Social Networks and Archival Context Cooperative: Pilot Phase

A Proposal to The Andrew W. Mellon Foundation

Project Overview

The Institute for Advanced Technology in the Humanities (University of Virginia) in collaboration with the U.S. National Archives and Records Administration (NARA) 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 international archival description and access cooperative. Funded by grants from the National Endowment for the Humanities (2010-2012) and the Andrew W. Mellon Foundation (2012-2015), SNAC has demonstrated that scholarly historical research can be substantively transformed by a research tool that 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. It became clear early in the project that the biographical data extracted and assembled from archival resource description constituted a valuable independent resource that could (and should) be maintained and further developed cooperatively. As a permanent maintained resource, it would not only benefit researchers but also contribute to the economy and efficacy of archival processing and description. Drawing on funding from the Institute for Museum and Library Services (2011-2013) and the Andrew W. Mellon Foundation (2014-2015) and critical assistance from key experts and institutions, the SNAC collaborators[2] have built broad community support for an archival cooperative, and developed a comprehensive plan for establishing it. From the planning discussions and activities, it has become clear to all involved that establishing a sustainable cooperative is exceptionally socially and technologically complex, and thus an incremental, phased approached is essential. Given the complexity of the undertaking, we propose to fully establish the Cooperative over a four-year period, beginning with a two-year pilot phase. We are requesting $999,466 for the two-year pilot phase, in order to establish the initial Cooperative legal, administrative, and governance infrastructure; to begin transforming the research and demonstration technological platform to support ongoing batch and cooperative manual maintenance; and to explore sustainable business models that will enable NARA to assume full administrative responsibility for the Cooperative, including budgetary authority.

Project Justification

Converting SNAC into a permanent cooperative resource and history research tool will be a substantial contribution to the scholarly research and communication economy, making the description of archives more efficient and effective, and significantly ameliorating the challenge of discovering, locating, and understanding the resources that document our shared history. **Scholars** who depend on historical manuscripts and records, and the archivists and librarians who curate, describe, and provide access to them, have been uniformly enthusiastic about SNAC's results to date and its potential to offer greater results through cooperative building and maintenance. **Researchers** have welcomed SNAC for its research economies: SNAC's Historical Research Tool provides integrated access to distributed primary (archival) and secondary (published) resources, eliminating or at least substantially ameliorating the need to track down resources in multiple archival catalogs. Painstakingly piecing together these networks constitutes an essential and time-consuming historical research activity. At the same time, SNAC makes explicit what has been, at best, implicit in archival description: the social-professional-intellectual networks within which the lives and work of the people documented in historical resources took place. SNAC, in essence, reveals a global social-document network that connects the past to the present. Ed Ayers, President of the University of Virginia and a Civil War historian, wrote that:

> SNAC promises to change the way history is imagined and written! For all that the digital revolution has revolutionized, the heart of research lies within the primary record embedded in archives large and small. The

[1] See project site http://socialarchive.iath.virginia.edu

[2] See appendix 7 for backgrounds of the collaborating institutions.

> 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), describes SNAC's potential:

> 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.

Archivists and librarians have embraced SNAC because it demonstrates that the descriptive data they create can be reconfigured and used to enhance discovery and understanding of historical resources in ways that were previously unimaginable. It also offers the possibility of sharing one another's work to realize complementary economic benefits, eliminating duplicative work and producing more effective description. Michael Rush, Processing Archivist, Beinecke Rare Book and Manuscript Library, Yale University; and Co-chair 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 collections 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.

RELATED PROJECTS

While there are cooperative authority programs that bear some similarity to the SNAC Cooperative, there is no program that addresses the specific descriptive requirements of the archival professional community. The existence of national and international library cooperative authority control programs (such as NACO and comparable national programs, and the Virtual International Authority File (VIAF)) make it reasonable to explore whether it would be feasible for the archival community to collaborate with the library community in the use of these existing programs rather than establish its own program. Many of the corporate bodies, persons, and families, after all, are common to the holdings of both communities, and with those of museums as well. Archival records, however, document the lives and work of people and organizations in all walks of life, not just the fraction of people who write books or have books written about them or create or collect objects found in museums. Farmers, business people, health professionals, teachers, social workers, community organizers, civil rights attorneys and workers, religious leaders, and many, many others are represented in archival holdings. Birth certificates, marriage certificates, military records, property records, and other types of records ensure that just about everyone is represented. The vast majority of the people documented in records are simply not present or associated with the holdings of libraries and museums. On March 11, 2015, 24.8% of nearly 3.5M SNAC identity descriptions matched VIAF cluster records. Of the SNAC identity descriptions matching VIAF, 32.2% are persons and 13.2% are organizations.[3] Archivists certainly do not and will not formally describe the majority of the people documented in their holdings, but many of them will be described, and many of these descriptions will be unique to the archival domain, and thus the sole responsibility of archivists.

Further, the centrality of describing the provenance of records leads archivists to describe individuals, families, and organizations in a manner different than in libraries. Common to both is authority control proper: recording the

[3] Family identities are not matched because the vast majority of them consist of a surname and the word "family," and the surname alone is insufficient for distinguishing unique families. Nearly 100% of exact string matches on families are "false positives."

names used by and for the entity, and designating which among them is to be preferred for facilitating access and display. As important as name control is, the archival imperative to establish the provenance of records involves documenting the *lives* of individuals and the *history* of families and organizations that created records, frequently in great detail. The archivists' objective is to describe the people documented in archival records in sufficient detail that users will be able to understand the *social context* in which the related historical records were created and used. Archival description of people is commonly far more elaborate than is accommodated by the brevity assumed in library authority control systems. Though it is conceivable that library authority control could be enhanced to accommodate the archival requirements, the library cooperatives exist first and foremost to serve the specific needs of the library community—as they should. There is no compelling motivation for the library community to enhance these systems to accommodate the rather complex descriptive needs of the archivists.

Archivists need a cooperative program that ensures that their primary professional curatorial responsibilities are fulfilled. Such a program needs be fully integrated into the management and processing of archival resources. While distinct from existing library cooperatives, the archival cooperative will necessarily collaborate with the library and museum communities, to align identities across the curatorial domains, share data with one another, and provide common pathways to resources across the cultural heritage spectrum.

In addition to providing a platform for archivists to share one another's research and description, the SNAC Cooperative will provide researchers with two unprecedented forms of access: *integrated* intellectual access to distributed primary (archival) and secondary (published) resources by and about persons, families, and organizations; and access to the social, professional, and intellectual networks within which the people lived and worked. The descriptions of people in SNAC provide additional access to resources through links to OCLC's ArchivesGrid, Digital Public Library of America (DPLA), and links to alternative descriptions of people in VIAF, LCNAF, and Wikipedia. The links to DPLA provide users with access to digitized cultural heritage resources, and this form of access will be enhanced further through links to Europeana, HathiTrust, Google Books, and to digital humanities thematic and prosopographical research collections.

The SNAC Cooperative will be a substantive contribution to the national, and indeed, international digital platform, making the description of archives more efficient and effective, and significantly ameliorating the challenge of discovering, locating, and understanding the resources that document our shared history. The SNAC Cooperative will not itself offer digital cultural heritage but rather descriptions of the people who created and are documented in the heritage, with links to both descriptions of cultural heritage (WorldCat records and EAD-encoded finding aids) and cultural heritage content aggregators such as DPLA, HathiTrust, Europeana, and others. As such SNAC will provide a unique pathway to the cultural heritage content through the descriptions of the people who created or are documented in that heritage. As with DPLA and Europeana, SNAC will provide metadata linked to cultural content served by others, but it differs in two fundamental ways. First, the core metadata being assembled and curated describes people and not the things people have created. This description of the people constitutes a resource in itself, resembling a biographical dictionary. Second, the description of the people provides a hub that integrates access to cultural heritage whether or not that cultural heritage is in digital form. What all cultural heritage resources have in common is that they are the product of people living their lives and working, and understanding the lives of the people provides context for understanding the resources.

PROJECT OBJECTIVES, STRATEGY AND WORK

Over the course of more than two years of planning to transform SNAC into a sustainable international archival cooperative, it has become clear to all participants that this is an *exceptionally* complex social, intellectual, and technological undertaking. Fully realizing SNAC's benefits will require a thoughtful step-by-step approach. While the inaugural participating archivists and librarians each bring essential knowledge and experience, it will take time to build shared understanding and competence that will equip the participants to perform the familiar in a new, unfamiliar environment and to assume responsibility for the governance of the cooperative. Transforming the current technological infrastructure, which builds and maintains the SNAC data algorithmically so that it will also support a team of human editors demands careful planning and adequate development cycles. The same can be said for building an effective administration and a self-sustaining business model. All aspects of community building, administration, technological development, adjustments in local systems, and so on, will necessarily be incremental. The result will be iterative, with each step informed and shaped by the preceding steps. Given the complexity of the challenge, we are beginning with a pilot phase, to be followed by a second phase that permits review, correction, and a more confident and capable implementation.

