Social Networks and Archival Context

A Proposal to The Andrew W. Mellon Foundation

PROPOSAL SUMMARY

The Institute for Advanced Technology in the Humanities (University of Virginia; IATH) in collaboration with the School of Information (University of California, Berkeley; SI/UCB) and the California Digital Library (University of California; CDL) propose to vastly expand *Social Networks and Archival Context* (SNAC)[1], a research and demonstration project. Funded in part by a grant from the National Endowment for the Humanities, the initial pilot phase of SNAC has demonstrated the potential for transforming scholarly historical research by dramatically improving access to resources that document the lives, work, and events surrounding historical persons, and by providing unprecedented access to the biographical-historical contexts of the people documented in the resources, including the social-professional networks within which the people lived and worked. In the next phase of the project, we propose to vastly expand the quantity and diversity of the source data, and extend and expand the research. While the immediate objectives of the project are to significantly refine and improve the effectiveness of the methods used in building and make an innovative research tool Internet-accessible, the long-term objective is to provide both methods and data as a solid foundation for establishing a sustainable national archival authorities program cooperatively governed by and maintained by the archive and library professional communities.[2]

The SNAC project is endeavoring to address the longstanding research challenge of discovering, locating, and using distributed historical records, and at the same time, to build an unprecedented resource that provides access to the socio-historical contexts in which the records were created. SNAC offers scholars studying the lives, work, and events surrounding historic persons three unprecedented forms of access: 1) integrated access to distributed primary (archival) and secondary (published) resources by and about persons, families, and organizations; 2) access to biographical and historical descriptions; and 3) access to the social and professional networks within which people lived and worked. [3] As more than one historian has observed, the data assembled and interrelated in each SNAC record would take a year or more to compile in the current research environment, with successful discovery and assembling of the data highly dependent on persistence and serendipity. Indeed it is likely that some of the information found in the SNAC records might never be discovered using current methods.

The SNAC research tool is using advanced technology to perform a series of processing steps that lead to a large collection of descriptions of persons, families, and organizations. These descriptions are then used to build a prototype public access system and historical resource. In the initial step, descriptions of people are extracted from descriptions of their archival records and assembled into an international standard representation, Encoded Archival Context–Corporate Bodies, Persons, and Families (EAC-CPF). Alternatively, original authority descriptions in a variety of formats are transformed into EAC-CPF. The resulting standardized authority records are matched against one another, and matching records are merged and then matched against existing library and museum authority records from which additional information is extracted and added to the EAC-CPF descriptions. The resulting authority descriptions are linked together in compelling new ways to bring together the descriptions of people interrelated with one another, and to resources by and about them. The authority descriptions (or records) resulting from this process are then used to build a novel SNAC historical resource and access system.

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

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

[3] See project prototype http://socialarchive.iath.virginia.edu/xtf/search

In the current phase of SNAC, the archival data employed is limited to 30,500 finding aids contributed by three archival consortia (Online Archive of California; Northwest Digital Archives, and Virginia Heritage) and the Library of Congress. In order to achieve both the near and long term objectives of the project, both the quantity and diversity of the data need to be increased, and the research agenda expanded. We are currently requesting $574,824 of funding for a two-year project. The finding aids will be increased to more than 148,000 contributed by fifteen consortia, over thirty leading American repositories, and the ArchivesHub in the U.K. In addition, OCLC WorldCat will contribute one to two million MARC collection-level archival descriptions, and the National Archives and Records Administration (NARA), Smithsonian Institution, New York State Archives, British Library (BL), Archives nationales (France), and Bibliothèque nationale de France (BnF) will contribute over 375,000 original archival authority records. The Virtual International Authority File (VIAF),[4] and The Getty Vocabulary Program[5] will contribute approximately 16.2 million authority records that will be used to enhance the archival data. By expanding the quantity and diversity of the data, the project will be able to further develop its processing, indexing, and display methods, public interface design, as well as address the challenge of scale.

PROPOSAL NARRATIVE

Background of Institution

The central purpose of the University of Virginia is to enrich the mind by stimulating and sustaining a spirit of free inquiry directed to understanding the nature of the universe and the role of mankind in it. Activities designed to quicken, discipline, and enlarge the intellectual and creative capacities, as well as the aesthetic and ethical awareness, of the members of the University and to record, preserve, and disseminate the results of intellectual discovery and creative endeavor serve this purpose. In fulfilling it, the University places the highest priority on achieving eminence as a center of higher learning.

Founded in 1992 at the University of Virginia, the Institute for Advanced Technology in the Humanities (IATH) is one of the world's leaders in transforming humanities research through the application of computing and network technologies. IATH is a research unit within the University of Virginia and reports to the Provost's Office. The University of Virginia funds IATH, and additional funding comes from grants and gifts. Since IATH was founded, it has been awarded over $12 million in grants. Our goal is to explore and develop information technology as a tool for scholarly humanities research. To that end, we provide our Fellows with consulting, technical support, applications development, and networked publishing facilities. We also cultivate partnerships and participate in humanities computing initiatives with libraries, publishers, information technology companies, scholarly organizations, and other groups residing at the intersection of computers and cultural heritage. The research projects, essays, and documentation presented here are the products of a unique collaboration between humanities and computer science research faculty, computer professionals, student assistants and project managers, and library faculty and staff. In many cases, this work is supported by private or federal funding agencies. In all cases, it is supported by the Fellows' home departments; the College or School to which those departments belong; the University of Virginia Library; the Vice President for Research and Public Service; the Vice President and Chief Information Officer; the Provost; and the President of the University of Virginia.

[4] Founded by OCLC Research, the Bibliothèque nationale de France, Library of Congress, and the Deutsche Nationalbibliothek, VIAF is an international cooperative aggregation of authority records. The contributors currently include 20 national or major research libraries and The Getty Vocabulary Program. See http://viaf.org/

[5] The Getty Vocabulary Program will contribute the Union List of Artist Names (ULAN). While versions of the ULAN records will appear in VIAF, the ULAN records contain biographical data that is not accommodated in the VIAF record format.

IATH's projects are primarily faculty-driven, though many projects involve digital library research and development in collaboration with the Library and other institutions around the world. With an abiding commitment to long-term preservation and access to humanities collections and research, the Institute is dedicated to the promotion and use of international archive, museum, library, and humanities standards and to the use of open source software.

