

## Social Networks and Archival Context Project: A Case Study of Emerging Cyberinfrastructure

Tom J. Lynch <tlynch6\_at\_csc\_dot\_com>, Computer Sciences Corporation

### Abstract

For digital humanists planning to build tools for cyberinfrastructure several variables ought to be defined for each project. Pay close attention to the balance of traditional methods and new ways of conducting research. When gathering resources to do the job, seek contributions of different domain experts. Also, careful consideration of a tool's intended scope will help refine the required resources needed to complete a project. This case study illustrates how one project, the Social Networks and Archival Context Project (SNAC), has defined these variables. The process of building a new tool also benefits from an awareness of older infrastructure that has come before it. SNAC illustrates this awareness in the way it has taken advantage of previously existing infrastructure, both cyber and not, by extending its purpose and building new features on top of it.

When speculating about the future of scholarship, it seems certain that research and teaching will continue to be affected by the evolution of traditional infrastructure into digital forms. This is as true for the humanities as it is for other disciplines and digital humanists in particular should have an important role to play in shaping the infrastructure for their domains. Digital humanists, who can both identify the needs of mainstream humanities scholars and suggest acceptable computational solutions to those needs [Juola 2008], are essential participants in the development of a new digital cyberinfrastructure for humanities research. 1

The American Council of Learned Societies (ACLS) report on cyberinfrastructure defines cyberinfrastructure as 2

the layer of information, expertise, standards, policies, tools, and services that are shared broadly across communities of inquiry but developed for specific scholarly purposes: cyberinfrastructure is something more specific than the network itself, but it is something more general than a tool or a resource developed for a particular project, a range of projects, or, even more broadly, for a particular discipline. [ACLS 2006]

This definition leaves a lot of room to imagine the technological future of the humanities, allowing researchers and educators to decide for themselves the appropriate tools and services that will meet their needs. While the report states that cyberinfrastructure is something more general than a tool or resource, tools will be a necessary component. For digital humanists planning to build tools for cyberinfrastructure, it is helpful to pay close attention to the balance of traditional methods and new ways of conducting research, and to the contributions of different domain experts, when gathering resources to do the job. Careful consideration of a tool's intended scope will also help refine the required resources needed to complete a project. The following case study will illustrate how one project has successfully balanced these variables and it may serve as a model for designing others. I will discuss each of these variables more, but I also want to introduce the idea of an infrastructure stack on top of which a successful cyberinfrastructure tool rests. The stack serves as a foundation for the new technology by linking it to the past. The sense of the word "stack" I use is the one that refers to a pile of objects, one on top of the other. In this case the stacked objects are abstract systems and technologies. The oldest technology rests at the bottom and newer technologies pile up to the top. If you remove one from the stack, the objects above it and supported by it could not exist without the missing foundation. To illustrate this

concept and elaborate on the balancing act of tool design, I offer a case study of The Social Networks and Archival Context (SNAC) Project, which is an emerging example of cyberinfrastructure. SNAC serves this purpose by demonstrating the ability to balance many of the variables necessary for cyberinfrastructure and the project is built upon an easily definable infrastructure stack. I will begin this case study with a description of SNAC and its prototype user interface.

## Introducing SNAC

The SNAC project aims to provide scholars with improved access to distributed historical records with new discovery tools. The heart of the project is a very large (and still growing) dataset of archival creator descriptions expressed in the international metadata standard Encoded Archival Context-Corporate Bodies, Persons, and Families (EAC-CPF). The records describe people and corporate bodies who are the creators of primary humanities resources that have been collected into archives and libraries around the world. The descriptions contain biographical data, lists of associated resources, and lists of associations with other people. This curated set of data enables SNAC to build new discovery tools for researchers. SNAC's goal is to provide improved access to the resources that document the lives, work, and events surrounding historical persons, and provide unprecedented access to the biographical-historical contexts of the people documented in the resources, including the social-professional networks within which they lived and worked [IATH n.d. b].

3

SNAC's emphasis on social networks is, of course, no accident. Social networks are a hot topic. However, social networks have been around far longer than Facebook. Simply put, a network is a set of relations between objects; a social network the set of relations between people. Kadushin describes social networks as having been

4

at the core of human society since we were hunters and gatherers. People were tied together through their relations with one another and their dependence on one another...Kinship and family relations are social networks. Neighbors, villages, and cities are crisscrossed with networks of obligations and relationships. [Kadushin 2012]

What is relatively new, Kadushin continues, are "systematic ways of talking about social networks, depicting them, analyzing them, and showing how they are related to more formal social arrangements such as organizations and governments." [Kadushin 2012]

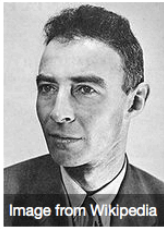
From Facebook to Flickr, Twitter to Tumblr, the social networking concept is embodied everywhere online through many widely used Web applications. Taking advantage of the relationships captured in its dataset, SNAC has developed a social network web application that has helped the project earn the nickname "Facebook for the dead" [Seal 2012]. The comparison with the ubiquitous social networking site is not without some utility. Social networking sites like Facebook enable users to articulate and make visible their social networks [Boyd and Ellison 2008], which is the organizing principle for SNAC's data. Boyd and Ellison define social network sites as web-based services that allow individuals to construct a profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system. In its own way, SNAC provides these same features for historical figures.

5

**Oppenheimer, J. Robert, 1904-1967** [Alternative names](#)

**Dates:** birth 1904-04-22 death 1967-02-18  
**Authority Source:** VIAF, LC, WorldCat  
**Nationality:** United States  
**Language:** English  
**Gender:**

**Biographical notes:**



Physicist (quantum theory and nuclear physics). On the physics faculty at California Institute of Technology and University of California, Berkeley in theoretical physics, 1929-1947; director of Los Alamos Scientific Laboratory, 1943-1945; chairman of the General Advisory Committee of the Atomic Energy Commission, 1946-1952; director of the Institute for Advanced Study at Princeton, 1947-1966.  
 Died 1967

**Links to collections**

- Archival Collections 201
- Related Resources 100
- Related External Links 6

**Related names in SNAC**

- People 175
- Families 0
- Organizations 35

**Visualize:**

- List collection locati
- View source EAC-CP

**Subjects:**

- Atomic bomb
- Atomic bomb--M aspects
- Atomic theory
- Electrodynamics
- Electrons
- Mesons

