SNAC Cooperative

An Archival Description Cooperative: The Vision

The initial steps towards realizing an international archival description cooperative were taken in January 2013. Daniel Pitti (SNAC R&D and Cooperative Planning Project Director), Laura Campbell (retired Chief Information Officer, Library of Congress), Anne Van Camp (Director of the Smithsonian Institution Archives), Don Waters (Program Officer, Andrew W. Mellon Foundation), and Clifford Lynch (Director, Coalition for Networked Information) met with David Ferriero (Archivists of the United States) and several of his staff: Deborah Wall (Deputy U.S. Archivist), William Bonsanko (Chief Operations Officer), Mike Wash (Chief Information Officer), Pamela Wright (Chief Office of Innovation), and John Martinez (Office of Innovation). At this meeting, the United States National Archives and Records Administration (NARA) committed to host the cooperative. All also agreed that while NARA would serve as the secretariat for the cooperative, the technological infrastructure would be developed and maintained outside of NARA. In the following months, the California Digital Library (CDL), University of California Office of the President, committed to host the technological infrastructure, to be developed and maintained by the cooperative community as an open source undertaking.

Though these critical first steps have been taken, it is clear that it will take many years to develop and fully establish a sustainable international archival description cooperative program. It is an exceptionally complex undertaking that will involve a wide array of interrelated administrative, social, legal, and technological details. As the administrative host, NARA will, among other things, need to address the business operations, manage and promote membership, manage public and research community relations, oversee community-based governance, and coordinate and direct the ongoing development and maintenance of the technological infrastructure. It will require staff with a variety of skills and competencies: financial operations, communications and promotion, archival description, community outreach, a firm understanding of established and emerging technologies, and more. The development and maintenance of the technological infrastructure will also require a variety of different specialized knowledge and skills, such as user experience, data migration and refinement skills, identity resolution expertise, and design, database, markup, and graph programming skills.

NARA's and CDL's commitments establish the basic foundation on which to establish and build the Cooperative, but before articulating a plan for the next steps, we first need a clear understanding of our vision for the Cooperative.

A. International Scope

The world's historical records are all interconnected. While there are national borders, and social, legal, and cultural boundaries, people and nations find ways to interact with one another, in ways peaceful and not. Borders and boundaries prove over time to be porous, as commerce, ideas, and creative works move back and forth. People migrate, building new lives, new social networks, and all the while maintaining ties to social networks they left

behind. All of this interaction is documented in the historical records that facilitate it. While establishing an international rather national cooperative increases the complexity of an already complex undertaking, an international level of cooperation will allow users to take full advantage of the interconnections, to the economic and professional benefit of archivists and researchers. Thus the Cooperative should be international in scope.

B. Reliable Identity Data

The Cooperative's core objective will be to build a large collection of increasingly reliable descriptions of persons, corporate bodies, and families (CPF entities), each of which will function as a node in a vast network of interconnected descriptions of other CPF entities (both those described within Cooperative data set as well as outside) and descriptions of cultural resources, including archival records as well as library and museum resources. The outcome will be a vast, global social-document network, which will provide both the means to locate resources and a context for understanding them.

Data reliability is critical for processing archivists as well as researchers. A reliable description is a description that accurately identifies a unique CPF entity; two or more CPF entities should not be combined in the same description, and there should not be more than one description for each unique CPF entity. Such identity resolution represents the most basic challenge in authority description. Names are weak identifiers: many CPF entities may share the same name, and any given entity may have more than one name. Additional evidence is necesary to accurately identify each entity. This evidence may be ambiguous, conflicting, incomplete, or simply not available. Not surprisingly, authority work is considered one of the most labor intensive, expensive resource description activities: identity resolution often requires painstaking research and careful evalution in order to locate and weigh the necessary evidence, and further time and effort to accurately document the evidence in the description.