Rationale

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

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

In the first phase of the project, SNAC has designed and developed or adapted open source software to extract and normalize the names and descriptions of people and assemble the data into archival authority descriptions; to match or resolve identities (determining whether two or more identical or matching strings are for the same entity or two or more entities) and for merging authority records determined to be for the same entity; and to implement a functionally rich public prototype system that provides integrated access to resources by and about described entities and access to biographical-historical information about those entities. The first phase of SNAC has demonstrated that it is feasible to extract the description of people from archival description and re-purpose this descriptive data to build a novel and powerful research tool.

While all of the software developed has been sufficiently effective to demonstrate the feasibility of the methods and technology being employed, both the quality and the quantity of the products are not sufficient for establishing a sustainable archival authorities program. The archival source data represents a fraction of the data available, and the narrow focus of the data set (three similar consortia and the Library of Congress) has constrained the general applicability of the software (particularly the extraction and matching software). Further, while great strides have been made in the research, it is sufficiently complex to require ongoing work and refinement of existing techniques, and for incorporating new techniques. Much work remains to be done on the existing software to increase the range of archival description it will accommodate, to increase the qualitative effectiveness of the techniques, in order to build a rich, high quality foundation for establishing a national program.

In addition to refining and extending ongoing work, other critical areas of research and development are essential in order to maximize the effectiveness, accuracy, utility, and general applicability of the research products being developed in SNAC. In the second phase of SNAC, the following *new* areas of research and development will be addressed: developing programs for extracting and assembling authority descriptions from MARC archival descriptions; developing programs for transforming original archival authority descriptions into a common, standard archival descriptive format; accommodating authority descriptions in the public interface in languages other than English; developing techniques for normalizing and adding coordinates to geographic names in biographical-historical descriptions; developing timeline-map rendering of chronological biographies or histories (lists of date, place, and event); developing an public interface that will enable scholarly users of the prototype to query social-professional networks; developing graphical displays of complex, dense networks; and developing graphical displays of organizational charts, and sequential displays of organizations merging or dividing.

While many of the names found in archival descriptions are explicitly identified as such, many names in biographies and histories, and in lists of correspondence are not. These names are either in natural language contexts, or in semi-structured descriptive contexts intermixed with non-name elements. The names of correspondents, in particular, are especially important for maximizing the populating of social-professional networks. In the second phase of SNAC, a major focus of research and development will be the development of effective National Language Processing (NLP)/Name Entity Recognition (NER) programs in order to increase the number and improve the quality of the names extracted from the finding aids.

Related Projects

To the best of our knowledge, SNAC is the only project utilizing a set of archival descriptive data to build a scholarly research tool on this scale. Though both SNAC and the People of the Founding Era (PFE)[6] focus on names and social networks, SNAC differs from PFE in three significant respects. First, SNAC is relying on archival description as the primary source of data, augmented by library authority records. PFE is deriving its name information from documentary editions. Second, SNAC is not restricted to a particular period of history but is inclusive, from B.C.E. to the present day. PFE is focused, as its name suggests, on a specific time and place. Third, SNAC is fundamentally about access to and context for understanding archival resources (records and manuscripts). PFE is focused on facilitating the understanding of the socio-historical context of the founding era. SNAC and PFE have had preliminary discussions about future collaboration and the director of PFE has agreed to serve on the SNAC advisory board.

There are also a number of other scholarly projects that bear some similarity to SNAC. Most prominent among them is the Research-oriented Social Environments (RoSE), led by Alan Liu at the University of California, Santa Barbara.[7] RoSE is developing social software to enable students to build "social-document" networks, similar to those being assembled in SNAC. The SNAC project is collaborating with RoSE in two ways. First, it has made its data available to RoSE. Second, the two projects are sharing work on the graphical display of social-document networks. SNAC also has a long term interest in the social software developed in RoSE, as this is likely to influence transforming SNAC into a national archival authorities cooperative. The director of RoSE has agreed to serve on the SNAC advisory board.

Also worthy of mention are Phylo,[8] a project devoted to documenting the social-professional and intellectual networks of Philosophy, and the Crowded Page,[9] a project devoted to the social networks of two literary and artistic communities. The focus of both of projects is social networks, and both are limited to specific domains, and thus differ from SNAC in that the latter endeavors to be universal in scope.

Project Description

Persons, families, and organizations, over the course of their lives, generate and accumulate records that document their lives. These records are primary evidence used by scholars to understand people, their social and professional relations with others, the work they produced, and the events in which they participated. Such records come in many forms: letters or correspondence, notebooks, diaries, school records,

[6] http://documentscompass.org/projects/pfe/

[7] http://transliteracies.english.ucsb.edu/category/research-project/rose

[8] http://phylo.info/

[9] http://www.crowdedpage.org/

photographs, video, manuscripts (literary and non-literary), financial records, and so on. When the records of a person, family, or organization are collected by an archive, library, or museum, they are maintained and described as an integral whole, a collection.[10] In order to preserve and provide access to these records, archivists and librarians describe the collection of records in a finding aid or guide.

Finding aids describe collections progressively and hierarchically, beginning with a description of the whole, and then important parts of the whole (called a series), and parts of the parts. These descriptions sometimes terminate in the description of individual items, though this is exceptional. In describing the records in a collection, the cataloger provides the name and a detailed biography or history of the creator of the collection. In addition, the cataloger also *selectively* explicitly provides the names of other persons, families, or organizations documented in the records. Many other names are embedded in the description, though not identified as such.

Currently, the names, social-professional relations, and biographical-historical description lie buried in finding aids that are themselves available only in dispersed systems. The core objective of SNAC is to extract the names, social-professional and resource relations, and biographical-historical data from the dispersed archival descriptions; reassemble the data in standard archival authority descriptions; and use these authority descriptions to build a prototype publicly accessible system that will provide integrated or union access to distributed primary resources and, at the same time, access to biographical-historical information that will enable users to identify and learn about persons, families, and organizations, their histories, and the social networks in which they lived and worked.