Figure 1. SNAC profile for J. Robert Oppenheimer

Figure 1 shows an example of a personal profile in the online SNAC prototype (<http://socialarchive.iath.virginia.edu/snac/search>), in this case the one for J. Robert Oppenheimer. The profile contains descriptive metadata such as life dates, occupations, subjects to which the person is related, alternative forms of his or her name, and a biographical history. The central column contains “Links to collections” and “Related names in SNAC,” where lists of connections from the personal profile to archival collections, other people, corporate bodies, and resources are articulated as well as providing links to traverse the connections. For example, Oppenheimer is listed as the creator of 38 collections housed at the Library of Congress and several university libraries. Additionally, SNAC lists another 163 archival collections in which Oppenheimer is referenced. For each of these collections SNAC provides a link to its entry in WorldCat, the online catalog of the world’s largest network of library content and services [WorldCat n.d.]. In addition to these external links, SNAC also links internally to the profiles of other people, families, and organizations that share a connection to a given profile. Oppenheimer, for example, is listed as having been associated with or corresponded with 175 other people and 35 organizations, all of which have a profile in SNAC.

The profile pages, the links between them, and the lists of external related resources are not created manually by experts. Another characteristic SNAC shares with social networking sites is the bottom-up, rather than top-down, approach to organizing its content. Individuals in the online community define the content and structure of a typical social network site instead of professional information providers [Kolbitsch and Maurer 2006]. SNAC allows the properties of its data (such as the names, dates, correspondence relationships, and creator attributions captured in the EAC-CPF records) to self-organize the information presented in the system.

## Project Variables

Using SNAC as an example this case study will show what a software tool may look like within humanities cyberinfrastructure and describe its relationship with the other components listed in the definition of cyberinfrastructure from the ACLS report quoted above (information, expertise, standards, etc.). Alongside discussing SNAC specifically, I will address cyberinfrastructure more generally by defining a set of variables to consider when approaching the design of a new tool. When reviewing the existing literature about humanities cyberinfrastructure a picture emerges that helps define these variables. Imagine a three-dimensional graph where examples of cyberinfrastructure technology could be plotted. The three axes of the graph represent the following opposing extremes:

- X: Created by humanists ← → Created by non-humanists
- Y: Small boutique projects ← → General-purpose applications
- Z: Traditional humanities infrastructure ← → Scientific cyberinfrastructure

I will discuss each of these axes in turn through highlighting ideas from the literature, as well as show how SNAC is a real-world example of how one project has found its place on the graph by striking an appropriate balance for its context

6

7

8

9

on each axis. Plotting a multitude of projects on the axes imagined here is beyond the scope of this article, but I do believe SNAC falls close to the middle of the graph, hence its use as an illustrative example.

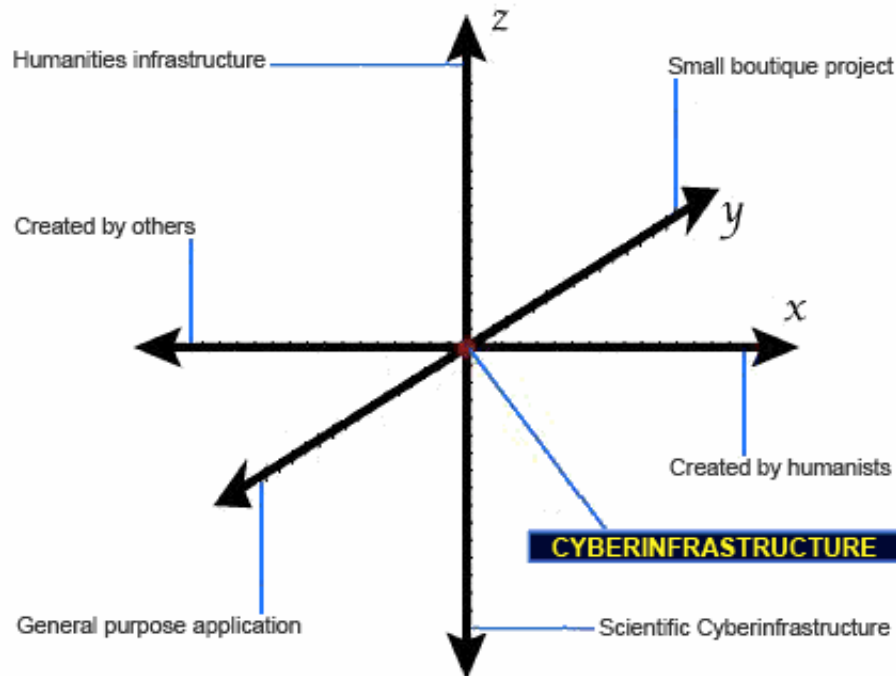


Figure 2. A trio of variables that affect humanities cyberinfrastructure

## X: Created by humanists ← → Created by non-humanists

Due to its complexity and necessarily broad range of required skills, the development of cyberinfrastructure often requires collaboration between humanists and non-humanists. First and foremost, scholars of the humanities must lead the development of their own cyberinfrastructure, despite the fact that tool development is not yet rewarded on par with more traditional scholarly outputs [Schreibman and Hanlon 2010]. They alone can accomplish the intellectual task of modeling scholarship for new digital environments [Drucker 2009]. They alone can judge its usefulness and assure that the technology will serve their needs [Hunt et al. 2011]. However, few individual humanities experts (let alone any expert) will have the range of expertise required to build cyberinfrastructure, so librarians, archivists, programmers, and computer scientists will bring essential complementary skills to the task [Borgman 2009]. Technology developers and humanities researchers working alongside each other are likely to push both the technical and the intellectual boundaries farther, yielding greater innovation [Collins et al. 2012]. Casting the net even wider, the ACLS report, which lists collaboration as a necessary characteristic of cyberinfrastructure, calls for “cooperation with librarians, curators, and archivists; the involvement of experts in the sciences, law, business, and entertainment; and active participation from and endorsement by the general public” [ACLS 2006]. However, it should be noted that, while necessary, collaboration is itself a challenge that requires careful resolution of methodological differences and regular communication about each collaborators’ perspective [Chuk et al. 2012].

In the case of SNAC, its emphasis on social networks is conducive to attracting collaborative partners. The study of social networks has been identified as a cross-cutting agenda where the digital scholarship of humanists, technology researchers, and others converge in a way that encourages collaboration and, in particular, brings humanists to a seat at the table where scholarly infrastructure decisions are being made [Borgman 2009]. The social network is a topic that lends itself to interesting collaborations among many different disciplines with a scope not so narrow as to constrain research nor so expansive as to be meaningless [Friedlander 2009].