C. Technological Infrastructure

1. Batch Ingest of Identity Data

To date, SNAC has relied exclusively on batch processing of existing archival descriptive data. The bulk of the CPF descriptive data is derived from existing archival record descriptions, either EAD-encoded finding aids, or MARC descriptions in OCLC WorldCat that are archival, broadly defined. Based on complex algorithms, the data documenting CPF entities found in the record descriptions is extracted and assembled into EAC-CPF instances. An additional source of CPF descriptive data is derived from existing authority descriptions, such as the descriptions of nearly 300,000 CPF entities that are associated with the archival holdings of the British Library, or the tens of thousands of descriptions associated with the holdings of NARA.

SNAC has ingested the bulk of the data available in WorldCat and a substantial portion of the data available in EAD-encoded finding aids in the U.S. and the U.K., but new descriptive data will contintue to emerge from these sources. Even so, this data represents only a fraction of the data available globally. Thus the Cooperative will continue to employ batch processing and ingest methods to collect large sets of data as they are made available.

Currently many steps in the processing that could be automated depend on centralized, manual intervention. Source data is solicited. Arrangements are made for acquiring data from sources. Data from each source are evaluated with respect to quality, and if found

wanting in any respect, are manipulated to ensure that they can be processed. Steps in the process are then manually applied: extraction of data and assembling into EAC-CPF instances; instances are then put through a matching and merging process (identity resolution); and finally the resulting instances are made available in the public research system. Many of the steps can be automated, at least in part, and some responsibility for evaluating the quality of the source data and manipulating it can be placed in the hands of the source provider.

In a fully developed system, source providers would be able to independently perform a variety of tasks. The contribution process would begin with an application that can be reviewed and approved by Cooperative staff. An approved provider could use a web-based form to formally commit to providing data, sign an intellectual property agreement, and enter institutional information essential to the integration of the data into the Cooperative data collection. The provider could then upload the data and execute evaluation processes that provide with status updates and feedback. A variety of tools for fixing the more common issues would be available for use. If necessary, the provider could consult with the Cooperative staff. Once data is ready for processing, the provider would perform the extraction-CPF description assembling process, with the results made available to the Cooperative for final evaluation and ingestion into the Cooperative data collection.

2. Manual Maintenance of Identity Data

While computational methods are an effective means for separating identity description from record description, algorithms alone are insufficient for creating reliable identity data. In part this is due to the uneven quality of the source data: typographical errors; names poorly formed; incorrectly identified with respect to type of name (corporate body, person, or family) or mistagged as names; insufficient data to uniquely identify; ambiguous biographical or historical data for an identity; data tagged as biographical or historical data that is in fact not; and other quality issues.[1]

There are additional challenges and limits to what can be accomplished through computer processing alone. The description of CPF entities is intermixed with the description of archival resources. When more than one CPF entity is associated with a described resource, the biographical or historical data, or occupations, functions, subjects, and associated places, presented may be ambiguous, which is to say, it may not be clear which data belongs to which named CPF entity.

If the quality of the source data were better, this would greatly facilitate the effectiveness of the computational methods. Given the complexity and nuance of identity resolution, however, computational methods alone will never be sufficient. Further, the ambiguity inherent in the EAD and MARC encoding schemes undermines the overall quality of what can be accomplished computationally. In order to create high quality, reliable identity data, it is essential that trained professionals have the means to edit the data assembled through computational means and to directly contribute new data.

The fully developed Cooperative system will thus support manual maintenance of the CPF descriptions through an editing interface. The manual maintenance system must support a wide array of functions. Among the functions supported by the maintenance system will be the following:

[1] In fairness to those who created the descriptions, they never anticipated the data would be used in the way in which it is being used.