History of SNAC

The National Endowment for the Humanities (NEH) awarded a $348,221 grant in 2010 for the initial two-year phase of SNAC (May 2010-April 2012). The University of Virginia contributed an additional $70,026 in cost-share. The NEH and University of Virginia funds have supported acquiring and processing data, and the development of a public prototype (see a detailed description of work below). The Andrew W. Mellon foundation awarded the University of California, Berkeley $20,154 in November 2011 for SNAC related work. Because of overly optimistic assumptions in the NEH work plan, the full cost of SI/UCB's ongoing development of match/merge algorithms and programs and the use of them in processing was funded only through the first eighteen months of the current project. The Mellon funds have enabled SI/UCB to employ a graduate student to assist in the development and application of match/merge processing programs for the last six months of the first phase of the current project. The total funding for SNAC to date has been $438,401.

The SNAC began May 1, 2010. Most of the objectives (and all key objectives) of the first eighteen months were met or exceeded. While it is anticipated that the current project will meet all proposed objectives by its conclusion in April 2012, some of the more challenging research activities will demonstrate significant progress, but not mastery.

The current project work plan for the initial phase has two sets of related activities:

1. Developing software to derive names and biographical/historical data from EAD-encoded finding aids, migrating that data into EAC-CPF records, and matching and combining data from VIAF, LCNAF, and ULAN authority records to produce the final EAC-CPF records.

2. Developing a prototype access system, based on Extensible Text Framework (XTF), open source software developed at the California Digital Library (CDL). The prototype access system will has three components: 1) display of the EAC-CPF records; 2) browsing and search of the EAC-CPF records; and 3) a proof of concept

[10] The technical term for the set of records generated or accumulated by a person, family, or organization is a "fonds," but it is also frequently called an "archival collection" or simply "collection."

API that allows other interested institutions to embed the prototype's functionality into their site.

The first set of activities involves two separate steps in the processing. The first step, extracting EAC-CPF authority records from EAD-encoded finding aids, is being done at IATH. The second step, matching and combining the EAC-CPF records with one another and with authority records, is being done at SI/UCB. The second set of activities, developing a prototype access and historical resource, is being performed at CDL.

The following is a summary of the major tasks that have been completed in the first eighteen months of the current project:

- EAD-encoded finding aids have been acquired from the Library of Congress (LoC) (1,159), Online Archive of California (OAC) (~15,400), Northwest Digital Archive (NWDA) (5,160), and Virginia Heritage (VH) (8,390). A total of approximately 30,500 finding aids were collected and made available on the project server. (IATH)
- Authority files have been acquired and indexed using Cheshire, a probabilistic XML-based indexing system being utilized in the match/merge processing. 161,771 ULAN records, 5,091,703 VIAF records, and 8,619,143 LCNAF records were indexed and made efficiently available for match processing. A sample set of EAC-CPF records have been used to successfully test matching and record combining using each of the authority record sources.
- Successfully generated 175,637 EAC-CPF records from 30,500 finding aids: Library of Congress: 43,702 EAC-CPF records derived from 1,159 finding aids. Average per finding aid: 37.7; Online Archive of California: 91,811 EAC-CPF records derived from ~15,400 finding aids. Average per finding aid: 5.96; Northwest Digital Archive: 24,949 EAC-CPF records derived from 5,568 finding aids. Average per finding aid: 4.5; and Virginia Heritage: 15,175 EAC-CPF records derived from 8,390 finding aids. Average per finding aid: 1.8
- Successfully matched and combined the derived EAC-CPF authority records against one another for all personal, corporate, and family names and against the personal names in VIAF. 123,920 EAC-CPF records resulted from matching from 158,079 records.[11] 93,033 merged person records resulted from 114,639 person records; 30,161 merged corporate body records were merged into 41,177 corporate body records (EAC-CPF/EAC-CPF matching only); and 1,669 merged families records resulted from 2,263 family records (EAC-CPF/EAC-CPF matching only)
- A prototype access system has been developed, with a first public release in December 2010 (http://socialarchive.iath.virginia.edu/xtf/search). Development has employed use cases and identifying baseline functionality. Refinement of the access system continues. A major focus has been the development of a graph database for storing and accessing social network relations in support of visualizing relations between persons and between persons and corporate bodies. Ed Summers, at the Library of Congress, has voluntarily contributed code for migrating the graph data into RDF.

During the initial phase of the current project, several technologies have been developed.

- IATH has developed a complex XSLT program for deriving data from EAD-encoded finding aids. Because the finding aids vary in quality with respect to the accuracy and consistency of tagging names, and in the forms of the names, the XSLT makes extensive use of regular expressions to normalize names, for example, stripping non-name components from the ends of strings, regularizing capitalization, inverting personal names in direct order to indirect order. The XSLT also utilizes regular expressions for matching biographical-historical entries with name entries when collections have multiple creators.

[11] Note that the difference in the total EAC-CPF records processed occurs because the 158,079 represent an earlier iteration of the extraction processing and the 175,637 in the latest extraction are pending match/merge processing.

- SI/UCB has developed code written in the programming languages C and Python for using an existing open source indexing tool (itself written in C and Python developed at SI/UCB, Cheshire, and has developed and implemented a name matching algorithms in C and Python for personal, corporate, and family names. Also utilized in the match process is the open source PostgreSQL, an object-relational database. The major focus of matching, to date, has been personal names.
- CDL has developed the prototype public system by writing extensions in Java and XSLT to the open source XML publishing platform, XTF, also developed by CDL, and has utilized and developed graph data, retrieval, and rendering programs based on additional open source software: SPARQL, TinkerPop (Stack and Neo4J) and the JavaScript InfoVis Toolkit (JIT) for visualization.