The Institute for Advanced Technology in the Humanities (IATH), University of Virginia, leads the team developing

SNAC. IATH is known for collaborative research and technology development with the practical needs of humanities scholars as its primary motivation. The list of other participating institutions and individuals on SNAC's project team and advisory board suggests a wide range of disciplinary expertise is contributing to the project, including historians, English scholars, librarians, archivists, information technology developers, information scientists, computer scientists, and more [IATH n.d. c]. The breadth of knowledge required by the development of SNAC means no single discipline has all the answers to the project's research questions. Archivists, librarians, and information scientists are needed for their expertise in existing infrastructure surrounding primary resources, experience in information organization, and their knowledge of new data standards. Computer scientists and developers contribute by solving and implementing solutions for the myriad of technical challenges encountered during the course of SNAC's development. Historians and English scholars contribute their knowledge of humanities research methods and SNAC's user community, helping shape the public interface to the data and ensuring the new tools are designed to be of practical value to scholars. For the sake of brevity, these examples are probably oversimplified. The full picture of collaboration on SNAC is more complicated, especially since its participants often wear multiple hats and it is hard to encapsulate all the multidisciplinary challenges faced by SNAC in a few short sentences. However, the team does appear to have a mix of expertise well suited to the task of developing SNAC.

## **Y: Small boutique projects ← → General-purpose applications**

Given the great need for teams of collaborators, it should be evident that the scale of cyberinfrastructure research is large, the goals are many and multi-faceted, and the aim is to serve a large audience. However, attempting to do too much, to make a technology useful for everyone in the humanities, risks the possibility of making it useful for nobody, because a tool with a very broad audience in mind is unlikely to meet the requirements of specialized fields [Hunt et al. 2011]. At the other end of the scale, cyberinfrastructure of the past decade has led humanities researchers to create smaller, isolated projects, the contents of which cannot be easily combined for analysis without the development of a more robust infrastructure to link them together [Blackwell and Crane 2009] [Crane et al. 2009]. Creators of a cyberinfrastructure tool should have a specific audience in mind and build features into a tool that serves that audience well, but the tool should also be interoperable with other systems to support some unimagined uses by unanticipated users. A successful cyberinfrastructure of the future will likely consist of multiple systems that exist to support the specialized needs of each discipline, yet support the sharing of data between these systems [Hunt et al. 2011].

13

SNAC's focus on social networks and creator description limits its scope by reducing the types of data to be managed and the kinds of functionality designers might build into the tool for analysis. However, it does not restrict the size of the project. SNAC is poised to expand widely within a certain band of information, specifically the description of primary resource creators and the relations between them. SNAC has compiled over 2.6 million EAC-CPF records describing persons, families, and corporate bodies [IATH 2014] since the project began. In future phases of development, nearly 4 million EAC-CPF records may be generated, creating an unprecedented collection of data about the lives, work, and events surrounding historic persons [IATH n.d. b].

14

With its narrow, but deep, focus on creator description, SNAC intends to do a handful of tasks extremely well and integrate itself into existing infrastructure to offer researchers more functionality. The challenges involved in creating, managing, and providing access to this massive amount of data are a big enough task for one project, so SNAC will rely on others to extend its functionality. For example, SNAC does not intend to collect digitized primary resources or provide direct access to such resources, but rather prefers to leverage existing cyberinfrastructure by providing links to WorldCat (<http://www.worldcat.org/>) and the online finding aids of archival institutions to bring researchers closer to the resources they seek. Also, SNAC has a specific audience in mind that is large, but certainly not all-inclusive: anyone interested in historical research. SNAC will serve this audience in a few ways: integrated access to distributed primary and secondary resources about people and organizations; access to historical and biographical descriptions; and access to the social networks of people of historical interest [IATH n.d. b]. The emphasis on a small number of functions for a targeted audience built atop a unique dataset of unparalleled size strikes an appropriate balance on the Y-axis between boutique projects and general-purpose applications.

15

## **Z: Traditional humanities infrastructure ← → Scientific cyberinfrastructure**

Humanities cyberinfrastructure should be a step forward for research infrastructure otherwise there is no reason to build it. It should offer new solutions to old problems or enable humanists to ask new questions. Therefore, to build such systems there is a need to utilize methods and technologies that go beyond the traditional tools of the humanities. Traditional humanities infrastructure includes intellectual categories such as literary genres and linguistic phenomena, material artifacts like books and maps, buildings such as libraries and book stores, organizations including universities and journals, business models such as subscriptions and memberships, and social practices such as publication and peer review [Crane et al. 2009]. In the digital era, however, existing infrastructure must not be used uncritically as the default model for new infrastructure [Svensson 2011]. New infrastructure must go beyond the capabilities of what already exists, otherwise it is merely reinventing the wheel. Humanists may have something to learn from scientists and engineers who have already defined, designed, and deployed much more in the way of cyberinfrastructure, but this would leave the humanities in the position of building on the technologies constructed for other disciplines [Borgman 2009] and this approach should also not be uncritically adopted [Svensson 2011]. A balance must be struck with new technologies that push the boundaries of scholarly activities, yet remain accessible and meet real needs. Humanities scholars will be unlikely to engage with new technology for its own sake [Collins and Jubb 2012] and their adoption of any new technologies will be highly influenced by existing research practices developed when using traditional infrastructure [Collins et al. 2012].

SNAC pushes the boundaries of traditional infrastructure through its use of digital data. The creator description data that drives the project is a standards-based, internally consistent dataset with high potential for interoperability with other systems and purposes. Data of this sort (and the tools that can use it) provide a desirable characteristic of cyberinfrastructure: that access to the data should be seamless across repositories [ACLS 2006]. Having the potential for data sharing and re-use baked into the initial design of the project is an approach more readily found in scientific research. Also, the rigor with which the data is curated opens the door for innovative approaches to research and the tools that support it. This sort of data-centric approach enables structures and relationships to emerge in the system that would otherwise go unnoticed without a lucky combination of serendipity and painstaking manual research by humans. For example, during the first phase of research and development, the SNAC prototype included an experimental radial graph feature that demonstrated the ability to visualize and explore the relationships between people and organizations represented in the data. See Figure 3.