The objectives of the pilot implementation of the Cooperative and the activities essential to realizing the objectives can be broadly characterized as social and technological. The social objectives involve developing the knowledge and expertise of NARA's staff to assume administrative and political oversight of the cooperative, and developing a core group of inaugural members of the Cooperative with a shared understanding who can creatively participate and contribute to its governance and assist in recruiting and training new members. The technological objective is to effect a major transformation of the research and demonstration technological platform into a platform that will support ongoing cooperative maintenance of the SNAC description and access data. Given the complexity of both objectives, it is anticipated that the pilot will provide a solid, credible foundation on which the Cooperative can be fully established with an additional two years of effort.

Establishing an international cooperative program within the context of a U.S. federal agency presents its own legal, administrative and cultural challenges. Under the leadership of David Ferriero, the current Archivist of the United States (AOTUS), NARA is undergoing a transformation into a dynamic and modern agency with an outward-looking leadership role in providing innovative access to historical resources. As a major component of this strategic realignment, NARA has agreed to serve as the Secretariat of the Cooperative, and in this capacity shall provide leadership and oversight regarding the business, governance, and technical infrastructure the Cooperative. The first essential step in taking on this new role, establishing the Cooperative and the Cooperative Secretariat within the legislated mandate of NARA, has been completed in a formal Charter document, augmented with a Mission Statement that broadly defines the purpose of the Cooperative (See Appendix 1 and Appendix 2). To date, NARA has no experience in establishing and hosting a cooperative program, and thus the next essential steps in assuming operational leadership of the SNAC Cooperative will require developing NARA staff with the knowledge and skills essential to the leadership role.

The development and training of NARA staff to assume the responsibilities of the Secretariat will be a core objective of the pilot. Laura Campbell, the retired Associate Librarian and Chief Information Officer of the Library of Congress, will serve as the *interim* Director of the Cooperative, and in this capacity will work closely with NARA staff, the co-principal investigators of the pilot, and other project staff. Based on her extensive experience in founding and managing cooperative programs at the Library of Congress, Campbell has played an essential role in planning the Cooperative, and, over the course of the pilot, will be instrumental in building the knowledge and experience of NARA staff to gradually assume full responsibility for Cooperative administration and governance oversight.

In addition to developing the proficiency of NARA staff to assume the oversight role, there must be complementary development within the Cooperative community. Many professionals in archives, libraries, and museums have enthusiastically embraced the potential benefits of aggregated description and access demonstrated to date in SNAC, and, further, embrace the idea that the resources amassed should be cooperatively built and maintained in order to fully realize these benefits. The core inaugural members, alongside of the Cooperative staff at NARA, IATH, and CDL, must have a common understanding of the key objectives and begin gradually assuming shared responsibility for governing the Cooperative. In this capacity, the cooperative members must share in informing and guiding the technological development; in developing editorial policies and best practices; in developing the membership policies and a governance structure that is democratic, representative, and addresses the needs of the Cooperative; in representing the Cooperative within their affiliated institutions and in the community at large; and finally, in the ongoing curation of the Cooperative data aligned with and linked to local descriptions of holdings.

Achieving these social objectives will require a shared fuller intellectual understanding, as well as exploring and adjusting to current technological practices to take advantage of a new model of archival description and access. The pilot principal investigator, Daniel Pitti, has extensive international experience in the development and implementation of archival standards, including EAC-CPF and EAD, knowledge of both library and archival authority control, and the experience in the design and development of library and archive description systems as well as complex digital humanities projects. Alongside of Campbell, Pitti will work with NARA and CDL staff and pilot members in an educational and consulting role, assisting in the development of a shared understanding with a particular focus on the essential, intellectual, social and technological components of the Cooperative. This will be an ongoing activity that begins with establishing a core understanding, and then, building on this, to collaboratively developing both the vision and the practical operation of the Cooperative.

Inaugural members

The institutional members of the pilot Cooperative came to our attention based on their experimenting with the implementation of EAC-CPF, often in imaginative ways, because of expressing a strong interest in participating in the pilot, or often, both. The pilot members represent archives, libraries, museums (art and natural history), government archives, and institutional archives. The following institutions have accepted the invitation to participate in the pilot Cooperative (see Appendix 9 for letters of commitment):

- American Institute of Physics
- American Museum of Natural History
- George Washington University
- Getty Research Institute
- Harvard University
- Library of Congress
- National Archives and Records Administration
- New York Public Library
- Princeton University
- Smithsonian Institution (2 representatives)[4]
- Tufts University
- University of California, Irvine
- University of Miami
- Yale University

All of the pilot members come with a keen interest in contributing to the realization of the vision of the SNAC Cooperative and a good base understanding of the core intellectual and technical standards. Some of the members also have demonstrated experience with relevant activities, such as using EAC-CPF as a conduit for contributing to Wikipedia and developing methods based on OpenRefine for improving the quality of name data in archival finding aids. [5] OpenRefine could benefit the Cooperative by facilitating efficient, manual enhancement of name data in finding aids that are contributed to the Cooperative for name extraction, improving the quality of the Cooperative automated processing by "pre-reconciling" and linking names in finding aids with identity descriptions in SNAC.[6] Both of these innovative activities may lead to long- if not near-term benefits to the Cooperative and the archival community.

There will be one lead representative to the Cooperative for each participating institution, though additional staff in each institution may also participate under the supervision of the lead. Each participating institution will contribute data to the Cooperative, both existing locally maintained archival description, and also data contributed through the Cooperative's editing interface. The participating institutions also agree to use persistent identifiers on all local publicly accessible description to which SNAC is linked. The lead representative will also be responsible for the following:

[4] The Smithsonian Institution (SI) represents a large number of semi-autonomous units that have archival records. The description data of the Archives of the Smithsonian Institution resides in a distinct archival management system, while the EAD-encoded archival holdings of eleven units are aggregated in a cooperatively maintained public access system. Given the complexity and diversity of SI, it will have two representatives.

[5] OpenRefine is a desktop tool intended to help people improve the quality, consistency, and factual accuracy of data. For more information on OpenRefine, see http://openrefine.org/.

[6] See further discussion below on Identity reconciliation in SNAC and its relation to OpenRefine.

1. Receive training in archival authority control in general, and specifically in relation to Cooperative EAC-CPF profile to ensure all participants have the same base understanding.

2. Provide Cooperative staff with evaluation and suggestions on the following:
   - Quality evaluation of SNAC Cooperative data, in particular issues surrounding the quality of identity resolution: false negatives and false positives being of primary concern. Quality review input will be used to inform refining of algorithms used in automated matching and merging.
   - The public research interface, ideally engaging reference archivists and librarians, as well as users.
   - Development of training programs for future Cooperative editors. Cooperative staff in collaboration member representatives will develop the training programs.[7]
   - Development of editing standards, policies, and best practices for curation of Cooperative data.
   - Development of technical communication standards for Cooperative data.
   - Development of controlled vocabularies (occupations, functions, subjects, place names, etc.) that will enhance the maintenance economy of Cooperative data, the effectiveness of Identity Reconciliation techniques, and the research potential of the social-document network represented in the Cooperative data.[8]

3. Participate in the design and use of an editorial interface. The Cooperative will develop a maintenance platform for the data, with an editorial interface that will facilitate revision, adding, splitting, merging, and designating obsolete or inactive EAC-CPF instances. Participants will provide input on the design of the editorial interface, and as it is iteratively developed, will use it to edit EAC-CPF data, all along providing input into the ongoing development and refinement.

4. Provide and support public outreach to the local user community, including promoting the Cooperative and possibly conducting local workshops for researchers.

5. Participate in developing the governance of the Cooperative, and, as called upon, participate in the governance as it unfolds.

During the pilot, the Cooperative will cover the travel costs of the lead participant for attending meetings. Additional representatives may also attend, though the member institutions must cover travel expenses for any additional representatives attending.

Community building

In order to achieve the community building objectives of developing a shared understanding and competency, and of developing an expanding community support, NARA will host a number of in-person meetings with the pilot participants. The meetings will serve to solicit input from as well as educate pilot members regarding the EAC-CPF standard and best practices for content input and maintenance workflow, provide input into the development of the editing user interface, and, as the interface becomes active, use it to edit and maintain the Cooperative data. All SNAC Team members recognize that it is critical for the pilot participants to contribute to what works best in their local environment and how that relates and contributes to scaling the cooperative to many participating institutions. These meetings will also help to foster, inspire and stimulate a sense of communal ownership within pilot participants. Participants will take part in planning and modeling activities for the future operations and sustainability

[7] The Society of American Archivists (SAA) has expressed interest in offering Cooperative training through its education program. It is anticipated NARA staff and interested members of the Cooperative will do the training.