- User and permissions management.
- User login and authentication.
- Tracking of user transactions.
- Cooperative member reports (for example, lists of descriptions associated with the member's holdings).
- Administrative reports.
- Creation of new CPF instances.
- Revision of existing CPF instances.
- Adding and deleting data fields.
- Linking CPF instances to other CPF instances.
- Linking CPF instances to a wide variety of external resources:
    - Finding aids and catalog records.
    - CPF descriptions in external authority files such as VIAF, ISNI, NACO/LCNAF, GND (Deutsche Nationalbibliothek) and Autorités BnF
    - Cross-domain digital resources such as Digital Public Library of America (DPLA) and Europeana
    - Digital humanities projects and programs, such as the Walt Whitman Archive and Chaco Research Archive
    - Allied social-document programs such as Kalliope, People of the Founding Era, Collective Biographies of Women.
    - Internet-accessible resources such Wikipedia.
- Splitting CPF descriptions when two or more identities have been incorrectly combined.
- Move data from one CPF description to another when data as been incorrectly associated with an identity.
- Merge CPF descriptions when two or more are found to be for the same identity.
- Flag as obsolete CPF descriptions that in fact do not describe a valid CPF entity.
- Flag as obsolete/split CPF descriptions that are split into two or more CPF descriptions.

3. Integration into Archival Management Systems

Archival context description is an integral component of archival processing. As such, it will be integrated into archival management systems such as ArchivesSpace. Integration into archival management systems will provide computer-assisted description of archival holdings and facilitate the efficient linking of local record description with CPF contextual data.

The same technical maintenance infrastructure that supports the Cooperative editing interface will also support editing interfaces integrated into archival management systems. All of the Cooperative editing functions described above will be available in the archival management systems.

When processing and describing an archival collection, archivists will be able to consult the Cooperative authority file for established CPF identities. When the identity has already been established, the processing archivist will be able to select the identity description, linking it through a persistent identifier to the collection being described and linking the collection description through a persistent identifier with the identity description. If the archivist needs to revise the identity description in any manner, the revision can be performed within the archival management system.

Integration of a Cooperative editing interface directly in the tool used to manage and process archival holdings will maximize the economic benefits of the shared Cooperative contextual data. It will also provide an efficient means to create the association between local holdings and cooperative data, associations that will make the holdings discoverable in the global context, and provide enhanced, reliable context data along side of the description of records in the local access system.

4. Public Interface: History Research Tool

Scholars, teachers, students, and others 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. Discovering, locating, and using these distributed historical records to understand the past are daunting, time-consuming activities.

The international archive community has made great strides in ameliorating this challenge through the development of public catalogs that integrate access to state or regional holdings. One significant international program within the European Community, for example, is Archives Portal Europe.[2] Nevertheless, these access systems, while decreasing the number of places researchers must explore, are isolated from one another. Researchers must still search scores of different archival access systems to find all relevant records. In order to know where to look, they need to know in advance where records are likely to be found. The process involves painstakingly accumulating and following clues, hunches, and leads, and is often a frustrating, time-consuming process of elimination. Employing such methods, it may take a scholar years of persistent, focused work to locate relevant resources, knowing that many clues are buried in complex and detailed archival descriptions.

The Cooperative public access system will serve as a History Research Tool that will not only ameliorate the research challenge of discovering and locating archival resources, but also reveal the social and intellectual connections of the people documented in the resources. Tracing the social networks of the people documented in the resources will improve understanding of the resources and the broader context within which they were produced, and will also provide pathways to related resources that might otherwise be overlooked. Cooperatively maintained social-document data will provide researchers with unprecedented research economy, enabling them to go to one rather than many places to discover, identify, and locate resources. Further, the social network data will enable them to easily see the social, professional, and intellectual connections between the people documented in historical records.

Cooperatively maintained descriptions of persons, organizations, and families will serve as the primary foundation for the History Research Tool. Users will be able to search the biographical-historical data using the familiar keyword search. For more sophisticated searching, they will be able to limit searches to particular components in the description: names, biographical information, places, occupations, and more. Alternatively, users will be able to browse alphabetically through names, either all types of names, or limited to people, organizations, or families. Search results can be filtered through facets such as occupation, associated places, location of associated archival resources, and so on.