There are several outstanding tasks that will be completed in the last quarter of the current project. The most of important of which are listed below:

- Match and merge processing of personal names employing the LCNAF and ULAN.
- Match and merge processing of corporate names employing the LCNAF.
- Interrelating and "flagging" EAC-CPF records that are candidates for merging but lack sufficient information for identity resolution (to facilitate later human resolution).
- Implement dynamic searches in public prototype of WorldCat, Flickr, and perhaps other resources (WorldCat Identities?) using name entries.
- Implement links in public prototype to WorldCat bibliographic records for titles of published works found in VIAF records.
- Implement in public prototype, provisionally, Linked Open Data availability of EAC-CPF records using RDF mapping developed in Italy.
- Package extraction and match/merge code for release as open source. (Note: The software for the public interface is currently available as open source).
- Implement persistent identifiers for named entity records and, based on them, implement EAC-CPF to EAC-CPF relations links (rather than the provisional search currently employed)

SNAC: Second Phase

In the first phase of SNAC, the current project has focused on extracting and assembling the authority descriptions from 30,500 finding aids and augmenting the derived authority records with additional data from library and museum authority records. In the proposed second phase, the number of finding aids will be increased to more than 136,000, and will be augmented by one to two million MARC collection-level archival descriptions contributed by OCLC WorldCat. The WorldCat collection-level descriptions provide only a brief, top-level description of a collection, and not the detailed hierarchical description found in finding aids. The use of MARC collection-level description has been almost exclusively restricted to U.S. repositories. Though not as extensive as finding aids, collection-level descriptions nevertheless typically contain the name of the creator of the collection, and frequently include a brief biographical-historical description of the creator, occupation, and the names of other people with whom the creator is most prominently related. Because creating collection-level descriptions was common in the U.S., the one to two million MARC descriptions provide comprehensive national coverage of (minimally or fully) processed archival holdings. In addition, the National Archives and Records Administration (NARA), Smithsonian Institution, British Library (BL), Archives nationales (France), and Bibliothèque nationale de France (BnF) will contribute over 375,000 original archival authority records in a variety of formats. These archival authority records will be augmented with additional data from library and museum authority records: 16 million Virtual International Authority File (VIAF) records; and 120,000 Union List of Artist Names (ULAN) records. Please see Appendix One for a complete list of data contributors.

There are three primary areas of research and development in SNAC that have been initiated in the first phase of SNAC that are ongoing and iterative and thus will continue in the second phase: 1) extraction of data from EAD-encoded finding aids; 2) Identity Resolution (identifying when two or more matching or similar name strings represent the same entity), including matching algorithms and use of contextual data; and 3) development of techniques for extracting, recording, indexing, searching, and displaying social-professional graphs or networks, in particular techniques for addressing the graphical display of connectors or nodes with large numbers of links to other nodes).

SNAC processes the source data and creates the archival authority records that are the content of the prototype public historical resource and access system in three steps, each step being the responsibility of one of the three project partners. IATH is responsible for acquiring and managing all of the data from the contributing institutions. IATH is also responsible for extracting archival authority data from the contributed EAD-encoded finding aids and MARC cataloging records and assembling them into Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF) records. It is estimated that between 1.5 and 4 million EAC-CPF descriptions will be produced in the processing of the 1,136,000-2,000,000 finding aids and MARC records. IATH also transforms contributed archival authority descriptions that are received in an alternative format. Once the finding aids have been processed, the result is a set of EAC-CPF records. Each contains a single identified name, along with identification of the source finding aid or catalog record, and, in the case of creators, any biographical information, dates of existence, language or languages used, links to related people, etc., that are found in the source. Since EAC-CPF descriptions are derived independently from each EAD or MARC description, there may be multiple EAC-CPF instances representing the same entity. A key challenge, then, is to identify multiple EAC-CPF descriptions that represent the same entity and combine them into a single description.

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

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

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

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

The increase in the quantity and diversity of the source data makes it possible to further develop and test existing methods and techniques, but also to expand the research agenda. The increase in quantity of data will present issues of scale in the processing of the records, the performance of the interface, the rendering of large amounts of graph data (social-professional relations and relations between people and related resources. The following lists the major new research made possible by the increase in quantity or diversity of the data:

- Developing techniques for extracting and assembling authority descriptions from MARC archival descriptions.
- Developing techniques for transforming original archival authority descriptions into the EAC-CPF standard (NARA, Smithsonian Institution, and BL).
- Linguistic diversity (the inclusion of French descriptions) will present challenges in assembling, indexing, and displaying EAC-CPF records.
- Developing techniques for normalizing geographic names and dates in and developing timeline-map renderings of biographies and histories.
- Further develop NLP/NER techniques in order to increase the number and improve the quality of the names extracted from the finding aids.
- Developing a social network searching interface
- Developing graphical displays of complex, dense networks that enable browsing of the networks and access to descriptions of objects (people or resources) found in the networks.
- Developing graphical displays of organizational charts, and sequential displays of organizations merging or dividing.

In order to improve the name entity identification (NER) and extraction processing, and the accuracy and reliability of the matching (Identity Resolution), an outside consultant (Wisser) will conduct a manual quality review of sample data representing the entire processing cycle from extraction through the final matching that produces the EAC-CPF records used in the public prototype system. The consultant will be given a sample set of finding aids and WorldCat MARC records; the EAC-CPF records derived from in the initial processing step; the EAC-CPF records resulting from the EAC-CPF to EAC-CPF matching and merging; the final set of

EAC-CPF records after matching against the library and museum authority records; and all matching library and museum authority records. The consultant will evaluate 1) recall and accuracy of the extraction; 2) recall and accuracy of the derived EAC-CPF record; 3) accuracy and reliability of EAC-CPF to EAC-CPF matching; 4) thoroughness and accuracy of the merging of matching records; 5) accuracy and reliability of matching EAC-CPF to library and museum authority records; and 6) the thoroughness and accuracy of identifying non-duplicating data in matching authority records and adding them to the final set of EAC-CPF records. The methods employed by the consultant will be based on methods developed by OCLC Research in its building of VIAF and WorldCat Identities, as well as other research activities. OCLC Research has developed NER and Identity Resolution *scoring software* that can be used in conjunction with the manual review to provide accurate quantitative reports on the recall and precision of the NER and Identity Resolution used in SNAC. OCLC Research as committed to sharing documentation of the manual review methods and to making the scoring software available to the SNAC project. The consultant will devote ten days in the first and second years of the project to performing the manual review. The scoring will be performed at IATH.

The programmers working on XSLT extraction and assembling of EAC-CPF records; NLP/NER programs for person, corporate, family, and geographic names; match and merge programs; and the prototype public system will be responsible for thoroughly documenting all code developed by the project (including the methods for integrating the code into existing open source software used in the project. The programmers will have the assistance of a technical documentation editor. In the last quarter of the project, two outside reviewers will perform a quality review of the code and documentation and provide written reports.