[8] It is worth noting that the archival community has for many years wanted controlled vocabularies for occupations and functions. In lieu of such vocabularies, archivists and librarians have relied on selecting terms from a variety of sources or merely invented terms. Such non-standard practice has an adverse impact on using the terms to facilitate access. For the Cooperative, the non-standard use of the terms has an adverse impact on access as well as the use of the terms in computer-assisted Identity reconciliation. The Cooperative will make it possible to standardize these critical vocabularies.

of SNAC. This communal sense of investment will serve to embed SNAC more deeply within the archival and library community.

The SNAC project team will also utilize these meetings to gather a baseline understanding of what it means for pilot members to participate in SNAC at its inception. We will gather information on practical considerations such as man-hours, technical logistics, initial workflow, and degree of content standards or best practice knowledge. We will also start to brainstorm and define with the community how SNAC's content and technical infrastructure impacts and adds value to local practice and overall community practice. At the end of the pilot period we will have determined the best methods to evaluate and refine community goals and objectives.

One additional and important objective of the meetings is to work with the pilot participants in advocating on behalf of the Cooperative, communicating the benefits for both resource curators and users of the resources and sharing the knowledge and experience acquired through their participation. We may identify conference, workshops, or other community educational or experience sharing opportunities where pilot participants can find opportunities to share their own SNAC "story"—what has been learned, what has been the most valuable part of their experience, and what can be explored in the future with the wider community.

Meetings

NARA will host four three-day meetings over the course of the two-year pilot. The following provides an overview of the objectives and content of the four meetings. Between the face-to-face meetings, communication will be facilitated through email, conference calls, webinars, and online updates. For more in-depth detail, see Appendix 3, Community Meeting Agendas.

> **Meeting One** (September, 2015): Kick off meeting. Introductions to all major SNAC project areas, goals, and objectives. Introductions among all pilot institutions, gather and share baseline pilot participant information to understand individual environments and commonalities. Develop a clear understanding of what infrastructure now exists and what will be developed jointly to ensure that all participants understand the goals of the pilot.
>
> (Remote activities: Updates and informational sessions via webinars, email, other online engagement)
>
> **Meeting Two** (May, 2016): Content standards training, user interface wireframe demonstration, workflow process brainstorming and follow-up assignments. Discuss data quality issues, discuss and demonstrate best practices resources from the Secretariat, initiate brainstorming among members for future long-term operations, sustainability planning, and governance structure. The afternoon of day 3 will be for the SNAC Program Team only.
>
> (Remote activities: User interface testing with a small number of pilot participants)
>
> **Meeting Three** (November, 2016): User interface launch and training, communication and opportunities for community presentations.
>
> **Meeting Four** (March, 2017): Pilot phase wrap up, technical planning for future development, content standards approval and release, post-pilot continuity, wrap up assessment activities, and conclusions regarding future publication, reporting, and conference presentations. The afternoon of day 3 will be for the SNAC Program Team only.

Technology

The long-term technological objective for the Cooperative is a platform that will support a continuously expanding, curated corpus of reliable biographical descriptions of people linked to and providing contextual understanding of the historical records that are the primary evidence for understanding their lives and work. Building and curating a reliable social-document corpus will involve a *balanced combination* of computer processing and human identity verification and editing. The next step towards realizing the long-term objective is to transition from a research and demonstration project to a production web service. From a technical perspective, this means transitioning from a multistep human-mediated batch process to an integrated transaction-based platform. Instead of the data being passed along from one programmer to another, the architecture will automate the flow of data in and out of the different processing steps by interconnecting the processing components, with events taking place in one component triggering related events in another. For example, the addition of a new descriptive record will lead to automatic updating of graph data in Neo4J and updating the indexed data in the History Research Tool. The coordinated architecture will support both the batch ingest of data *and* human editing of the data to verify identities and refine and augment the descriptions over time.

Using techniques developed in the research and demonstration phase of SNAC, computer processing will be used to extract and ingest existing name authority and biographical data from existing archival descriptions. Identity reconciliation, i.e. matching and combining two or more descriptions for the same person, organization, or family, has relied solely on algorithms in the research phase. While identity reconciliation techniques will continue to inform the reconciliation process, Cooperative professional editors, beginning with librarians and archivists though expanding over time to include allied scholars, will verify identities and curate the data. This two faceted approach, combining intelligent computer processing and professional editing, will enable building a large corpus of networked social-document data that is not constrained geographically or by historical period, and over time establishes an expanding core of reliable identities within the overall corpus. (See Appendix 4 for a diagram showing the relationship between certain/uncertain data and dense/sparse evidence for identity resolution.)

Current SNAC R&D Technology Platform

Current SNAC processing employs a complex sequence of steps that produces a collection of biographical descriptions used to produce the History Research Tool.

- Acquire source data from archives, libraries, and OCLC WorldCat (EAD-encoded finding aids; MARC21; and ad hoc authority and biographical data sets).
- Extract data essential to assembling descriptions of persons, organizations, and families into standardized descriptions (EAC-CPF).
- Load the name and, when available, the existence dates of the entity described in each EAC-CPF instance into the PostgreSQL database the multi-component platform that attempts to match and combine different descriptions of the same identity into a single description.
- Names in the PostgreSQL database are matched against more than 25 million identities in the Virtual International Authority File (VIAF), maintained by OCLC. Cheshire, an XML-based indexing tool developed at the University of California, Berkeley is a key component of this step. Based on a sequence of matching attempts, additional data is associated with each identity in the PostgreSQL database.
- Based on an evaluation of the match processing results, a final set of EAC-CPF instances is produced to serve as the basis for the History Research Tool. For instances that are deemed to be for the same identity, the different instances are combined based on complex merge algorithms; instances deemed to be possibly for the same identity but lacking sufficient confidence to be merged are related to one another as "maybe the same as;" and finally instances that do not match or weakly match are passed into the results as is.
- A subset of the EAC-CPF data is extracted and loaded into Neo4J, a graph database. The graph database serves two primary functions: support of graphical representations of the SNAC social-document network, and exposing a subset of the SNAC data as Resource Description Framework (RDF) Linked Open Data (LOD).[9]
- The resulting set of EAC-CPF instances is used to produce the History Research Tool. The key underlying platform is XTF, an open-source XML-based publishing tool developed at the California Digital Library.

Transforming the existing platform into a platform that supports both ingesting of large batches of data but also manual maintenance of the data will require a reconfiguration of major components of the current underlying technology. Two major existing components will be retained with minimal modification during the pilot: the "back end," the processing used to extract data from existing descriptive sources and assemble it into EAC-CPF instances; and the "front end," the History Research Tool. While the code and technology for these two components of the

[9] See Graph Data Store section below for a detailed description of use of LOD/RDF in SNAC.

SNAC infrastructure would benefit from additional development and refinement, each is sufficiently robust and functionally complete to remain largely unchanged during the pilot. The intermediate technologies used in loading and matching of the EAC-CPF will need to be thoroughly revised, retaining existing functionality but in a configuration that will support both batch and manual maintenance. One component, the processes used to merge or combine matching EAC-CPF instances will be deferred to a later stage of development (to be described below). Finally, two components that will be developed in the pilot are entirely new: an API to support both batch processing and an Editing User Interface, and development of the Editing User Interface itself. While the transformation of the underlying technology is underway, there will be a one-year pause in batch ingesting new data in order to focus programming resources on the essential development work needed to go forward. No large batches of new source data have been solicited for the pilot, although pilot member institutions will contribute batches of data for use in first testing and then bringing online the batch ingest function of the Cooperative. During the pilot, new sources of batch data will be solicited for the second two-year phase of establishing the Cooperative.

Data Maintenance Store

In the current processing stream, the EAC-CPF instances are placed in a read-only directory as the primary data store. A small number of select components (name strings) of each EAC-CPF XML-encoded instance are loaded into a PostgreSQL database. In order to support dynamic manual editing of the EAC-CPF instances, it will be necessary to parse the entirety of each EAC-CPF instance into PostgreSQL tables.[10] Parsing all (or most) components of the EAC-CPF instances into SQL tables is necessary because no open source native XML database will efficiently support the essential maintenance functionality required, in particular effectively managing editing transactions at the component-level of each EAC-CPF instance.[11] MarkLogic would enable maintaining the data in XML, but it is an expensive commercial platform. The most robust of the open source native XML databases, eXist, does not support transaction management. Further, EAC-CPF was not designed as a maintenance format, but rather as a communication format and, with this in mind, was intentionally designed to facilitate the serialization of the data into and out of SQL environments.[12]

The PostgreSQL Data Maintenance Store (DMS) will represent the core foundation of the SNAC technology platform. It will hold the crucial SNAC data, including: the parsed EAC-CPF instances, version control for modified data fields, editor authorization privileges, editor work histories (e.g., edit status on individual EAC-CPF instances), local controlled vocabularies (e.g., occupations, functions, subjects, and geographic names), and workflow management data. An open source authentication system will mediate editor/user access to the DMS. All other component subsystems will rely in large part on the DMS, with several of the DMS functions being generalized to be effective through the component APIs. Additionally, nearly all reporting for editors and administrators will be based on the DMS. We will use an open source reporting package, but the reports themselves will need to be generated by the DMS. With the DMS as the core foundation of the SNAC technology platform, evolving features of the component subsystems will often require further development of corresponding functions in the DMS.