[2] http://www.archivesportaleurope.net/

Each CPF description provides a constellation of data about a particular entity and relations to other CPF entities and archival resources, other cultural heritage resources, and complementary authority and biographical resources for the same entity. The History Research Tool will enable users to explore all of these links. Links to archival resources provide a comprehensive list of available historical records. To assist users in planning research travel, the locations of repositories holding relevant records will be presented in geographic visualizations. Links to DPLA, Europeana, and thematic scholarly research sites will lead users to digital historical resources.

D. Cooperative Administration

The National Archives and Records Administration (NARA) Office of Innovation serves as the SNAC Cooperative Secretariat. The Cooperative, under the direction of the Secretariat, provides ongoing administrative support for members and for the elected leadership of the Cooperative governance. NARA Innovation Office staff members have the responsibility of managing membership, communications, training, data contributions, meeting and event planning, archiving of decisions, statistics and reporting, and the Cooperative administrative web site.

1. Administrative Support

- Maintain and update the Governance Document, Strategic and Operational Plans as necessary
- Oversee elections: post candidate biographies; prepare, distribute and tally ballots, post results
- Establish and maintain a registry of all members and their contact information, terms of office
- Coordinate and maintain the Cooperative archives
- Recruit new members
- Coordinate and maintain the official calendar
- Serve on the Steering and Policy Committee and work closely with the standing committees
- Serve as a voting member of the Steering and Policy Committee representing NARA
- Administer funding in collaboration with the Cooperative Technology Host

2. Communication

- Oversee Cooperative communications
- Coordinate and maintain the postings to the Cooperative web site to include:
    - Member information, events, meetings, authoritative changes to policies and guidelines
    - Frequently Asked Questions
    - Workflow and Procedural Guidelines
    - White Papers, Articles of Interest, Member News of Interest
- Maintains Cooperative presence via social media such as Twitter, Facebook, and blogs

3. Training and Data Contributions

- Oversee the development of training for new members
- Facilitate training by helping to identify trainers and logistics
- Work closely with the standing Training Committee

- Maintain training documentation
- Coordinate training events and publish the schedule

4. Meeting and Event Planning

- Arrange ALA and SAA meeting logistics: meeting scheduling and locations
- Arrange Steering and Policy Committee and Operations Committee meetings
- Record minutes of the various committee meetings or coordinate the receipt of those minutes
- Provide support, where needed for Cooperative committees and ad hoc working groups

5. Statistics and Reporting

- Define key statistics and historical comparison data of value to the Cooperative membership
- Collect, compile and report on those metrics and statistics, where appropriate post on web site
- Draft the Cooperative Annual Report for delivery to NARA management and the Cooperative governance leadership
- Post final Annual Report to Cooperative web site

E. Cooperative Governance

As a cooperative program, members of the SNAC Cooperative community are responsible for its governance. The Cooperative is a democratic organization whose mission and policies are determined by its membership. Governance will be performed by elected and appointed representatives of participating organizations and programs, who will ensure that the Cooperativ'es mission and operations address shared objectives that benefit archives and the users of archival resources. Governance will also ensure that the policies and services of the Cooperative adapt and develop in response to new ideas and emerging opportunities.

The governance structure includes the Advisory Board, the Steering and Policy Committee, and the Operations Committee. Additional standing committees focused on technology, standards, and training will provide informed recommendations to the Steering and Policy Committee and the Operations Committee.

1. Secretariat

The Cooperative Secretariat is the National Archives and Records Administration. It provides leadership and oversight of Cooperative mission and goals. Its governance responsibilities include participating in and facilitating the workings of the Steering and Policy Committee, and Operations Committee. It manages and oversees the election of the members of the Steering and Policy Committee. The Secretariat creates and maintains the registry of member institutions and individuals including member identifiers, emails, telephone numbers and addresses; coordinates the list serves; arranges annual meetings; and records and maintains the Cooperative archive, including minutes of the Steering and Policy Committee and other governance committees; coordinates the communication mechanisms for the Cooperative via email, brochures, posters, newsletters and other social media; oversees the maintenance of documentation; coordinates training of new members; and coordinates the Program statistics.