The project will have an advisory board of eight scholars, archivists, and librarians. The scholars are leading members of the documentary editing, literary editing, history, and social history communities. The archivists and librarians are internationally recognized leaders in the professional community. The advisory board will meet once, late in the first year of the project. The scholars, archivists, and librarians will be asked to provide guidance in at least the following areas. First, what are the best ways to promulgate the use of SNAC in the scholarly communities that are most likely to benefit from it? Second, what are the best ways to build support for SNAC and for transforming it into an ongoing, sustainable, and cooperatively maintained program? Third, what opportunities are presented by SNAC for scholar, archivist, and librarian collaboration? Fourth, what are the most important and functions and features of SNAC for scholarly users and archivist and librarian users? Holding the meeting late in the first year will allow the project to have accomplished significant work that can serve as a basis for discussion, and will allow the project sufficient time to act on the advice of the board. See Appendix Five for a list of advisory board members.

Collaborators

IATH will be the lead institution in the project, and will collaborate with the School of Information, University of California, Berkeley (SI/UCB), and the California Digital Library, University of California (CDL).

In addition to overall management and coordination of the project, IATH will be responsible for acquiring and managing the 136,000 finding aids and approximately 21,000,000 archive, library, and museum authority records. Further, IATH will be responsible for extracting data from finding aids and assembling this data into EAC-CPF records; converting contributed original archival authority records into EAC-CPF; normalizing ill-formed names (for example, inverting personal names in direct order to indirect order), and dates in biographical or historical description; and development of timeline-map displays biographies and histories. IATH will work collaboratively with SI/UCB on NER for and normalization of personal, corporate, family, and geographic names, and georeferencing geographic names. IATH will also host the prototype public access and historical resource system.

SI/UCB will be responsible for matching the EAC-CPF records against one another, combining records that describe the same person, organization, or family, and then matching the resulting set of EAC-CPF records against VIAF, NACO/LCNAF, and ULAN authority records, adding additional data (alternative names, sex, country or countries of affiliation, titles of resources found in OCLC WorldCat, and, from the ULAN and NACO/LCNAF records, additional biographical-historical data). SI/UCB will also be responsible for assisting IATH in the development and use of Name Entity Recognition (NER) programs designed to locate and identify names not explicitly tagged as such in the finding aids. SI/UCB will work collaboratively with IATH on NER geographic name identification, normalization, and georeferencing.

CDL will be responsible for the ongoing development of the prototype public historical resource and resource access system. This development involves the indexing of the archival authority descriptions, faceted browsing, graphical display of historical social network, integration of timeline-map display developed by IATH, outgoing links to Linked Open Data biographical-historical resources related to the described entity, links to descriptions of archival (and other) resources, and finally exposure of the SNAC archival authority descriptions to facilitate use of the data by other projects. CDL will also employ face-to-face scholar and educator user testing to evaluate the usability and usefulness of functions and features.

Job Titles and Job Descriptions

Principal Investigator – Daniel Pitti will serve as the Principal Investigator. Pitti will be responsible for overall management of the project, and, with Martin, will supervise IATH staff working on the project.

Co-principal Investigator – Worthy Martin will serves as Co-Principal Investigator. Martin will assist Pitti in management of the project and supervision of IATH staff. In addition, he will oversee the development of the NLP/NER programs used in name and data extraction, date normalization and timeline-map display. He will collaborate with Pitti, Larson, DE/TP, and TMP in design and development of NLP/NER programs for personal, corporate, family, and geographic names. He will also supervise the TMP.

Project Coordinator – coordinating relations and communication with data providers, among project collaborators, planning of meetings; and coordinating and editing documentation. (IATH: Sarah Wells)

Data Manager and System Administrator (DM/SA) – acquire and manage several million EAD, MARC, VIAF, LCNAF, ULAN, and EAC-CPF records; maintaining platform for first and third steps of processing. (IATH)

Data Extraction/Transformation Programmer (DE/TP) – XSLT extracting of EAC-CPF instances from EAD and MARC descriptions; conversion of archival authority descriptions into EAC-CPF; Natural Language Processing name extraction from correspondence description; name normalization. (IATH, to be hired. See Appendix Four for job description.)

Timeline-Map Programmer (TMP) – date and geographic name normalization; development of timeline-map display of biographical-historical description. (IATH)

Lead Match/Merge Programmer – lead developer of match and merge algorithms and processing; and collaborate with Pitti, Martin, DE/TP, and TMP in design and development of NLP/NER programs for personal, corporate, family, and geographic names. (UCB: Ray Larson)

Graduate Student Programmer (GSP) – assist in development of match and merge processing. (UCB)

Prototype Access System Coordinator – coordinate the CDL's participation in the project and design of public prototype system, and quality evaluation of processing results. (CDL: Adrian Turner)

Prototype Access System Developer – develop the public prototype resource and access system. (CDL: Brian Tingle)

User Experience/Design Coordinator – conduct focused user studies with scholars. (CDL: Rachael Hu)

Quality Review Consultant – in each of two years, evaluate the accuracy of the extraction, matching, and merging processing based on a statistical valid sample. (Simmons College: Katherine Wisser)

Code and Documentation Reviewer Consultant I (CDRCI) – in the final quarter of the project, evaluate the XSLT code and documentation (Columbia University: Terrence Catapano).

Code and Documentation Reviewer Consultant II (CDRCII) – in the final quarter of the project, evaluate the all code and documentation developed in the project except the XSLT. (Washington & Lee University: Sara Sprenkle).

Principal Investigators

Principal Investigator – Daniel Pitti will serve as the Principal Investigator. He is Associate Director of IATH, the chief technical architect of both the EAD and EAC-CPF standards. In addition to his work with archives and libraries, he has extensive experience in the design and implementation of scholar-driven humanities research projects that employ advanced technologies.

Co-principal Investigator – Worthy Martin will serves as Co-Principal Investigator. Martin is an Associate Professor of Computer Science and Acting Director of IATH, and has been one of the prime information architects on numerous digital humanities projects through IATH.