Identity Reconciliation

A major focus of the SNAC R&D has been on identity reconciliation. A fundamental human activity in the development of knowledge involves the identification of unique "real world" entities (for example, a particular person or a specific book) and recording facts about the observed entity that when taken together uniquely distinguishes the entity from all others. Establishing the identity of a person, for example, involves examining

[10] PostgreSQL is a widely used and supported open source, SQL standards-compliant relational database management system. Using PostgreSQL as the maintenance platform for the authoritative EAC-CPF descriptions will ensure data integrity and provide robust performance for the large quantity of data (current and anticipated) in SNAC.

[11] A small number of complex EAC-CPF XML components may be stored and maintained outside of PostgreSQL using an alternative technology because their complexity may not be efficiently processed in an SQL environment. SNAC programmers are currently experimenting with a PostgreSQL database with 2.6 million EAC-CPF XML-encoded instances fully parsed into normalized tables in order to assess the optimum storage and maintenance configuration.

[12] The PI was the technical editor of the EAC-CPF schema.

available evidence, including the existing knowledge base, and recording facts associated with him or her. For a person, the facts would include names used by and for them, dates and places of birth and death, occupation, and so on. Establishing identities is an ongoing, cumulative activity that both leverages existing established identities and establishes new identities. Identity reconciliation is the process by which an encountered identity is compared against established identities, and if not found, is itself contributed to the established base of identities. The networked computing environment presents opportunities for using algorithm-based inference methods for comparing newly encountered entities with established identities to determine the probability that a new entity represents the same person or thing as an established identity. In this way, the ongoing expansion of the base of reliable identities is an interplay of human research, knowledge recording, and computational methods.

With the emergence of Linked Open Data (LOD) and the opportunity it presents to interconnect distributed sets of information, new names for entities are introduced, namely the URI's used to provide globally unique identifiers to entities. In order to exploit the opportunity presented by LOD, it necessary to include these URI's in the reconciliation process. SNAC assigns its own identifiers (ARKS) because doing so is essential to effectively managing the identities throughout the processing and maintenance. Even if this were not essential for managing the workflow, the majority of the identities in SNAC will not be found in other sources such as VIAF, and thus the SNAC identifiers and associated data that establish the identity are likely to be unique, at least in the near term.[13] For those identities that do overlap with VIAF, SNAC processing takes advantage of the VIAF reconciliation to associate the VIAF identifier as well as identifiers for Wikipedia and WorldCat Identity.

SNAC identity reconciliation processing, while sufficiently reliable for the research and demonstration purposes of SNAC, is inadequate for meeting the long-term objective of building a large corpus of networked social-document data that, over time, establishes an expanding core of reliable identities within the overall corpus. This objective will require a *calibrated balance* of computational methods and human verification of identities, and it must be possible to refine the balance over time as new understandings and insights into the processing arise, and as new data patterns are encountered in new sources data. To inform the refinement of the match processing in balanced relation with the human verification, an ongoing quality review regimen needs to be put in place that supports efficient evaluation of match algorithms that employs both benchmark data used in testing and revising algorithms, and feedback for revision of algorithms. Revision of the match processing will require the following:

- Developing an Identity reconciliation services module with API that supports batch match evaluation processing, evaluation of identities that are manually added through the Editing User Interface, and OpenRefine users interested in reconciling names against SNAC identities.[14]
- Establishing a quality review process that employs benchmark or "ground truth" data to be used in evaluating and refining match algorithms and a regimen of ongoing human review that is performed collaboratively by Cooperative staff and members. The benchmark data will be initially based on the in-depth match quality evaluation that was performed by an outside consultant during the SNAC research phase. The consultant's report, describing in detail both methods employed and results will also serve as the basis for training Cooperative staff and members.

The major components of the current match or identity reconciliation processing are Cheshire and PostgreSQL. For each EAC-CPF instance, Cheshire is used to perform a series of queries of the Virtual International Authority File (VIAF), with the results of the queries (positive matches, possible matches, and non-matches) being stored in the PostgreSQL database for use in later processing. The Cheshire index may be replaced because it has proven to be problematic to maintain, though the final decision will be based on ongoing testing of alternatives, such as ElasticSearch, to ensure that that any replacement for Cheshire will provide the performance and functionality required.

There are three major reconciliation results: reliable match; possible match; and no match. These determinations are based on string matching combined with the relative "identity strength" of each name string, that is, the extent to which the name string is likely to uniquely identify a person, organization, or family in a large population. Factors

[13] 24.8% of SNAC identities match VIAF identities.
[14] The SNAC reconciliation process will be exposed via a REST API which follows the OpenRefine Reconciliation Service API guidelines.

considered in determining the "identity strength" of a name string are the following: length of the string, the number and length of name components, the order of the name components, the presence or absence of life dates, how common the surname is, and others. Another way of understanding "identity strength" is as a determination of the quantity and quality of available evidence.[15] If two *matching* strings contain sufficient evidence, then two strings are deemed a reliable match; if the strings do not contain sufficient evidence, then the strings are flagged as a possible match.

The reconfiguration will enable the identity reconciliation processing to be integrated with the extraction/assembling processing, the data maintenance platform, and the editing user interface.[16] EAC-CPF extracted and assembled using existing archival descriptions (EAD-encoded finding aids, MARC21, or existing non-standard archival authority records) will be batch ingested into the SNAC Cooperative PostgreSQL database; when ingested, the Identity Reconciliation module will be invoked to flag reliable matches, possible matches, non-matches. The results of the identity reconciliation evaluation will be available to editors through the Editing User Interface to assist them in verifying identities. When editors create new identity descriptions or revise existing descriptions, the Identity reconciliation module will be invoked to provide the editors with feedback on likely and potential matches that may be otherwise overlooked when employing human-only authority control techniques.

Merge processing, that is the automatic merging or combining of EAC-CPF instances deemed reliably to be for the same identity, will be deferred until after the revision of the match processing and developing an effective quality evaluation regimen. The current merge processing combines two or more EAC-CPF instances into one. The combining is primarily cumulative, though redundant data fields are combined as a further step. Once human verification and editing is introduced, any automatic merging of records will necessarily need to respect the integrity of the judgment of the human editors. The merge algorithms will need to be based on policies developed in consultation with the Cooperative community, taking into account quality evaluation findings and the nature of the components of each description. The engagement of the archivists and librarians thus will be essential in developing an appropriate balance of computer and human maintenance of the data. An informed understanding of the issues will not be possible until the editing platform and editing interface are functional, and the community knowledgeably engaged.

Editing User Interface

Developing the Editing User Interface (EUI) is a primary objective of the two-year pilot. The SNAC developers have identified the essential functional requirements for the interface, and extensive user studies with archivist and librarians have been used to refine and substantially extend the requirement list. Because the EUI is dependent on the reconfiguration and development of the data maintenance and identity reconciliation modules, and the development of the Edit API, the development will first focus on engaging the pilot participants by means of wireframes of the EUI, in conjunction with rehearsing established research and description tasks, the order or orders in which such task are performed, and walk-throughs of the steps involved in manually adding, revising, merging, and splitting identity descriptions. These activities and the findings that result from them will inform the parallel development of the maintenance platform. When the underlying data maintenance platform is in place, development of the EUI will commence, informed by the activities described above. As the EUI becomes functional, the pilot participants will transition to iteratively testing and using it to perform editing tasks to ensure that the essential functions are supported and that this support makes the performance of the tasks logical and efficient. Those functions of the EUI that overlap with the History Research Tool will employ a common interface. The bulk of the EUI will be based on JavaScript running in modern web browsers.

[15] The relative identity strength of names is sometimes described in terms of "dense" names, and "sparse" names. This form of evaluation emulates human judgment, in that a person will likely evaluate a "dense" name (for example, "Schleich, Carl Ludwig, 1859-1922") as highly likely to be unique, and a "sparse" name (for example, "Adams, James") as requiring further research and additional evidence to verify the identity.

[16] For a full description of the integration of the components of the Cooperative platform, see below.