2. Advisory Board

The Secretariat and Steering and Policy Committee shall identify a set of advisors from the membership of the Cooperative, from allied cultural heritage programs, and from the research user community to provide oversight and guidance with respect to the mission and operations of the Cooperative and the relation of the Cooperative to allied cultural heritage and scholarly programs and service organizations. The Advisory Board will also provide topics and suggestions with respect to Cooperative strategic planning.

The Advisory Board consists of leading experts from the archival, library, and museum communities; representatives of strategically allied programs and services; and researchers and scholars representing a range of user communities: historians, genealogists, local historians, documentary editors, and the like.

The SNAC Advisory Board shall be active in developing relationships with other similar organizations related missions and seek their input on common issues and work solutions. This may include bibliographic data and standards initiatives that are national and international in scope. SNAC Advisors will provide comments and advice on SNAC policy issues keeping the Steering and Policy Committee informed on current developments in the broader community.

3. Steering and Policy Committee

The Steering and Policy Committee performs the governance of the Cooperative as a whole. This includes developing, reviewing and approving long-term strategies, plans, goals; initiating, reviewing, and approving policies in all areas; reviewing, and approving new initiatives within the Cooperative; initiating relations with allied programs and services; establishing criteria for membership; reviewing resource implications of technical and operational initiatives; and establishing special initiatives as needed.

The Steering and Policy Committee will have both permanent and elected members. The permanent members of the Committee will be representatives from the Secretariat and major participating national repositories in the international community. Other members of the Committee will be representatives elected by the voting members of the Cooperative. The elected officers of the Committee will be the Chair, Chair-Elect, and Past-Chair. The Steering and Policy Committee has ex-officio, non-voting members: representatives from the Operations Committee and the Technology, Standards, and Training standing committees.

4. Operations Committee

The Operations Committee is charged with efficiently monitoring and maintaining program functions and activities both locally and across the overall program. The Operations Committee responsibilities will include monitoring operational procedures and suggest needed changes; receiving recommendations from members; developing and maintaining documentation; participating and contributing to development of standards by proposing, reviewing and or commenting upon changes to rules, formats, and standards in close coordination with the Standing Committees.

The Chair of the Steering and Policy Committee in consultation with the Secretariat will appoint the members of the Operations Committee.

5. Standing Committees and Ad Hoc Working Groups

The Cooperative will have small, agile standing committees focused on critical components of the mission and operations: the Standing Committees for Technology, Standards, and Training. The Chair of the Steering and Policy Committee in consultation with the Secretariat will appoint the members of the committees, and as necessary, may also create ad hoc working groups to address strategic challenges and opportunities that may arise.

The Technology Committee's responsibilities will include reviewing and evaluating the maintenance and user interface technological infrastructure in order to ensure that the most appropriate, effective, sustainable technologies are being employed; monitoring and making recommendations with respect to emerging, promising technologies; recommending technology initiatives; developing strategies and specifications for improving the functionality of the technology infrastructure; making recommendations on developer staffing; and working with vendors and utilities that have critical allied relations with the Cooperative.

The Committee on Standards' responsibilities will include evaluating and recommending intellectual content standards; evaluating and recommending technology standards; representing the Cooperative in the development of critical technology and intellectual content standards; developing best practices guidelines and documentation for Cooperative data creation and maintenance; developing methods for ongoing evaluation of data quality and reliability; and making recommendations for improvements in data processing, in particular improvements in computation-based identity resolution.

The Committee on Training is responsible for assessing the need for training, identifying who needs training, and what type of training. This work will also include identifying a ready cadre that can both be trained as trainers. and provide training to new members. The Chair of the Training Committee shall work closely with the Secretariat to plan and schedule annual training workshops and with the SAA Training and Continuing Education Division. The Committee will also develop documentation that supports the creation of high quality Cooperative records and resources. The Secretariat and Training Committee maintain a roster of active trainers.

6. Funding

The National Archives and Records Administration provides staffing support for the Secretariat (a minimum of 2 FTE). Member fees and donations support ongoing development and maintenance of the technical infrastructure. Major technology initiatives will be funded through grants and in-kind support.