Length of Project with Timeline

The project will last from April 1, 2012, to March 31, 2014.

Overview

The archival source data sets used in the project will be quite diverse and will differ substantially in the challenge each presents for processing. While the data will be in diverse formats, it will also vary in the degree to which it is structured and curated.

The least problematic of the data will be the WorldCat MARC archival descriptions. While they will represent the bulk of the archival source data, they also will present the fewest processing challenges. The name and descriptive data are well delimited and thus identifiable, and the quality of the delimited data, the descriptive content, will have generally been subjected to more careful formulation than that found in finding aids. Thus it is anticipated that writing programs for extracting name and descriptive data from MARC archival descriptions will be relatively straightforward, and after careful design, development, and testing, will only need to be applied a minimum number of times. Manual review and scoring will be used to refine and enhance extraction techniques if necessary. The only exception is biographical-historical notes, which may have names not otherwise found in the record. If these notes prove to be sufficiently rich in name data, upon manual review of a sample set, then NLP/NER processing will be applied to them.

The original archival authority records contributed by NARA and BL will be in an ad hoc XML format. The NARA records will present some processing challenges, as NARA authority practice with respect to government agencies is unique. Nevertheless, EAC-CPF was developed to accommodate this unique approach, and thus migrating the data should be straightforward. BL authority data is based on ISAAR(CPF), the International Council for Archives authority content standard. EAC-CPF is specifically designed to accommodate ISAAR(CPF), and thus the migration of the BL data should be straightforward.

In preparation for the second phase of SNAC, a prototype XSLT transformation was successfully written to extract and migrate Smithsonian Institution Archives (SIA) MARC bibliographic-encoded agency histories into EAC-CPF. The New York State Archive agency histories employ the same practice as the SIA agency histories. With some refinement of the existing prototype, these two sets of records should not present any major processing challenges.

The EAD-encoded finding aids will represent the most processing challenges. While archivists have slowly begun the process of normalizing the use of EAD, there are many divergent practices. Further, while many of the names present in the finding aids will be identified as such and will be well or reasonably well-formed, many other names, particularly those found in prose biographies and histories and in the description of correspondence, are in either natural language contexts or in contexts that intermix names with other descriptive data. Adapting the XSLT programs to *each* set of finding aids will present a particular challenge. Further, in order to maximize the recognition and extraction of names, particularly the names of correspondents, NLP/NER programs will be installed, trained, tested, and to the extent that results are acceptably accurate and reliable, integrated into the extraction process. It is anticipated that the bulk of the work in extraction processing will be devoted to EAD-encoded finding aids.

While the overall processing of the data is completed in three steps executed sequentially, the development (refinement, enhancement, and integration of techniques and processes) for each of the three steps takes place simultaneously, and will take place continuously over the course of the two-year project, periodically interrupted to execute all three steps in succession, punctuated with a new release of the public prototype system. It is anticipated that the complete execution of the three steps from beginning to end will take place no less than one time for each quarter of the project.

Status meetings will be held in each quarter, alternating between IATH and CDL or UCB. At the beginning of Year 2, the Advisory Board will meet with the project staff at the University of Virginia to review and discuss the work done to date.

Case studies will be developed early in the first year of the project, and a face-to-face user study with scholars and educations will take place early in the second year of the project. A consultant will perform systematic manual review recall (NER) and precision (Identity Resolution) will take place in the summer of each year of the project.

Timeline

<u>First Quarter: April 1, 2012-September, 30, 2012</u>.

Position will be posted for Extraction/Migration Programmer (DE/TP). IATH currently has an open position, and thus there is no delay anticipated in posting and hiring. Position to be filled by July 1, 2012.

Web site will be updated and announcements made to various listservs and publication venues. Publication and disseminating information about the project will be performed periodically over the course of the project, with particular emphasis on milestone releases of the public prototype system. (Wells, with assistance of Turner, Tingle, Martin, and Larson)

Project dedicated server will be purchased, installed, configured. If administratively possible, this will be completed before the start date of the project. Saxon, NLP/NER, XTF, and additional software as necessary installed and tested. (DM/SA in consultation with Pitti and Martin)

Archival data acquisition will commence and continue through the six-month period. The initial focus of the data acquisition will be on reacquiring data from first phase contributors in order to update the data, and also because processing can commence on it immediately using existing extraction software. Focus will then shift to acquiring one to two million MARC WorldCat records. Given the quantity of WorldCat records, the storage and arrangement will be optimized to support incremental, efficient processing. Following the acquisition of the WorldCat records, NARA and BL original authority records will be acquired, and then agency records from Smithsonian Institution Archives and New York State Archives. Both of these are in the MARC bibliographic format. A small number of original archival authority descriptions in the EAC-CPF format along with a small number of EAD-encoded finding aids will be acquired from the BnF and Archives nationales (France). Last in the order of acquisition will be the remaining EAD-encoded finding aids. The order of acquisition is strategic and is intended to maximize the number of EAC-CPF records as early as possible in the project. It is also intended to ensure that various tasks assigned to the DE/TP continue without interruption. (DM/SA with assistance from Pitti and Wells)

First project status meeting.

XSLT and NLP/NER extraction and data migration development will commence. (Pitti until DE/TP hired; and then DE/TP under supervision of Pitti)

Ongoing data management and system administration task with respect to acquired data and EAC-CPF records produced. Ongoing system administration tasks in service of IATH, SI/UCB, and CDL programmers. (DM/SA with assistance from Pitti, Martin, and Wells)

Design and development of date normalization programs and, geographic NLP/NER, name normalization, and addition of coordinates data to identified geographic names. (TMP and Martin, coordinating with DE/TP for integration into extraction and normalization processing)

Consultant Wisser will meet for one day with Larson, Martin, and Pitti to clearly identify the objectives of manual qualitative evaluation of extraction and match/merge processing, and the methods to be employed in meeting the objectives.