Graph Data Store – Visualizations and Exposure of RDF/LOD

Neo4J, an open source graph database, is currently used to generate social-document network data as GraphML. The generated GraphML supports the graphic representations of the social-document network in the History Research Tool. In addition, it supports exposure of the SNAC data for third-party use through Resource Description Framework (RDF) Linked Open Data (LOD). The Neo4J component will not require major reconfiguration during the pilot, but will be used as an active source of the social-document network data in place of static GraphML representation. In this capacity, the Neo4J database will provide a number of services: serving graph data to drive social-document network graphs in the HRT; and serving and providing LOD through a SPARQL endpoint and RDF exports for third-party consumption. It will, however, be necessary to integrate both the data ingest and data serving functionality of Neo4J into the coordinated system architecture in order to ensure that the graph data and dependent services remain current with the evolving SNAC data.

While the initial focus of the Cooperative is on the curation of data, it is interested in using the social-document network data in public graphical displays of the networks and exposing the data as LOD in RDF syntaxes (RDF/XML and RDF compatible JSON-LD) for use by third parties. By exposing the data for use in a variety ways (including making available full EAC-CPF XML-encoded instances) the Cooperative will contribute to global open data initiatives that are intended, among other objectives, to promote use and reuse of data in innovative ways, and to interconnect and interrelate complementary resources that currently exist in isolation from one another.

Currently there is no existing ontology for archival description, and thus the classes and properties used in exposing graph data expressed in RDF is based on classes and attributes selected from existing, well-known and widely used ontologies and vocabularies: Friend of a Friend, OWL, SKOS, Europeana Data Model (EDM), RDA Group 2 Element Vocabulary, Schema.org, and Dublin Core elements and terms.[17] In the long term, it should be noted that the International Council on Archives' Expert Group on Archival Description (EGAD; chaired by the PI) is developing an ontology for archival entities and the description thereof. While the initial focus of EGAD necessarily is focused on developing a clear model of the world based on archival curatorial principles, once this work is completed, the group intends to collaborate in mapping the archival ontology to CIDOC CRM, which incorporates both museum and library description.[18]

Component Subsystem Integration

In order to integrate the component subsystems of the Cooperative technological platform we will develop a thin middleware component that routes each request from a client or from an intra-server process to the appropriate individual subsystem based on SNAC workflows we will establish. The middleware component invokes the required functions via calls to the subsystems: Identity Reconciliation, PostgreSQL database, Neo4J graph database, History Research Tool, and Editing User Interface. These subsystems deal with the variety of automated and semi-automated transactions required by the Cooperative platform.

We will develop the Cooperative middleware component as a LAMP application rather then using an open source Enterprise Service Bus (ESB).[19] We will use LAMP (with PostgreSQL) for efficiency of coding; flexibility enabled by

[17] Among the RDF vocabularies considered was BIBFRAME, an initiative led by the Library of Congress to replace the MARC21 format using graph technologies, specifically the W3C Resource Description Framework (RDF). BIBFRAME aspires to be "content standard independent," and to accommodate library, museum, and archival description. BIBFRAME development is still in the early stages, and most of the development work centers on accommodating data currently in MARC21. It is unclear at this stage in the development of BIBFRAME whether it will attempt to accommodate the data in EAC-CPF.

[18] "Toward an International Conceptual Model for Archival Description: A Preliminary Report from the International Council on Archives' Experts Group on Archival Description" in *The American Archivist* (Chicago: SAA), 76/2 Fall/Winter 2013, pp. 566–583. With Gretchen Gueguen, Vitor Manoel Marques da Fonseca, and Claire Sibille-de Grimoüard. Also available here: http://www.ica.org/13851/egad-resources/egad-resources.html.

[19] For more information on LAMP, see http:/en.wikipedia.org/wiki/LAMP_%28software_bundle%29#Variants. The article notes that: "A version where MySQL has been replaced by PostgreSQL is called LAPP, or sometimes by keeping the original acronym, LAMP (Linux / Apache / Middleware (Perl, PHP, Python, Ruby) / PostgreSQL)."

a very large number of available software modules; ease of maintenance; and clarity of software architecture (that is, the model-view-controller).[20] The result will be a lightweight and easy to administer software stack. (See Appendix 11 for diagrams illustrating the basic system architecture.) ESBs contain far more features than required by the Cooperative architecture, and because of the complexity these unwanted features present, it would be less efficient to configure and maintain the middleware using an ESB than simply selecting only the components needed by the architecture from a large library of existing open source LAMP modules.

The thin middleware component will be available via a RESTful API, allowing appropriate third party access to services. A simple example might be a dedicated MARC21-to-EAC converter where a MARC21 record is uploaded, data extracted and transformed into EAC-CPF, and returned in a single transaction. Another example is saving an identity record where the data is written to PostgreSQL, EAC-CPF is exported to and indexed by XTF, and the Neo4j database is updated. The three steps are sequenced by the thin middleware component.

With respect to potential third party applications, Brad Westbrook from Lyrasis has participated in the technical planning for the Cooperative, in anticipation of extending ArchivesSpace to include a SNAC editing interface. Adding this capacity will enable archives, libraries, and museums using ArchivesSpace to benefit from the Cooperative data, and to link the description of local holdings to SNAC descriptions of organizations, persons, and families. The PI has also had discussions with Susan Perdue at the Virginia Foundation for the Humanities and Director of Document Compass. Perdue is currently in the planning stages of developing an open source platform to support documentary editing. Incorporating a SNAC editing interface into this platform would be of significant benefit to documentary editors (as a reference resource), and would expand the range of professionals participating in the SNAC Cooperative. Finally, the Staatsbibliothek zu Berlin is keenly interested in interrelating Kalliope with SNAC, contributing 600,000 person and organization descriptions linked to 2.5 million descriptions of archival holdings.

Development Strategy

The development of the SNAC subsystems will be developed in parallel because of their interdependence. The three principal programmers all reside at IATH and will function as a team under the supervision of Co-PI Martin. While the programmers will have primary responsibility for one or more components (see Job Titles and Job Descriptions below), each will contribute to all component subsystems as required. The programming team (Co-PI and three programmers) will meet weekly, with the PI attending as necessary. The development will be iterative, with extensive prototyping and testing by programmers, project staff, and pilot members, and the ongoing work will necessarily need to be responsive to input from pilot members on features, editing policy, testing, and other issues as the pilot progresses.

Estimated programming effort for each of the five subsystems is as follows:

- Data Maintenance Store - .6 FTE
- Identity Reconciliation - .75 FTE
- Editing User Interface - .6 FTE
- Graph Data Store - .3 FTE
- Component Subsystem Integration - .25 FTE

Hardware Platform

The processing of SNAC data entails running batch processes that consume large amounts of RAM, and which benefit from inherent multi-threading in Saxon and other tools. Additionally, these processes typically produce large amounts of output. Typical processes will generate several gigabytes of data, and the larger, multi-day runs will create tens of gigabytes of output data. At the same time, the server must be capable of simultaneously handling http (web) requests in a timely manner. Thus the server needs 112GB of RAM, 16 (2 x 8 core) CPU, and 7.2TB of disk. A

[20] For more information on the model-view-controller see http://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller

dedicated server will be purchased and maintained at IATH for these tasks. Current research indicates that the Dell PowerEdge rack server offers an optimal combination of performance and storage.

JOB TITLES AND JOB DESCRIPTIONS

Note: the National Archives and Records Administration (NARA) is contributing substantial in-kind personnel. The personnel are listed below with percentage of commitment followed by "NARA funded." The total in-kind NARA staff contribution is 2 FTE.

Principal Investigator (PI) – Daniel Pitti will serve as the Principal Investigator. He is Associate Director of the Institute for Advanced Technology in the Humanities (IATH), University of Virginia. Pitti has extensive experience in the development of library and archival description standards and systems, including the design of two "linked-authorities" library systems (UCLA and UC Berkeley), and was the chief technical architect of both EAD and EAC-CPF. He has served as the PI for SNAC since 2010. Pitti will have ultimate responsibility for administering the grant, will serve on the Cooperative Steering and Policy Committee/Project Management Team, and will in addition provide expert consulting on governance, training, policy, data, and technical requirements. (40%)

Co-principal Investigator (Co-PI) – Worthy Martin, IATH, Director and Associate Professor, Department of Computer Science. Martin will serve on the Cooperative Steering and Policy Committee/Project Management Team and will supervise the programmers working on the Cooperative technological platform. Martin will also be jointly responsible with the PI for administration of the project, including financial and results reporting. (10% Academic Year, 12.5% summers)

Administration Coordinator (AC) – Sarah Wells, Scholarly and Technical Communications Officer, IATH. Wells has assisted the PI in the oversight and administration of SNAC since 2010. In this role, Wells will have among other duties the following: assist in the monitoring, managing, and reporting on project activities and fund accounting and serve as the liaison to the Cooperative Secretariat in planning and developing pilot communication, events, and training. Wells will serve on the Communication, Meeting, and Event Planning Team. (15%)