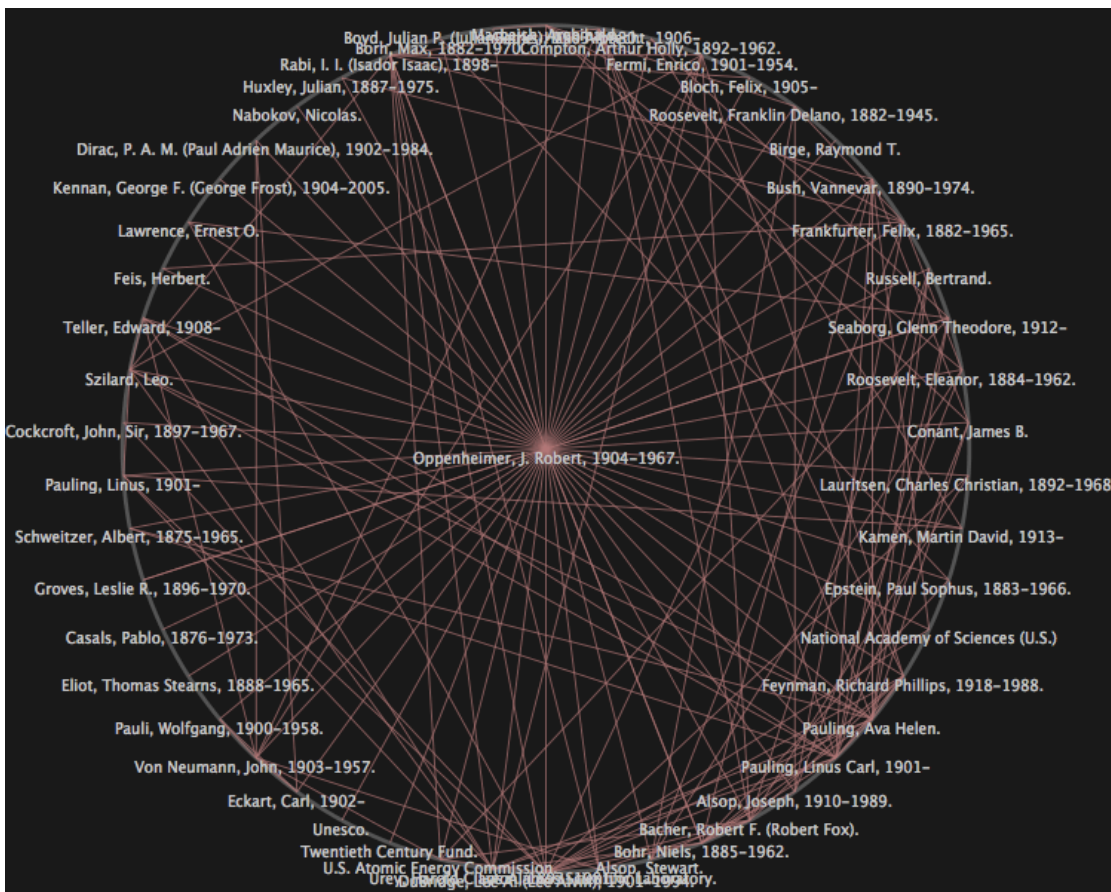


Figure 3. Radial graph demo for J. Robert Oppenheimer

The radial graph visualization initially puts the name of an individual at the center of a circle of nodes (other names) to which he or she is related in some way. The example in Figure 3 shows “Oppenheimer, J. Robert” as the central node with lines extending out to the names of all the other people and organizations he had been associated with in his lifetime. In addition to the lines drawn between the central node in the network and all the other surrounding nodes, relationships between the surrounding nodes in the circle are also indicated with a line. The visualization enables further exploration by responding to mouse clicks on any node in the graph by redrawing the visualization to include the immediate social network of the selected node. For an amusing example, from the graph in Figure 3, clicking on the node for “Eliot, Thomas Stearns” opens his social network and reveals that Oppenheimer, the father of the atomic bomb, was only two degrees of separation from Groucho Marx<sup>[1]</sup> (see Figure 4).