A statistically valid subset of EAC-CPF records with associated source data will be selected (TMP, Martin, in coordination with Larson). Consultant (Wisser) will manually review EAC-CPF records in relation to source data to evaluate the extraction, matching, and merging recall and precision. Various processes will be scored using open source software developed by OCLC Research. Findings will be distributed to enable focused development and refinement of extraction, matching, and merging processing. (TMP and Martin in consultation with Larson)

16 million VIAF authority record clusters will be acquired from OCLC Research and an additional 200,000 ULAN records will be acquired from the Getty Vocabulary Program. Each set of authority records will be indexed in Cheshire to facilitate subsequent match/merge processing. Using EAC-CPF records produced at IATH, EAC-CPF records will be matched against one another, and the resulting set matched against VIAF, and ULAN. Code will be developed to further process the resulting set to selectively access and explore LCNAF records using permalink identifiers acquired from the LCNAF data in the VIAF clusters. This step is to acquire descriptive data in the LCNAF records not otherwise found in the VIAF cluster records. Testing, refining, and enhancing the match/merge processing will be an ongoing activity. (Larson and GSP)

Development of the prototype public access system will continue. (Tingle in consultation with Turner, Hu, and as necessary, entire team). Hu, in consultation with Turner and Tingle and entire team will conduct a user engagements to further develop use cases to assist in designing and refining the features and functionality of the prototype. Research and development of social network extraction (entity to entity and entity to resource data extracted from EAC-CPF records into graph database and also migrated to RDF), search, and display, with incremental releases in public prototype. (Tingle with input from Pitti, Turner, Martin, and Larson)

Second Quarter: October 1, 2012-March 31, 2013.

Archival data acquisition will continue. (DM/SA with assistance from Pitti and Wells)

XSLT and NLP/NER extraction and data migration development will continue. (DE/TP under supervision of Pitti)

Testing, refining, and enhancing the match/merge processing continues. (Larson and GSP)

Development of the prototype public access system continues. (Tingle in consultation with Turner, Hu, and as necessary, entire team)

Second project status meeting.

Design and development of date normalization programs and, geographic NLP/NER, name normalization, and addition of coordinates data to identified geographic names. Design and development of timeline-map display for integration into public prototype. (TMP and Martin, coordinating with DE/TP for integration into extraction and normalization processing, and Tingle for integration of display into public prototype)

Third Quarter: April 1, 2013-September, 30, 2013.

Meeting of Advisory Board.

Any not yet acquired data will be acquired. (DM/SA with assistance from Pitti and Wells)

Face-to-face user studies with scholars and educators (Hu and Turner), with findings to be reported to Tingle and entire team.

Consultant (Wisser) will manually review EAC-CPF records in relation to source data to evaluate the extraction, matching, and merging recall and precision. Various processes will be scored using open source software developed by OCLC Research. Findings will be distributed to enable focused development and refinement of extracting, matching, and merging processing. (TMP and Martin in consultation with Larson)

Third project status meeting.

XSLT and NLP/NER extraction and data migration development will continue. (DE/TP under supervision of Pitti)

Testing, refining, and enhancing the match/merge processing continues. (Larson and GSP)

Development of the prototype public access system continues. (Tingle in consultation with Turner, Hu, and as necessary, entire team)

Design and development of graphical display of organizational charts and timeline display of corporate body name changes using NARA, SIA, and NY State Archives agency history. (TMP and Martin, coordinating with Tingle for integration of display into public prototype)

Fourth Quarter: October 1, 2013-March 31, 2014.

Project code and code documentation will be reviewed by outside reviewers. Reviewer reports will be distributed to Wells and all code developers to ensure that recommendations can be incorporated before the code is released as open source. (CDRCI and CDRCII)

XSLT and NLP/NER extraction and data migration development continues and terminates with a final extraction and assembling of EAC-CPF records. (DE/TP under supervision of Pitti)

Testing, refining, and enhancing the match/merge processing continues and terminates with a final processing of EAC-CPF records. (Larson and GSP)

Development of the prototype public access system continues. (Tingle in consultation with Turner, Hu, and as necessary, entire team)

Design and development of graphical display of organizational charts and timeline display of corporate body name changes is finalized. (TMP and Martin, coordinating with Tingle for integration of display into public prototype)

Final project status meeting.

Expected Outcomes

The project will have four significant outcomes. First, the software used in processing and publishing the data will be made available as open source. Second, research findings will be published and will contribute to the advancement of research in Name Entity Recognition and Identity Resolution, two related and challenging areas critical to the advancement of the Semantic Web and humanities text mining; and to ongoing international development of cultural heritage descriptive standards, in particular International Council on Archives (ICA) and on Society of American Archives (SAA) archival standards. Third, the prototype public system will be useful in itself as a powerful research tool, *and* will serve as a compelling demonstration that archival authority description, when aggregated and interrelated in a single system, can transform both the economy and nature of scholarly historical research by providing integrated access to primary and secondary resources as well as access to the social-historical contexts that document important person, lives, and events. This demonstration will be used to build support in the cultural heritage and scholarly communities for establishing a sustainable national archival authority cooperative. Finally, the methods developed and used in building, and much of the assembled data in SNAC, will serve as a solid data foundation for such a cooperative.

Intellectual property issues

The types of institutions providing data to the SNAC project are a mix of federal and state archives and libraries, academic or research (public and private) archives and libraries, archival consortia, a private museum, and OCLC. All of the academic and research archives and libraries are in the U.S., except the ArchivesHub (U.K.). The BL, BnF, and Archives nationales (France) are also contributing data.

All of the repositories and consortia have formally committed to provide data to the project. In addition, all of the repositories and consortia have given permission to the project to make Internet-accessible, both displayed in browsers and as Linked Open Data, authority records derived in whole or part from source data provided by them for the research and demonstration purposes of the project. Some but not all providers of data have granted permission to the project to transfer the authority records assembled in the project to a national archival authority cooperative program, if such a program is established.

Please see Appendix One for a list of the contributing institutions and Appendix Three for copies of the letters of commitment and permission.

The project will also release all software and programs developed in the project as open source. Given that some of the software and programs are built on or employ as components to open source software not developed in the project, the project software will be released under a variety of open source licenses that respect the licensing of the software and programs utilized.

