Social Networks and Archival Context
R&D to Cooperative

Library Science Talks
September 2017
CERN Geneva / Zentralbibliothek Zürich
Overview

• Archival records and the description of people
• R&D Objectives and Results
• From R&D to Cooperative: Objectives and Results
• Social Networks and Archival Context Cooperative (SNAC) within the global cultural heritage landscape
• Brief look at the soon-to-be-released, revised and enhanced SNAC public interface
Records

• People living and working together record information
• Such information may serve a variety of purposes
• Some times the recorded information is intended to be a reliable witness to human activity: birth and marriage certificates are examples
• But even when the information is not primarily intended to be a “record,” it is evidence of human activity
• If you want to understand a publication, a building, a work of art, an event ... historical records are essential!
Archival Description

• Archivists describe not only the records themselves ...
• But also the contexts in which the records were created, accumulated, and used
• A key component of the context is describing the people who created and used the records, as well as, selectively at least, the people documented in them
• Records are largely unintelligible without intellectually preserving their context through description
Archival Description

• Thus archivists describe the creators and the contexts in which they worked and lived

• Their names, of course, but facts about them too: when and where they were active, what they did, and with whom

• Traditional library authority control was about managing the headings or entry points that appeared in catalog records

• For archivists, it is more about identities of persons, corporate bodies, and families (CPF entities)

• Though library authority control is also increasingly about identities
Quick History Overview

• Research & Development
  • 2010-2015
  • NEH and Mellon Foundation

• Cooperative Planning
  • 2011-2015
  • IMLS and Mellon Foundation

• Transformation into a Cooperative
  • Phase One 2015-2019
  • Mellon Foundation
R&D Objectives

• Demonstrate that data describing people in existing archival description can be used ...
  • To address the challenge of finding/discovering/locating/understanding distributed historical resources
    • Integrated access to geographically dispersed historical records
    • Access to the social-profession networks that created and are documented in the records
  • To lay the foundation for an international cooperative for centrally maintaining the collectively created biographical-historical data
R&D Activities

• Data sources: 2.25M WorldCat archival descriptions, 190K EAD-encoded finding aids, 400K or so British Library and NARA authority records, agency records from Smithsonian Institution Archives and NY State Archives, and more ...

• From sources, extracted and assembled descriptions of corporate bodies, persons, and families (CPF entities)

• Identity Resolution within assembled set and against VIAF to create final set of CPF descriptions

• Created a prototype “history research tool” (HRT)
  • Social-professional-intellectual networks (CPF to CPF relations)
  • Links to archival resources documenting the CPF entities (integrated access)
Identity and Identity Resolution

• Extracting the data from MARC, EAD, and other sources presented challenges, as did the development of the History Research Tool (HRT)

• But the central challenge was and is identity resolution: two or more people with the same name; two or more names for the same person

• A challenge for people and computers

• Quality of computational resolution depends on the *a priori* quality of human resolution
Identity and Identity Resolution

- Names are weak identifiers
- Life dates help, but they are still not enough
- Additional evidence is needed for reliable resolution; the more the better
- Each additional fact makes the identification more certain, more reliable
- A persistent identifier is just another name
- Essential is the set of facts associated with each identifier: name or names, existence dates, affiliated places, occupations, functions, significant events ...

Daniel V. Pitti, University of Virginia Library
R&D Results

• Original Source Records: 6,719,064
  • 4,653,365 Persons
  • 1,868,448 Corporate Bodies
  • 197,251 Families

• Merged Records: 3,741,262
  • 2,466,425 persons
  • 1,077,588 corporate bodies
  • 197,249 families

• All linked to archives resources in ~4000 repositories
From R&D to Cooperative Program: Objectives

• Primary objectives are practical
  • Sharing description of archives
  • Making description more effective
  • Improving the economy of research

• Intended as a contribution to the humanities and social science research infrastructure

• Improving the scholarly communication research economy
  • For curators
  • For researchers

• For more detail
  • http://socialarchive.iath.virginia.edu/SNAC-C_rationale.pdf
From R&D to Cooperative Program: Transformation

• Social
  • Administration
  • Governance
  • Building a community of editors

• Technological
  • From a set of independent steps that led to an aggregation of identity descriptions
  • To a dynamic, human curated collection of identity descriptions

Daniel V. Pitti, University of Virginia Library
Administration

• The University of Virginia Library hosts both the secretariat and the technology infrastructure of the Cooperative

• Director and deputy director

• Two programmers

• Additional administrative assistance provided as needed by Library staff

• The long-term home is to be determined in conjunction with developing a business model to ensure sustainability without regular grant funding
Governance

• Building a community with shared understanding and purpose
• Transitioning from central R&D decision-making to community governance
  • Editorial policy and standards
  • Technology
    • Infrastructure
    • Relation to other archival description systems (local; ArchivesSpace)
    • Relation to other identity resources (VIAF, national authority files, Wikidata …)
  • Communication: within the community and outreach
  • Training (SNAC School) – building a community of expert editors: international, cultural heritage professionals and humanities scholars
• Operations Committee to coordinate
Cooperative Members

- American Institute of Physics
- American Museum of Natural History
- Archives, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India
- Archives nationales de France
- Brigham Young University
- California Digital Library
- Cecilia Preston (individual scholar)
- George Washington University
- Getty Research Institute
- Harvard University
- Indiana University–Purdue University Indianapolis
- Jane Addams Papers (documentary editing)
- Library of Congress
- Mojave Desert Archives
- National Archives and Records Administration
- New York Public Library
- Princeton University
- Smith College
- Smithsonian Institution
- Tufts University
- University of California, Irvine
- University of Miami
- University of Nebraska
  - Library
  - Walt Whitman Archive (documentary editing)
- University of North Carolina, Chapel Hill
- University of Oregon
- University of Virginia
- Utah State Archives
- Yale University

Daniel V. Pitti, University of Virginia Library
Technology Transformation

• From an R&D pipeline with three steps
  • Extraction of data and assembling of CPF descriptions
  • Identity Resolution – match/merge
  • History Research Tool

• To an integrated maintenance and publishing platform
  • Supporting human editing of CPF descriptions
  • Supporting batch ingest of new data from new members or “data donors”

• Robust History Research Tool

• LOD exposure of the social-document network
SNAC Cooperative Identities

- Human editors: evaluate, verify, add new evidence & create, edit, link ...
- Smart people
- Sources: archives, libraries, museums, scholarly research projects …
- MARC21, EAD, TEI, Local formats
- Smart algorithms
- EAC-CPF

Evidence: sparse → dense
Certain: uncertain → certain
SNAC and the Global Cultural Heritage Landscape

• Broad perspective
  • The scope of archives: any person or group of persons that has ever lived and has left a recorded trace
  • The archival social-document network can provide context for other cultural heritage communities

• VIAF, Wikipedia/Wikidata, national library authority files, museum authority files (ULAN)
• ORCID and ISNI
• Multiple authorities undermines the function of authority
• Multiple is a political reality and ethically right
• Alignment of the multiple authorities will be an ongoing challenge
• Identifying is an ongoing activity and negotiation
• Though we can never get it right (once and for all), we can continue to make it better based on the available evidence, and based on shared intellectual and ethical values

Daniel V. Pitti, University of Virginia Library