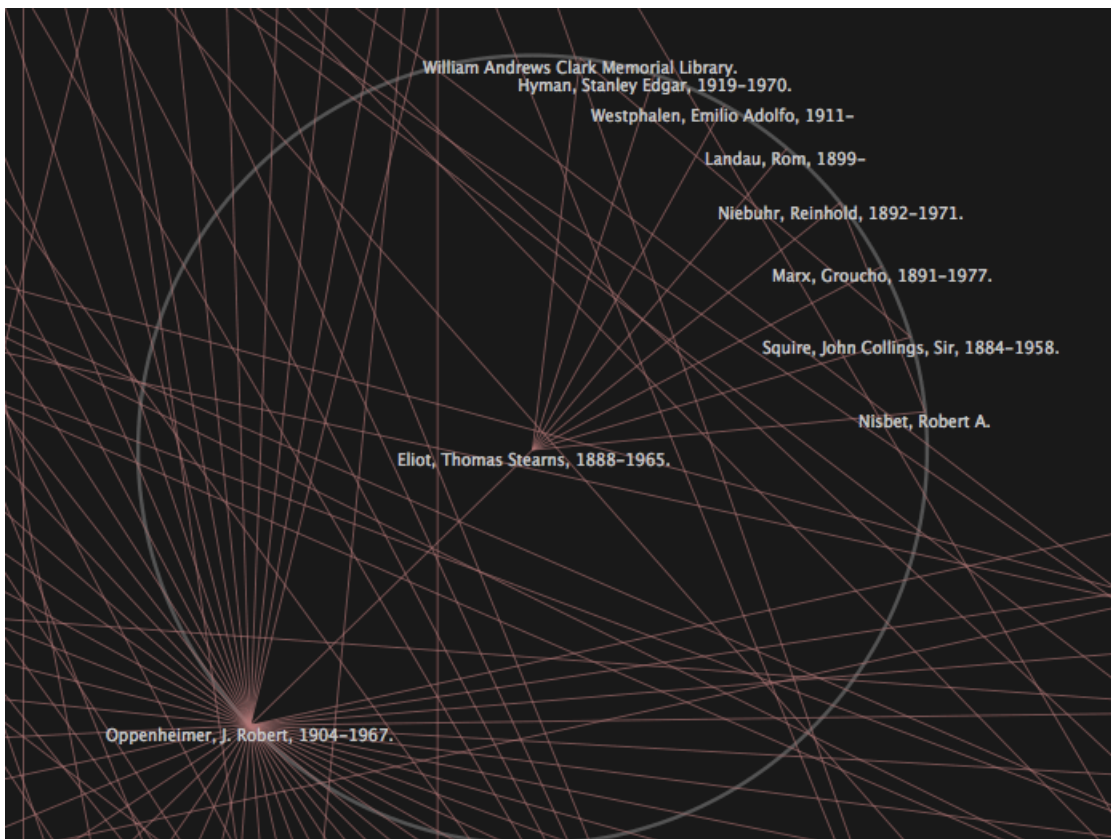


Figure 4. Extended network for J. Robert Oppenheimer via T. S. Eliot

While the current phase of SNAC development does not feature the radial graph, the feature is archived at another location (<http://socialarchive.iath.virginia.edu/xtf/view?mode=RGraph&docId=oppenheimer-j-robert-1904-1967-cr.xml>). It stands as an early experiment in what is possible with SNAC's data, providing a hint of what is to come in future features such as a graph visualization of social-document networks, visualizations of chronologies, and geographic displays of collection locations [IATH n.d. f]. SNAC's web prototype and its visualization tools like the radial graph are the public interface to (and a product of) a variety of innovative, yet practical research agendas tackled by the SNAC project. The research never sought to build and explore technologies just for the sake of creating new technologies, nor did it adopt any existing technology without a plan to apply its use to the current needs of SNAC's target user group. SNAC balances the utility of traditional methods that existing infrastructure enable with the exciting, but unproven potential of new technologies. It does so through the use of advanced technology to accelerate and extend traditional methods in new ways. While researchers may have always wanted to search for every occurrence of an individual's name in the archives, it has previously been impossible due to a lack of funding and time [Collins and Jubb 2012]. However, as some historians have already observed, a year's worth of work using the current infrastructure combined with a great deal of serendipity and persistence can be replaced by the data in a single SNAC record [IATH n.d. b]. Through this dramatic reduction in time required to conduct research SNAC demonstrates the benefits of a well-designed cyberinfrastructure tool.

19

## Humanities Infrastructure Stack

To build upon existing infrastructure an understanding of its layered history and the relationships between its various components may be beneficial. For example, when trying to introduce a new public utility or a new mode of transportation to a community, thorough knowledge of the existing layers of infrastructure may reward planners with insights such as knowing the gaps in existing service that a new one may fill, discovering ways to save development effort for a new project by integrating with current infrastructure, and understanding usage patterns that may help predict the best places to initially roll out a new service. Look at the streetscape of a typical city and you'll see the layers, old and new. Sewers, sidewalks, bike lanes, buses, and telephone poles line the streets. The poles are slung with wires that

20



carry electricity, telephone, cable TV, and the information superhighway. The infrastructure that surrounds us is the result of an iterative process of change and evolution over time. Humanities cyberinfrastructure will be no different.

SNAC is not possible without the many layers of humanities infrastructure that have come before it. SNAC is built on top of a long-standing infrastructure stack that has served the scholarly community in a reciprocal relationship, where it has been shaped and honed by scholarly use and in turn has also shaped how scholars do their research. Any attempt to build cyberinfrastructure ought to include a process of looking back while moving forward. Which facets of prior infrastructure continue to be important today? How can new technology be used to augment and extend those facets to create a new layer of the infrastructure stack? Of course, some may ask, is it even worth building? Certainly there are some who will have doubts about the value of cyberinfrastructure and this points to the importance of the Z-axis of the evaluation criteria. Cyberinfrastructure will not emerge unless it is designed in a way that balances traditional research methods along with the pursuit of advanced technology. One way to design technology that respects the traditions of the past is to look back at the infrastructure stack upon which the new technology is being built and not neglect the scholarly activity it supports. If one can build a technology that eliminates artificial or unnecessary restrictions on scholarly activity, freeing scholars to do what they really want to do — read, write, analyze, produce knowledge, and distribute it — then that technology will be successful [Juola 2008]. What follows is a discussion of several key pieces of the stack beneath SNAC.

21

## Names

Arguably one of the earliest forms of infrastructure supporting information for and about humans is the practice of naming people and their organizations. The abstract appellation is a shorthand way of referencing a whole person in narratives, news, rumors, documents, and data systems like SNAC. Look back at Figures 3 and 4 in this article and notice the primacy of names in SNAC's radial graph. Figure 5 shows a portion of the SNAC prototype's home page where a searchable list of names provides access to data contained in SNAC. The history of naming is long, but is perhaps most significantly related to SNAC in the way personal names reflect the growing complexity of society. For example, the legal and judicial systems of expanding towns in medieval Europe needed to identify individuals clearly in order to attach them to property, to tax and recruit them, and to prosecute them, eventually leading to compulsory naming in the modern era for these reasons [Wilson 1998]. All human organizations now have some bureaucratic interest in names and the official records of many institutions, from churches to governments to corporations and beyond, can end up residing in the next level of the infrastructure stack, the archive.

22

Search by:

All



Person



Family



Organization

enter a name or keywords

Browse or filter by:

Showing 1,568,481 results in person

Featured

Name

Occupation

Subject

Record types

Results

O			O 'Callaghan, Cornelius, 1st Viscount Lismore
A			O : Hara, Joseph, Curate of Vange, county Essex
B			O Briain, Donnchadh Cairbreach, active 17th century
C			Ó Broin, Pádraig,
D			Ó Broin, Pádraig, 1908-
E			Ó Broilcháin, Tomás.
F			O Caiside, Niall, active 17th century, Irish poet
G			O Caiside, Pilip, active 17th century
H			O Caiside, Proinnsias óg, active 17th century
I			O Caiside, Proinnsias, active 17th century
J			O Caoimh, Eoghan, active 1807-1808, Reverend; poet
K			
L			
M			
N			
O			
P			
Q			
R			

Figure 5. Names are a primary access point in SNAC

## Archives

Archives are institutions that collect the unique records of corporations and the papers of individuals and families, the unselfconscious by-products of corporate bodies conducting business and people living their lives [Pitti 1999]. Archives are part of our culture's memory infrastructure, collecting material of legal and historical value. "Any record of human experience can be a data source to a humanities scholar" [Borgman 2009]. Historians list just about every conceivable type of material as important to their work, including primary source material such as reports, wills, deeds, diaries, letters, and so on, all materials to be found in the archives [Case 1991]. What tool can a scholar use to find the needles in the haystacks of archives? That would be the next level of the infrastructure stack.