IATH Developed Software

The extraction software will be based on XSLT 2.0 and XPath 2.0, W3C Recommendations. The programs will be entirely original except for the use of third-party XSLT functions provided by the FunctX XSLT Function Library. FunctX is available under a GNU-LGPL license. The extraction programs will be released under an ECL 2.0 license. Components of the extraction software will address normalization of dates and normalization of geographic names. These components will be written in XSTL, possibly with components of the components written in Java. These normalization components will be designed to be of general use outside of the context of the SNAC project, and will thus be made available as separate components under an ECL 2.0 license.

Software to render timeline-maps of biographies or histories of named entities will be an adaptation of the open source timemap (available under a MIT license). timemap is a Javascript library to help use a SIMILE timeline with online open source mapping software such as OpenLayers and OpenStreetMap. IATH will utilize and may modify third party open-source, software, code' and libraries in conjunction with timemap. When modifying third party code or libraries, the IATH will make its derivative works available under the same terms as the original third party software, code, or libraries. It is anticipated that IATH will use utilize PostgreSQL, Ruby, Ruby On Rails, Javascript, and JSON in developing the SNAC timeline-maps.

NLP/NER software will be based on available open source software. Candidates for use are two *trainable* NLP/NER suites of software: the Stanford Name Entity Recognizer (NER) and Extracting Metadata for Preservation (EMP). Both the Stanford NER and EMP suite are available under a GNU GPL v2 license. Other NLP/NER software may be employed if neither of these candidates proves sufficient in achieving the projects quality objectives. The software and programs developed in SNAC adapting or training the NLP/NER software will be made available in accordance with the underlying GUN GPL v2. licenses.

SI/UCB Developed Software

The role of SI/UCB in this project will be to further develop and enhance the name matching and merging system developed in the initial SNAC project and through the "Connecting Lives - Context Mining for Information on People, Organizations and Families" currently being funded by Mellon at SI/UCB. The system that will be built for the name matching and merging system is primarily written in the Python programming language and includes various open source Python utility libraries for natural language processing and string processing as well as original code. In addition the current version of the system makes use of "MongoDB," under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) License. The current system also uses the Cheshire information retrieval system, under a BSD-style license from the Regents of the University of California, which allows free non-commercial and educational use. The items currently in use (including MongoDB and Cheshire) may change, as other options and requirements are found in the project (e.g., issues of scaling), but any new software adopted for database management or searching will need to offer equivalent licensing for open and free non-commercial use.

CDL Developed Software

The public prototype access system is being developed on the Extensible Text Framework (XTF), an open-source platform for publishing W3C XML-encoded documents and data developed by the CDL. Commissioned by the CDL to be the primary access tool for its collections, XTF provides a powerful, flexible platform for providing access to digital content. It consists of Java and XSLT 2.0 code that indexes, queries, and displays W3C XML-encoded documents and data. It incorporates other open-source software -- in particular Saxon, an XSLT processor, and Lucene, a text search engine developed by the Apache Project -- and is itself freely available for developers to download, install and configure their own instance of the software. CDL will utilize standard UNIX-based programming and editing applications to configure XTF.

CDL customization of XTF and display of records in the access system will primarily be done using W3C eXtensible Stylesheet Language-Transformation (XSLT), Java, HTML, and CSS.

XTF is developed under a Berkeley Software Distribution (BSD) license. For the SNAC prototype system, CDL will utilize and may modify third party open-source code and libraries in conjunction with XTF. When modifying third party code or libraries, the CDL will make its derivative works available under the same terms as the original third party code or libraries.

The social network graph data retrieval and rendering programs are based on additional open source software: SPARQL, TinkerPop (Stack and Neo4J) and the JavaScript InfoVis Toolkit (JIT). For SNAC social network and display programs, CDL will modify third party code or libraries, and make the derivative work available under the same terms as the original third party code or libraries.

Long-term sustainability

While the immediate goal of the project is not to build permanent resources and tools, the tools developed in the project, the archival authority descriptions, and the prototype public system will be maintained by IATH until a sustainable national archival authority cooperative can be established. In this regard, in parallel with SNAC, IATH is leading community-based development of a blueprint to address the business, governance, and technological requirements for establishing a national archival authorities cooperative. The Institute for Museum and Library Services (IMLS) is funding the development of the blueprint and the National Archives and Records Administration will host three meetings related to the effort.[12] See Appendix Six for a copy of the abstract for the Building a National Archival Authorities Infrastructure.

Reporting and Evaluation

The project will submit a report for each of the two years of the project. These reports will provide detailed description and evaluation the project activities, as well as a detailed financial report that will include documentation on grant expenditures. Assuming an April 1, 2011 start date, the first report will be submitted no later than June 30, 2013. The second and final report will be submitted no later than June 30, 2014.

The project activity and evaluation reports will be organized around the three primary activity areas:

- Name entity recognition (NER) and extraction; and assembling from extracted data or deriving through transformation EAC-CPF-encoded archival authority descriptions. (IATH)
- Authority description matching (Identity Resolution); and merging or combining archival authority descriptions. (SI/UCB)
- Developing prototype public historical resources and resource access system. (CDL)

[12] http://www.iath.virginia.edu/news/news_naac_announcement.html

An outside consultant will evaluate the first two of the three activities. The methods employed by the consultant will be based on methods developed by OCLC Research in its building of VIAF and WorldCat Identities, as well as other research activities. OCLC Research has developed NER and Identity Resolution scoring software that can be used in conjunction with the manual review to provide scoring with respect to the recall and precision of the NER and Identity Resolution. The manual review will be given a statistically valid sample of EAC-CPF records with accompanying source data, and will examine each step of the processing with respect to actual results and correct results.

Two outside reviewers will evaluate the code and documentation developed by the project in the last quarter of the second year of the project. An expert in XSLT (Catapano) will evaluate XSLT code and documentation developed to extract data from archival descriptions and assemble them into EAC-CPF records. An expert in software design and development (Sprenkle) will examine all other code developed in the project. The reviewers will provide reports to Pitti, and summaries of the reports will be provided in the final report of the project.

The project will also provide a detailed quantitative description of the number of finding aids and original archival authority records processed, the number of EAC-CPF records derived or assembled, the number of EAC-CPF records in the final set after identity resolution and merging of matching records.

The prototype public access system will be evaluated through face-to-face scholar and educator user studies. These studies will provide insight into how best to design the public interface, but also details on what features are effective (or not).

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

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