Cooperative Director/Project Manager (CD/PM) – Laura Campbell, retired Associate Librarian and Chief Information Officer at the Library of Congress. Campbell will be responsible for oversight and management of all Cooperative activities and operations, serve as the chief representative of the Cooperative, and supervise and mentor NARA staff engaged in Cooperative activities. Campbell will chair and serve on the Cooperative Steering and Policy Committee/Project Management Team. (45%)

Deputy Director/Project Manager (DD/PM) – John Martinez, Director, Business Architecture, Standards, and Authorities Division, Office of Innovation, NARA. Martinez will work closely with the CD/PM in providing oversight and management of Cooperative activities and operations and will serve on the Cooperative Steering and Policy Committee/Project Management Team, and chair the Communication, Meeting, and Event Planning Team. In addition, he will have responsibility for, among other tasks: development and maintenance of member registry; maintenance and updating of documents such as strategic and operational plans; overseeing communications of all types; meeting and event planning; statistics and reporting; management of training; and preparation of annual report to the Archivist of the U.S. and the Cooperative membership. (50%, NARA funded)

Director of Governance and Training (DGT) – Jerry Simmons, Authority Cataloging Team Lead, Office of Innovation, NARA. Simmons will be the lead on Cooperative governance and training. In collaboration with members and pilot staff, Simmons will develop the initial training, editorial policies, and procedures, each fully documented. Simmons, with input from the members, will develop governance structure and policy recommendations for the Cooperative Steering and Policy Committee/Project Management Team. As called upon, will participate in meetings of the Cooperative Steering and Policy Committee/Project Management Team; and will serve as a member of the Communication, Meeting and Event Planning Team. (50%, NARA funded)

Administration, Governance, and Training Assistants (AGTA) – Staff from the NARA Authorities Cataloging Team (Amanda Ross, David Schlanger, and Amber Thiele) will assist the Administration and Governance directors as needed.) (Combined effort 100%, NARA funded)

User Experience Design Lead (UEDL) – Rachael Hu, CDL. User Experience Design Manager. Hu will represent user in the design and implementation of the Editing User Interface, and will work with the technical team to prioritize and phase feature development. As called upon, will participate in meetings of the Cooperative Steering and Policy Committee/Project Management Team, and serve on the Communication, Meeting, and Event Planning Team. (15%)

History Research Tool Programmer (HRTP) **–** Brian Tingle, Technical Lead, Access and Publishing, CDL. In collaboration with the Co-PI and programmers, Tingle will implement changes in the History Research Tool to ensure that public data is current with data as it is maintained. He will also assist in transferring maintenance responsibility for the HRT to TDP3. (5% for year one)

Technical Development Programmer (TDP1) – Tom Laudeman, SNAC Data Extraction/Transformation Programmer, IATH. Laudeman will be responsible for the development of the PostgreSQL component of the Cooperative platform. With TDP2 and TDP3, he will be jointly responsible for the development of the APIs and coordination layer of Cooperative platform. (100%)

Technical Development Programmer (TDP2) –IATH (to be hired). Programmer will be responsible for development of the Identity Reconciliation module. Will be responsible for integration of the Neo4J graph database into the Cooperative coordinated platform. With TDLP1 and TDP3, this person will be jointly responsible for the development of the APIs and coordination layer of Cooperative platform. (See Appendix 6 for a draft job description.) (100%)

Technical Development Programmer (TDP3) – IATH. Programmer will be responsible for development of the Editing User Interface. In collaboration with TDP1 and TDP2, will be jointly responsible for the development of the APIs and coordination layer of Cooperative platform. (TDP3 will be one of two existing IATH programmers with expertise in developing UIs.) (50%)

System Administrator (SA) – IATH. Will be responsible for set up of the SNAC Cooperative server and its ongoing maintenance. (5%)

Consultant on Identity Reconciliation Processing – Ray Larson, Professor, School of Information, University of California, Berkeley. Larson is the designer of Cheshire and oversaw the development of SNAC research phase identity match and merge processing. He will provide expert advice to the TDP2 in the development of the Identity Resolution module. (16 days)

Consultant on Visual Design – Anne Chesnut, graphic designer. Chesnut designed the current SNAC web site. She will be responsible for the design, development, and implementation of the Editing User Interface, in conjunction with the UEDL and technical staff. (80 hours)

Administration and Management of Pilot and Cooperative

While the PI and Co-PI will have ultimate administrative responsibility and authority for the performance of the work proposed in the pilot, financial oversight, and reporting, the Cooperative Director/Project Manager (CD/PM) and Deputy Director/Project Manager (DD/PM) will have delegated responsibility for managing and supervising the operations of the pilot. This administration and management structure will ensure that all essential administrative responsibilities are modeled in the pilot Cooperative Secretariat in anticipation of the Secretariat seamlessly assuming full authority after the Cooperative is fully established.

Delegated responsibilities will include ensuring that all pilot activities are being conducted in a timely and efficient manner and that project milestones and objectives are being met. The Director and Deputy Director will share responsibility with the PI and Co-PI in monitoring the pilot budget. The CD/PM and DD/PM exercise direct oversight of the DGT, and the UEDL. Oversight will include monitoring progress, determining when assigned tasks have been performed and objectives met in a timely manner.

The pilot will have a Cooperative Steering and Policy Committee/Project Management Team. The objective for this group will be to put in place a transitional body that will first and foremost be responsible for coordinating the management of the pilot, but will also serve as the basis for a Steering and Policy Committee after the Cooperative is fully established. Initially the Committee will have four members: CD/PM, DD/PM, PI, and Co-PI. (See Appendix 5 for CVs of these members.) The Steering Committee will be chaired by the CD/PM. The group will meet via teleconference no less than twice each month of the pilot, supplemented with email as needed. Other members of the project team, the Director of Governance and Training (Simmons) and the Technical Development Manager, in particular, may be asked to participate in the meetings, at the request of any member of the group.

Over the course of the pilot, institutional members will participate in shaping the long-term mission and composition of the Cooperative Steering and Policy Committee, and membership in the transitional group in an advisory capacity.

A Communication, Meeting, and Event Planning Team will be chaired by the DD/PM. This team will be responsible for planning public communications and promotion, meetings of the members, and public events. Members of the team will also include the AC, DGT, and UEDL.

LENGTH OF PROJECT WITH TIMELINE

The project will last from July 1, 2015 to June 30, 2017.

TIMELINE				
PRE-PILOT (APRIL 1-JUNE 30, 2015)	**ADMINISTRATION**	**GOVERNANCE**	**MEETINGS**	**TECHNICAL SYSTEM**
	NARA resource allocation (DD/PM)	Develop preliminary standards, best practices, workflow (DGT)	Pre-meeting planning logistics (PI, DD/PM, AC)	
	Community meeting planning (CD/PM, DD/PM)	Develop training for content (DGT)		Community meeting planning (PI, Co-PI)
	Plan for program assessment and evaluation (cd/PM, DD/PM)	Plan for training for system (DGT)		Plan for training for system (UEDL, TDP1)
	Set up overall project management infrastructure (CD/PM, DD/PM)	Set up pilot governance structure (DGT)		Plan for program assessment and evaluation (UEDL)
	Develop reporting expectations and mechanisms (CD/PM, DD/PM)			Refine user requirements, specifications, and wireframes (UEDL)

	Legal consultation (intellectual property and data ownership) (CD/PM, DD/PM)			Refine technical system architecture (UEDL, TDP1, TDP2, HRTP)
	Develop and finalize membership levels, rights, and responsibilities (throughout pilot) (CD/PM, DD/PM)			Refine technical specifications for system (UEDL, HRTP, TDP1, TDP2)
	Sustainability planning (throughout pilot) (CD/PM, DD/PM)			Participate in development of training for content (UEDL, TDP1)
	Fiscal management (throughout pilot) (PI, co-PI, AC)			
	Develop preliminary promotional and outreach strategies; establish social media channels and branding (DD/PM)			
YEAR ONE (JULY 1- DEC 31, 2015)	**ADMINISTRATION**	**GOVERNANCE**	**MEETINGS**	**TECHNICAL SYSTEM**
				Hire TDP3 (PI and Co-PI)
	Set up internal and external communication infrastructure and practices (DD/PM)	Develop training for system (DGT)	Meeting #1 (Kick-off)	Develop training for system (UEDL, TDP1, TDP2, TDP3)
	Plan for customer service infrastructure (CD/PM, DD/PM)	Continue to develop preliminary standards, best practices, workflow (DGT)	Virtual Communication (October-May)	Set up technical project management infrastructure
	Continue to conduct sustainability planning (CD/PM, DD/PM)	Continue to develop training for content (DGT)		Visual design of UI including Wireframes (UEDL, TDP1)
	Fiscal management (PI, co-PI, AC)	Conduct training for content (DGT)		Set up programming environment (TDP1,