23

## Finding Aids

A finding aid is a printed description of all the records left in an archive with a common creator or source. A finding aid contains a description of the creator, functions performed by the creator, and the records generated by the creator through the performance of those functions [Pitti 2006]. The finding aids provide context to archival collections and detail their contents [Daines and Nimer 2012]. Reading through finding aids has been listed by some historians as the most frequently used method for locating primary source material [Dalton and Charnigo 2004]. The archival descriptive principle that organizes material based on its provenance (who created the collection) causes personal and organizational names to be a key access point for finding information in archives. Of course, this creator attribute of a finding aid is not the only name. Deeper sections of the finding aid describing the contents of a collection will also contain a wealth of names. Experienced researchers often know the names of key people or organizations connected to their topics of interest and pursue source material linked to those names to provide context [Duff and Johnson 2002]. Reading finding aids and collecting names found therein is a method for building up a list of leads to new sources. Finding aids have been an integral part of humanities infrastructure, but for a long time the descriptive practice did not conform to expectations of consistency one might expect from such a system. Archivists initially developed specialized

24

local practices for describing their collections, believing their collections to be too unique to employ standardized practices. However, over time, archivists came to realize that although the contents of their collections are unique, the properties, characteristics, and needs of these collections are not [Theimer 2011]. This realization, combined with the energy behind automating digital forms of description, led to the adoption of the structured standard embodied in the next level of the infrastructure stack.

## Encoded Archival Description

The Encoded Archival Description (EAD) Document Type Definition (DTD) is a standard for encoding archival finding aids using Extensible Markup Language (XML) [Library of Congress 2011 a]. Development of the data standard began in 1993, currently exists in a version released in 2002, and has been adopted by archives to enhance access to their collections by displaying the finding aids online and enabling full text indexing for easier keyword searching [Pitti 1999]. The schema is intended to represent traditional print finding aids in machine-readable form and contains elements to hold the same information one would find in those finding aids, such as the creators, their functions, biographical information, and descriptions of individual records in a creator's collection [Pitti 2006]. While EAD transformed archival description and brought it into the digital age, the print conventions retained in the presentations of online finding aids has been criticized as inappropriate for the Web and archivists were urged to explore different ways of presenting finding aid content online [Daines and Nimer 2012]. Archivists themselves increasingly recognize that the structure of the traditional finding aid, which is embodied in EAD, is inflexible and inefficient when dealing with complex, interrelated records [Pitti 2004]. The next level up in the infrastructure stack, the aforementioned EAC-CPF, provides the key to revolutionizing archival description and online access. However, before I address that, there's one more part of the stack that bears mentioning.

25

## Authorities

Before reaching the top of this mostly archival infrastructure stack a side step to library infrastructure is necessary. An authority record is "a tool used by librarians to establish forms of names (for persons, places, meetings, and organizations), titles, and subjects used on bibliographic records" [Library of Congress 2011 b]. Of particular interest to SNAC are the collection of authorized names (such as "Twain, Mark, 1835-1910") and the list of cross-references that lead to the authorized names (such as "Snodgrass, Quintus Curtius, 1835-1910" and "Clemens, Samuel Langhorne, 1835-1910"). For reasons that will be described shortly, SNAC relies on authority files to flesh out the biographical descriptions of people and organizations in its data.

26

## Encoded Archival Content – Corporate bodies, Persons, and Families

Encoded Archival Content – Corporate bodies, Persons, and Families (EAC-CPF) is the top of the infrastructure stack that supports SNAC. EAC-CPF Schema is a standard for encoding contextual information about persons, corporate bodies, and families related to archival materials using XML [Staatsbibliothek zu Berlin, n.d.]. EAC-CPF solves a problem inherent in traditional finding aids and EAD, that is the difficulty of representing complex records with a single document. Archival records are often of mixed provenance or the records of the same provenance can be dispersed over numerous archives. This situation causes unnecessary redundancy and duplication of effort [Pitti 2004]. EAC-CPF solves this problem by enabling the separation of creator description from record description. Maintaining a unique, centralized creator record not only reduces redundancy and duplication of effort, but also provides an efficient means of linking creators to the functions and activities carried out by them, to the dispersed records they created or to which they are related, and to other related creators [Pitti 2006].

27

As I mentioned before, one of SNAC's primary objectives is the creation of a large (and growing) collection of EAC-CPF records that power its prototype access system. As its original National Endowment for the Humanities (NEH) proposal describes, SNAC takes three steps to generate creator descriptions in EAC-CPF [IATH n.d. a]. First, preliminary EAC-CPF records are derived from name and biographical components of EAD. The structured nature of the EAD records enables information to be migrated directly into the EAC-CPF format through automated EXtensible Style Sheet Transformations (XSLT). A less simple part of this first step is the extraction of names from semi-structured and free

28

form text fields that require natural language processing techniques.

In the second step, duplicate EAC-CPF records that describe the same person, family, or corporate body are automatically matched and merged into a single record. This process presents a challenge for SNAC. It's not unusual for people and corporate bodies to use different names and leave records behind under those various names. Additionally, the EAD records, while structured, often contain text data created manually by humans. Consequently, names may contain misspellings or the same name may be entered in a variety of ways, such as "Twain, Mark, 1835-1910," "Twain, Mark," "Mark Twain," or "M. Twain." For more information on this step and further technical details about the entire SNAC project, see Larson and Janakiraman [Larson and Janakiraman 2011].

29

The third step involves matching the set of EAC-CPF records against authority records from the Library of Congress Name Authority File (LCNAF) and the Getty Vocabulary Program Union List of Artists' Names (ULAN). Doing so will enable the project to set the authoritative form of a name as the primary name in the EAC-CPF record as well as incorporating a list of alternative forms. Also, biographical/historical data found in the ULAN files will be added to the data in the EAC-CPF record to form a more complete description of the person, family, or corporate body.

30

As of August 2014, SNAC has derived more than 2.6 million EAC-CPF records by extracting data from 2.2 million WorldCat archival descriptions and nearly 300,000 British Library authority records. Future phases of development expect to derive data from nearly 190,000 more descriptions of historical collections from other government and academic archives and libraries [IATH 2014]. SNAC's prototype access system and the radial graph visualization tool are the current examples, but more may be developed. Perhaps, in time, such tools will come to be considered as essential to humanities research infrastructure as the earlier layers of the stack that support them.

