alone, and members that will contribute data and processing staff time to testing and using the professional-user maintenance interface. The PCT will also establish a mandate for and recruit members for an advisory board, drawing upon prominent researchers and scholars, archive and library community leaders, and representatives of programs with complementary programs.

The planning documents (drafts and final versions) as well as description of SNAC to provide context for understanding them will be made publicly available on the SNAC: Cooperative Program Planning web site. Availability of the documents will be announced on various web sites including the EAD listserv (archivists and librarians), and the CenterNet (digital humanists) listserv. In addition, participants recruited for the pilot phase implementation of the Cooperative will individually be made aware of the planning documentation. Finally, the SNAC team has submitted a session proposal for SNAC, including the Cooperative Program Planning, for the Annual Meeting of the Society of American Archivists in Washington, D.C. in August 2014.

The primary immediate benefit of the project will be that the foundational documents, detailed plans, and charter participants will be in place, and ensure that the Cooperative can be launched on a firm foundation that will enable efficient and effective implementation that leads to a sustainable Cooperative. The long-term benefit of the project will be providing a solid foundation on which to establish a Cooperative that will offer researchers dramatically improved efficiencies in locating primary resources, new insights and understandings of the historical contexts in which the records were created, and a novel aggregation of data that will open up new social-historical and historical social network research; and will enable professional curators of primary resources to more efficiently and thoroughly describe the resources and better serve the research communities interested in them.

After the first six months of the planning process, the University of Virginia in collaboration will submit a request for a two-year pilot phase implementation of the Cooperative to the Foundation, contingent on an invitation to do so. While NARA will host the Cooperative, providing staff support for administration and governance, the technical infrastructure for the project will be hosted outside of NARA, and will need funding in order to provide a foundation for firmly establishing a sustainable self-sustaining business and operational model. Placing the technological infrastructure outside of NARA will permit the development to take place without the costly overhead associated with government contracts. In addition, the pilot phase will provide the means for establishing a public-private partnership between NARA and the professional archive, library, and museum communities.

While it is anticipated that the University of Virginia will continue to be the lead institution until the Cooperative is firmly established, though over the course of the planning it may emerge that it is more appropriate and practical for another institution to assume the leadership role.

Intellectual Property Issues

The planning project will not produce software, though will produce detailed technical specifications for software development. The technical specifications document will be published with a Creative Commons Attribution-

ShareAlike (CC BY-SA) license. The project will also produce a charter document, and an Intellectual Property policy. These documents will be made available without copyright or use restrictions.

NARA General Counsel has been consulted in the preparation of this proposal and foresees no significant obstacles to NARA being able to sign an IP agreement conforming to Mellon Foundation policy.

The Cooperative Intellectual Policy developed in the planning will address the Foundation's policy and will be formulated in consultation with the Foundation.

Long-term Sustainability

The comprehensive plan for an archive description and access cooperative hosted by NARA is intended to serve as the basis for requesting funding from the Foundation for a pilot project that will establish the Cooperative. During the pilot phase of the project, the planning documents will be further refined and adjusted based on lessons learned in the implementation.

Investigating the requirements and developing a strategy for the long-term sustainability of the Cooperative, will be a major activity in the planning process, as discussed above in the Project Description.

Reporting and Evaluation

The project will submit a final report at the end of the project period. This will provide detailed description and evaluation of the project activities and documentation on grant expenditures. The final report covering project activities and grant expenditures for the grant period, will be submitted no later than June 30, 2015.
The final report will cover these activity areas:

- Complete drafts of all planning documents
- Recruitment of inaugural Cooperative members and advisory board members
- Administrative/Governance Team (AGT) and Technical Architecture Team (TAT) final reports.
- Final findings from user requirement studies.
- A financial narrative.

Financial reports will be provided in the form of Excel spreadsheets, augmented and annotated as necessary. The spreadsheet will provide, in relation to the budget, a line-by-line list of expenditures for the project. It will provide information on the investment of the funds, and the return on the investment.

Daniel Pitti, Principal Investigator for the project, will be responsible for soliciting the necessary data from the project partners and for assembling the final reports. In addition to the project partners, he will be assisted in the reporting by the Planning and Documentation Coordinator and IATH's administrative assistant.