				TDP2, TDP3)
	Implement and maintain promotional and outreach activities; revise branding as needed; revise branding as needed (DD/PM)			Set up quality assurance and technical testing infrastructure (TDP1, TDP2, TDP3)
				Database schema and related API (TDP1, TDP2, TDP3)
				Development of maintenance and editing UI (throughout pilot, UEDL, HRTP, TDP1, TDP2)
				Create basic reconciliation functionality (TDP2)
				Set up bug tracking system (TDP1, TDP2, TDP3)
				Set up infrastructure to assess future UI additions, changes (UEDL, HRTP, TDP1, TDP2)
				Build core web application (TDP1, TDP2)
				Web application integrated with workflow automation (TDP1)
				Deploy core authentication and security (TDP1, TDP3)

				Create new specifications for merge functionality (PI, TDP2)
				Create core reconciliation API (TDP2)
Year One (Jan 1-June 30, 2016)	**Administration**	**Governance**	**Meetings**	**Technical System**
	Administer internal and external communications (DD/PM)	Continue to develop preliminary standards, best practices, workflow (DGT)	Meeting #2 (Content Standards, UI Wireframe, and Workflow)	Continue development of maintenance and editing UI (UEDL, TDP3, TDP1)
	Set up customer service infrastructure (DD/PM)	Conduct training for content (DGT)		Add web dashboard, more detailed workflows, authorization (UEDL, TDP1, TDP3)
	Manage reporting and overall milestone tracking (CD/PM, DD/PM)	Continue to develop training for system (DGT)		Reconciliation API continued development (TDP2)
	Continue to conduct sustainability planning (CD/PM, DD/PM)			Develop training for system (UEDL, TDP1)
	Fiscal management (PI, co-PI, AC)			Refine all APIs, especially add features to support EUI (TDP1, TDP2, TDP3)
	Maintain promotional and outreach activities (DD/PM)			Integrate off-the-shelf reporting system (primarily TDP1)
				REST API development, testing, and third party integration (TDP1, TDP2, TDP3)

				Neo4j and graph database related API and integration (TDP2)
Year Two (July 1-Dec 31, 2016)	**Administration**	**Governance**	**Meetings**	**Technical System**
	Administer internal and external communications (DD/PM)	Continue to develop preliminary standards, best practices, workflow (DGT)	Meeting #3 (User Interface Launch and Training)	Accessibility review of system (UEDL, TDP1, TDP2, TDP3)
	Set up customer service infrastructure (DD/PM)	Develop lessons learned and adjustments to content training (DGT)		Workflow automation testing, refinement (UEDL, TDP1, TDP2, TDP3)
	Manage reporting and overall milestone tracking (CD/PM, DD/PM)	Conduct training for system (DGT)		Continue development of maintenance and editing UI (TDP1, TDP2, TDP3)
	Continue to conduct sustainability planning (CD/PM, DD/PM)			Conduct training for system (UEDL, TDP1)
	Fiscal management (PI, co-PI, AC)			Finalize reconciliation process and related API (TDP1, TDP2)
	Manage external partnerships (e.g. Lyrasis) and develop third-party use agreements (CD/PM, DD/PM)			Continue database schema and API development (TDP1)
	Maintain promotional and outreach activities (DD/PM)			
Year Two (Jan 1-June 30, 2017)	**Administration**	**Governance**	**Meetings**	**Technical System**
	Administer internal and external communications (CD/PM, DD/PM)	Establish and enact post-pilot governance structure (DGT)	Meeting #4 (Evaluation and Wrap-up)	Wrap up development phase and plan for next funding phase (UEDL, TDP1, TDP2, TDP3)

	Establish customer service infrastructure (DD/PM)	Establish post-pilot educational program (DGT)		Additional user engagement as needed for future planning (UEDL)
	Continue to conduct sustainability planning (CD/PM, DD/PM)	Establish plan for post-pilot community engagement (DGT)		Final validation of all APIs and user interfaces (UEDL, TDP1, TDP2, TDP3)
	Fiscal management (PI, co-PI, AC)	Publicly release best practices guidelines, online resources, and other informational products (DGT)		Develop reports: admin, editing (UEDL, TDP1, TDP2, TDP3)
	Plan for future promotional and outreach strategies (DD/PM)			
	Manage reporting and overall milestone tracking (CD/PM, DD/PM)			
	Internally and externally report findings from program assessment and evaluation (CD/PM, DD/PM)			
	Create final grant reporting as needed (CD/PM, DD/PM)			

EXPECTED OUTCOMES

The SNAC Cooperative pilot will have two major categories of expected outcomes. The first category is social, and covers administration, governance, and the member community.

- SNAC Secretariat at NARA is established with staff fully proficient in administering the ongoing operations and governance of the Cooperative.
- A governance structure is in place that includes at a minimum a Steering and Policy Committee; additional structural components will be developed in collaboration with members and may include, among others, committees to address editorial policy, technology standards and best practices, and research user services.
- A core thirteen-member community is established with a shared understanding of and commitment to the cooperative's objectives and functions. The inaugural members will be able to proficiently curate Cooperative data, and participate in Cooperative governance, and recruiting and training new members.

- Fifteen to twenty new members will have been recruited for the second two-year phase of establishing the Cooperative. The number of new members will be contingent on the Cooperatives' capacity to train new members.

The second category is technological.

- Editing User Interface (EUI) will support adding new descriptions; revising existing descriptions; revising relations in the social-document network; merging descriptions; spitting descriptions; and declaring descriptions obsolete.
- Data Maintenance Store (based on PostgreSQL) provides support for primary storage and maintenance of data, and required services for dependent subsystems.
- Identity Reconciliation subsystem is in place to provide feedback through the EUI and to be used in batch ingest of descriptions derived from external sources.
- Graph Data Store (based on Neo4J) is in place to support both graphical social-document network displays in History Research Tool and exposing graph subset of SNAC data as RDF/LOD.
- RESTful API in place to support third party editing interface applications.
- Component Subsystem Integration middleware integrates Cooperative platform subsystems.

At the end or near the end of the two year pilot, the Cooperative staff and members will have developed a strategic plan for an additional two-year effort that will consolidate the foundation that has been put in place in the pilot, address gaps, identify collaboration opportunities, expand Cooperative membership, and begin the process of transitioning to a self-sustaining business model.

During the recruitment of inaugural members, it became clear that there was significant potential for expanding the member community. One outcome of the pilot will be to recruit new members for the second two-year phase of establishing the Cooperative. Major institutions that have expressed an interest in becoming members of the Cooperative are the British Library, the Bibliothèque nationale de France, Archives nationales France, and the Staatsbibliothek zu Berlin. Thus one outcome of the pilot will be a list of potential new members with contact information. Based on informal expressions of interest, we are confident that we will be able to recruit fifteen to twenty new members.

The technological platform outcome will support third party applications from professional and scholarly communities that will expand the means to participate in the Cooperative. It is anticipated that Lyrasis, New York Public Library (NYPL) and Yale University will collaborate to revise ArchivesSpace to incorporate SNAC editing functionality, and to revise the archival access public access systems of NYPL and Yale to incorporate SNAC data. The Virginia Foundation for the Humanities is planning the development of an open source platform to support the full processing lifecycle of documentary editions, and plans to incorporate a SNAC editing interface into this platform. Finally, the SNAC technological platform will enable the Staatsbibliothek zu Berlin to develop the means to contribute data from Kalliope to SNAC. Kalliope is a collaborative project that has descriptions of nearly 600,000 persons and organizations linked to 2.5 million descriptions of letters, manuscripts, personal documents, albums, diaries, lecture notes, photographs, posters, movies, screenplays, music, and notebooks representing the holdings of more than 400 repositories in central Europe.

It is anticipated that at the end of the two-year pilot the long-term benefits of the Cooperative for both the professional curatorial community and the research user community will not be fully realized and established. Nevertheless, the social and technological outcomes will have established a solid foundation upon which an additional two years of effort will be able to complete establishing the Cooperative as a sustainable, broadly supported program that is a substantial component of the international scholarly communication infrastructure.

LONG-TERM SUSTAINABILITY

In order for the Cooperative to be sustainable, it will require a long-term business model that secures ongoing resources. NARA's commitment to serving as the Secretariat of the Cooperative and, in that role, committing staff to administer operations and governance *substantially* ameliorates the challenge. The primary ongoing costs of sustaining the Cooperative will be ongoing support of its technological infrastructure and travel and events associated with Cooperative governance. The challenge of securing ongoing support for the technological infrastructure may also be ameliorated by in-kind contributions to the open source software.[21] The NARA Chief Financial Officer has affirmed that during the pilot NARA will establish a trust fund account that will ensure that the Secretariat will be able to manage Cooperative funds held in an account dedicated to the activities of the Cooperative.

During the pilot, the Cooperative Steering and Policy Committee/Project Management Team (CSPC/PMT), chaired by the CD/PM, will develop, consulting with Cooperative staff and pilot members, will develop estimates of the scope of ongoing resources essential to a sustainable business plan. The source or sources of funds to support ongoing technological and governance costs have not been identified, but source options will be actively explored the members of the CSPC/PMT.