31

## Conclusion

This case study has discussed the development of cyberinfrastructure at the level of building tools to enhance research. I used SNAC as an example because it allowed me to illustrate several variables that ought to be considered when designing a research tool for humanities cyberinfrastructure: size and scope of the project, members of the project's team, and choice of technologies. SNAC aims to be a project of great size focused on a narrow mission — creating a network of archival data. To meet the specific research needs of its anticipated community of users SNAC will leverage the power of its data to provide access to research materials through integration with existing infrastructure. This goal could not be achieved without a diverse group of collaborators, from technology specialists to humanists, each representing the various domains of expertise that contribute to the project. This variety of influence on SNAC's design contributes to decision-making that weighs the merits of traditional infrastructure and advanced cyberinfrastructure. SNAC adopts and develops technologies that best meet the needs of humanities researchers while pushing the limits of what is possible. One key way designers of SNAC ensure that the needs of current scholars are met is to heed the heritage of the project's infrastructure stack. The creators of SNAC have understood the infrastructure of the past and present, identified what has served users well and which aspects need improvement, and thought creatively about how to augment existing services and solve current problems. Further examples of new cyberinfrastructure are likely to evolve naturally from the layers of existing infrastructure in this way as opportunities to build new layers are realized in the future, just as the infrastructure layers of the past were once new solutions to old research challenges. Research infrastructure both reflects and inspires humanities thinking and methods. As old obstacles to research crumble through the development of new infrastructure, new methods evolve and generate new challenges to bedevil researchers. The developers of infrastructure respond with technology and ingenuity, creating new services to meet these challenges. The cycle continues to this day, where digital technologies are inspiring new methods of research and new tools are being designed to facilitate, enhance, and accelerate those methods while eventually setting the stage for the next evolution in humanities thinking.

32

While I have discussed mainly tool building, cyberinfrastructure is much more than that. Cyberinfrastructure is the digital manifestation of a research culture. However, even while designing cyberinfrastructure at this grand scale, some of the variables that contribute to the success of a single tool may also contribute to successful shaping of a digital culture. SNAC can serve as an illustrative example in this regard, too. To better serve the entire research community SNAC

33

intends to grow from the size of a research project to a self-sustaining national program. The Institute for Museum and Library Services (IMLS) has funded a project named Building a National Archival Authorities Infrastructure to help meet this goal. A proper mix of collaborators that represent the breadth of the project's stakeholders is essential to bring this plan to fruition. The IMLS funding will support a meeting of leaders in the archive, library, museum, scholarly, and funding communities to determine the requirements of building a sustainable National Archival Authorities Cooperative (NAAC) [IATH n.d. d]. This cooperative would go far beyond SNAC in scale, but share the same narrow scope as SNAC. It intends to do a few things very well, but at a national scale: provide integrated access to the records held in government, business, and research archives across the United States as well as access to the social networks of the people whose lives are documented in those records. At this level of cyberinfrastructure building, researching new technology plays less of a role than developing new policy recommendations for the business, governing, and technical requirements for NAAC. However, a great deal of emphasis is placed on the top of SNAC's infrastructure stack, the EAC-CPF standard [IATH n.d. e]. Writing about EAC-CPF several years ago, Pitti predicted that this "standard for creator description will facilitate building international, national, regional, and institutional biographical and historical databases that can serve as resources" [Pitti 2004]. To that end, the IMLS funding will also support workshops and scholarships to attend them in which EAC-CPF will be taught to increase understanding and expertise in archival authority control. Broad adoption of this standard by the archival community is a necessary component to building the immense collection of data which tools like those developed in the SNAC project require in order to become a major contribution to cyberinfrastructure.

As progress continues with SNAC, NAAC, and other related research efforts, it appears that tools like those discussed in this case study will continue to grow and serve their intended community of scholars. In time, perhaps new projects will emerge to fill gaps in other disciplines. As this happens, we will get closer and closer to the goal set forth by the ACLS [ACLS 2006] report on cyberinfrastructure to create an integrated digital representation of the cultural record, connect its disparate parts and make the resulting whole more available to one and all, over the network.

34

## Notes

[1] Thanks to Daniel V. Pitti (Associate Director, Institute for Advanced Technology in the Humanities) for demonstrating this social network example during a personal conversation.

## Works Cited

- ACLS 2006** American Council of Learned Societies (ACLS), *Our Cultural Commonwealth: The Final Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences* (2006). [http://www.acls.org/uploadedFiles/Publications/Programs/Our\\_Cultural\\_Commonwealth.pdf](http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf). Accessed March 30, 2012.
- Blackwell and Crane 2009** Blackwell, Christopher and Gregory Crane. "Conclusion: Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital Age." *Digital Humanities Quarterly*, 3 (2009). <http://digitalhumanities.org/dhq/vol/3/1/000035/000035.html>. Accessed March 29, 2012.
- Borgman 2009** Borgman, Christine L. "The Digital Future is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly*, 3 (2009). <http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html>. Accessed March 29, 2012.
- Boyd and Ellison 2008** Boyd, Danah M. and Nicole B. Ellison. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication*, 13 (2008): 210-230.
- Case 1991** Case, Donald Owen. "The Collection and Use of Information by Some American Historians: A Study of Motives and Methods." *The Library Quarterly*, 61 (1991): 61-82.
- Chuk et al. 2012** Chuk, Eric, Rama Hoetzlein, David Kim, and Julia Panko. "Creating socially networked knowledge through interdisciplinary collaboration." *Arts and Humanities in Higher Education*, 11 (2012): 93-108.
- Collins and Jubb 2012** Collins, Ellen, and Michael Jubb. "How do Researchers in the Humanities Use Information Resources?" *Liber Quarterly*, 21 (2012): 176-187.
- Collins et al. 2012** Collins, Ellen, Monica E. Bulger, and Eric T. Meyer. "Discipline matters: Technology use in the humanities." *Arts and Humanities in Higher Education*, 11 (2012): 76-92.

- Crane et al. 2009** Crane, Gregory, Brent Seales, and Melissa Terras. "Cyberinfrastructure for Classical Philology." *Digital Humanities Quarterly*, 3 (2009). <http://digitalhumanities.org/dhq/vol/3/1/000023/000023.html>. Accessed March 29, 2012.
- Daines and Nimer 2012** Daines III, J. Gordon, and Cory L. Nimer. "Re-Imagining Archival Display: Creating User-Friendly Finding Aids." *Journal of Archival Organization*, 9 (2012): 4-31.
- Dalton and Charnigo 2004** Dalton, Margaret Stieg, and Laurie Charnigo. "Historians and Their Information Sources." *College and Research Libraries*, 65 (2004): 400-425.
- Drucker 2009** Drucker, Johanna. "Blind Spots: Humanists must plan their digital future." *The Chronicle of Higher Education*, (3 April 2009).
- Duff and Johnson 2002** Duff, Wendy M. and Catherine A. Johnson. "Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives." *The Library Quarterly*, 72 (2002): 472-496.
- Friedlander 2009** Friedlander, Amy. "Asking questions and building a research agenda for digital scholarship." In *Working Together or Apart: Promoting the Next Generation of Digital Scholarship*, Washington, DC (2009), pp. 1-15.
- Hunt et al. 2011** Hunt, Leta, Marilyn Lundberg, and Bruce Zuckerman. "Getting beyond the common denominator." *Literary and Linguistic Computing*, 26 (2011): 217-231.
- IATH 2014** Institute for Advanced Technologies in the Humanities (IATH). "News." SNAC: The Social Networks and Archival Context Project. <http://socialarchive.iath.virginia.edu/news.html>. Accessed August 12, 2014.
- IATH n.d. a** Institute for Advanced Technologies in the Humanities (IATH). "NEH Proposal." SNAC: The Social Networks and Archival Context Project. [http://socialarchive.iath.virginia.edu/NEH\\_proposal\\_narrative.pdf](http://socialarchive.iath.virginia.edu/NEH_proposal_narrative.pdf). Accessed April 15, 2012.
- IATH n.d. b** Institute for Advanced Technologies in the Humanities (IATH). "Mellon Proposal." SNAC: The Social Networks and Archival Context Project. [http://socialarchive.iath.virginia.edu/Mellon2011\\_proposal\\_narrative.pdf](http://socialarchive.iath.virginia.edu/Mellon2011_proposal_narrative.pdf). Accessed April 29, 2012.
- IATH n.d. c** Institute for Advanced Technologies in the Humanities (IATH). "Project Team." SNAC: The Social Networks and Archival Context Project. <http://socialarchive.iath.virginia.edu/SNACI/staff.html>. Accessed April 22, 2012.
- IATH n.d. d** Institute for Advanced Technologies in the Humanities (IATH). "Meetings." Building a National Archival Authorities Infrastructure. [http://socialarchive.iath.virginia.edu/SNACI/NAAC\\_meetings.html](http://socialarchive.iath.virginia.edu/SNACI/NAAC_meetings.html). Accessed January 29, 2013.
- IATH n.d. e** Institute for Advanced Technologies in the Humanities (IATH). "Introduction." Building a National Archival Authorities Infrastructure. [http://socialarchive.iath.virginia.edu/SNACI/NAAC\\_index.html](http://socialarchive.iath.virginia.edu/SNACI/NAAC_index.html). Accessed February 5, 2013.
- IATH n.d. f** Institute for Advanced Technologies in the Humanities (IATH). "Forthcoming Features." SNAC: Social Networks and Archival Context. [http://socialarchive.iath.virginia.edu/forthcoming\\_features.html](http://socialarchive.iath.virginia.edu/forthcoming_features.html). Accessed August 13, 2013.
- Juola 2008** Juola, Patrick. "Killer Applications in Digital Humanities." *Literary and Linguistic Computing*, 23 (2008): 73-83.
- Kadushin 2012** Kadushin, Charles. *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press, New York (2012).
- Kolbitsch and Maurer 2006** Kolbitsch, Josef and Hermann Maurer. "The Transformation of the Web: How Emerging Communities Shape the Information we Consume." *Journal of Universal Computer Science*, 12 (2006): 187-213.
- Larson and Janakiraman 2011** Larson, Ray R. and Krishna Janakiraman. "Connecting Archival Collections: The Social Networks and Archival Context Project." In Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt (eds), *Research and Advanced Technology for Digital Libraries*, Springer Berlin/Heidelberg (2011), pp. 3-14.
- Library of Congress 2011 a** Library of Congress. "Encoded Archival Description: Version 2002 Official Site." The Library of Congress (November 1, 2011). <http://www.loc.gov/ead/>. Accessed May 4, 2012.
- Library of Congress 2011 b** Library of Congress. "Frequently Asked Questions." Library of Congress Authorities (October 4, 2011). <http://authorities.loc.gov/help/auth-faq.htm>. Accessed May 5, 2012
- Pitti 1999** Pitti, Daniel V. "Encoded Archival Description: An Introduction and Overview." *D-Lib Magazine*, 5 (1999). <http://www.dlib.org/dlib/november99/11pitti.html>. Accessed March 29, 2012.
- Pitti 2004** Pitti, Daniel V. "Creator Description: Encoded Archival Context." *Cataloging and Classification Quarterly*, 38 (2004): 201-226.

**Pitti 2006** Pitti, Daniel V. "Technology and the Transformation of Archival Description." *Journal of Archival Organization*, 3 (2006): 9-22.

**Schreibman and Hanlon 2010** Schreibman, Susan and Ann M. Hanlon. "Determining Value for Digital Humanities Tools: Report on a Survey of Tool Developers." *Digital Humanities Quarterly*, 4 (2010).  
<http://www.digitalhumanities.org/dhq/vol/4/2/000083/000083.html>. Accessed March 30, 2012.

**Seal 2012** Seal, Rob. "'Social Network for the Dead' Set to Expand Due to Mellon Grant." "UVA Today," (19 April 2012).  
<http://www.virginia.edu/uvatoday/newsRelease.php?id=18148>. Accessed April 22, 2012.

**Staatsbibliothek zu Berlin, n.d.** Staatsbibliothek zu Berlin. "EAC-CPF Homepage" Encoded Archival Context Corporate Bodies, Persons, and Families. <http://eac.staatsbibliothek-berlin.de/>. Accessed May 5, 2012.

**Svensson 2011** Svensson, Patrik. "From Optical Fiber To Conceptual Cyberinfrastructure." *Digital Humanities Quarterly*, 5 (2011). <http://digitalhumanities.org/dhq/vol/5/1/000090/000090.html>. Accessed March 30, 2012.

**Theimer 2011** Theimer, Kate. "What Is the Meaning of Archives 2.0?" *The American Archivist*, 74 (Spring/Summer 2011).

**Wilson 1998** Wilson, Stephen. *The Means of Naming: A social and cultural history of personal naming in western Europe*. UCL Press, London (1998).

**WorldCat n.d.** WorldCat. "What is WorldCat?" WorldCat. <http://www.worldcat.org/whatis/default.jsp>. Accessed April 27, 2012.



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.