One possible source of funds is membership fees. The ongoing assumption in the Cooperative planning has been that initial membership in the Cooperative would not involve membership fees. The rationale for this assumption is that the Cooperative would not initially provide pilot members with substantial substantive benefits in processing efficiencies and enhanced access to their holdings. Rather the pilot members would be willing to make substantial commitments of staff time to assist in defining and establishing the Cooperative based on its *potential to provide benefits*, with the long-term understanding that ongoing membership may involve fees once the value of the Cooperative is established for them.

During the pilot, the Cooperative members will, among other contributions, be asked to assist in determining the potential value of the Cooperative to members. At this early stage, the value of the Cooperative as well as resources essential to establishing a sustainable business model are not fully understood. Developing such an understanding and a sustainable business model that addresses it is an objective of the pilot. Ultimately the value of membership will be contingent on the Cooperative being substantively established and operating, with concomitant commitments on the part of member institutions to adjust procedures and processing systems to realize the full value of the Cooperative. It is at this point of development that we will be able to demonstrate that the Cooperative can make the description archival holdings more efficient and effective and significantly ameliorate the time-consuming challenge of discovering, locating, and understanding the resources that document our history. Membership fees as a source of funding will be explored over the course of the pilot with the inaugural members.

In the long term, membership fees alone may not be sufficient for a sustainable business model. For ongoing development and maintenance of the Cooperative infrastructure, it may be and likely will be necessary to secure additional funding. During the pilot, the PI, Co-PI, CD/PM, and DD/PM, in consultation with David Ferriero, Archivist of the United States, and Pamela Wright, NARA's Chief Innovation Officer, will explore securing support from foundations and gifts. Among the foundations and individual philanthropists to be explored are the following: Alfred P. Sloan Foundation; The Pew Charitable Trusts; the MacArthur Foundation; W.K. Kellogg Foundation; the Walton Family Foundation; Bill & Melinda Gates Foundation, and David Rubenstein.

CD/PM Campbell has extensive experience in raising funds for research and technology that make wide access to knowledge and information possible. While at the Library of Congress, she was instrumental in raising funds for American Memory, the National Digital Preservation Program, and the Educational Outreach Program for K-12. The PI and Co-PI have extensive experience in the development of funding proposals for digital archives, library, and humanities projects and programs.

IP ISSUES

The SNAC Cooperative (archival descriptive) data will be made publicly available under a CC0 (Creative Commons with no use constraints) license. This policy will align the Cooperative with the policy of the Digital Public Library

[21] The SNAC PI has been approached by "programming-archivists" eager to contribute in some way.

of America, and embrace the government and academic trend towards making information available for use without constraints. The Cooperative has and will ensure that all data contributors and Cooperative members grant written permission to make contributed data available under the policy of the Cooperative.

All Cooperative documentation (policy, governance, training, editing guidelines, and other educational or supporting documents) will be publicly available under a CC0 license.

The Cooperative will release all software and programs developed as open source using GitHub. Given that some of the software and programs are built on or employ open source software components not developed in the project, third-party open source software will be released according to its license.

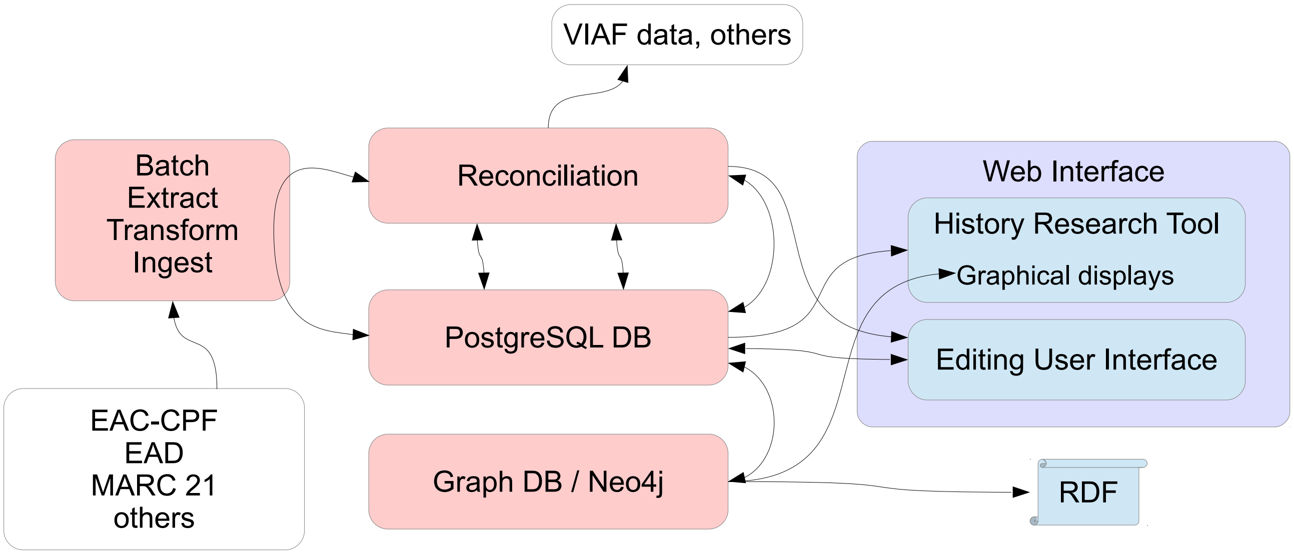
All source code produced by the project shall be Copyright the Rector and Visitors of the University of Virginia, and the Regents of the University of California, and shall be made available under the Open Source Initiative certified BSD 3-clause license.

Reporting

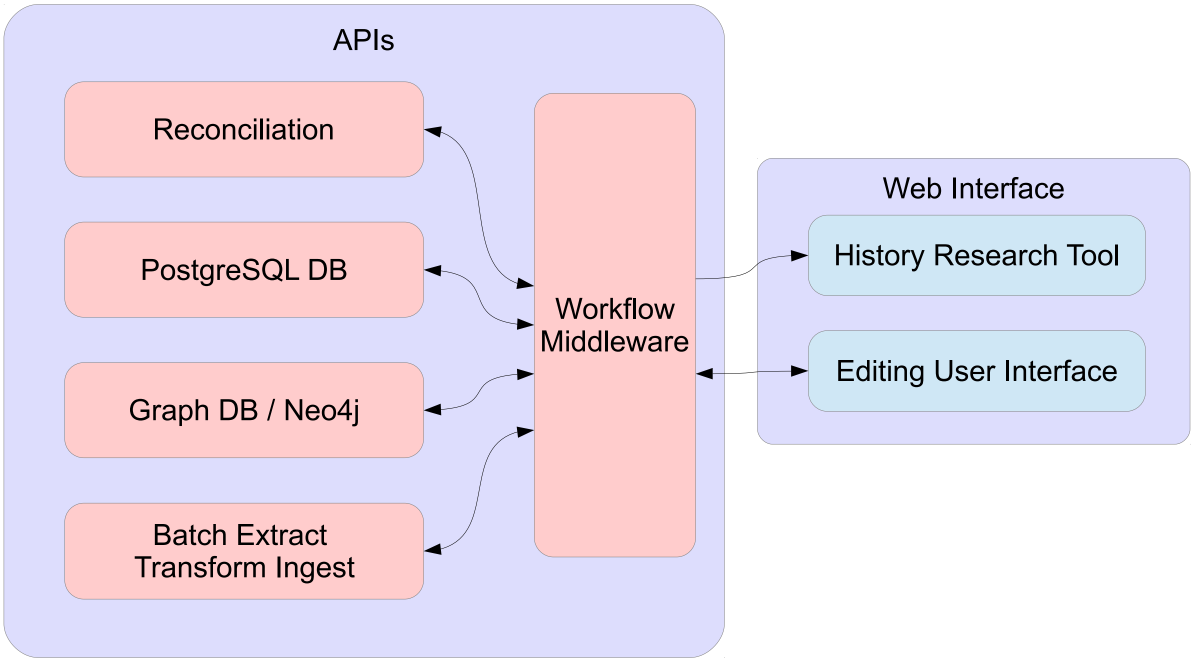
The University of Virginia will provide the Mellon Foundation with interim and final reports according to the schedule specified by the Foundation. Narrative reports assessing the progress and success of the project and commenting on the financial reports will be produced by PI Pitti, with assistance from Administration Coordinator Wells, and will be submitted by Pitti. Financial reports will be produced and submitted by the University of Virginia Officer of Sponsored Programs.

Appendix 11 System Architecture

Relations of SNAC system components (subsystems)



Relations Among Components


APIs
Reconciliation
PostgreSQL DB
Graph DB / Neo4j
Batch Extract
Transform Ingest
Workflow
Middleware
Web Interface
History Research Tool
Editing User Interface