National Archival Authorities Infrastructure

Social Networks and Archival Context
&
National Archival Authorities Cooperative

SNAC

- 2010-2012
  - National Endowment for the Humanities
  - Preservation and Access, Research and Development grant
- 2012-2014
  - Mellon Foundation

Project Team

- Daniel Pitti (PI) and Worthy Martin (Institute for Advanced Technology in the Humanities, University of Virginia)
- Adrian Turner and Brian Tingle (California Digital Library, University of California)
- Ray Larson (School of Information, University of California, Berkeley)

Project Objectives

- Archival finding aids currently intermix description of records with description of the creators of records and persons evident in the records
- Further the ongoing process of transforming archival description using advanced technologies
- By facilitating the *separation* of the description of *people* from the description of *records*
- Using EAC-CPF, an International archival authority control standard
- Goal: enhance the economy and effectiveness of archival description to improve access to and understanding of archival resources

Rationale for Separation

- Authority control of forms of names
- Flexible description
- Integrated access to cultural heritage
- Biographical/historical resource
- Social/historical context (social-professional networks)
- Cooperative authority control (more later)

The Data 2010-2012

- EAD-encoded finding aids
  - Library of Congress (1,546)
  - Online Archive of California (~15,400)
  - Northwest Digital Archive (5,160)
  - Virginia Heritage (8,390)
- Authority records
  - Library of Congress: NACO/LCNAF (3.8M personal names; 900K corporate names)
  - Getty Vocabulary Program: Union List of Artist Names (293K personal and corporate names)
  - Virtual International Authority File (16M+ personal names, corporate, uniform titles, jurisdictions)

Methods and Processing

- Extract EAC-CPF records from existing EAD-encoded archival descriptions
  - Extracting both creators and referenced CPF names
- Match EAC-CPF records against one another and against existing authority records (ULAN, VIAF, LCNAF); merge records for the same entity
  - Enhance EAC-CPF by normalizing entries, adding alternative entries, titles (VIAF), and historical data (ULAN)
  - Key challenge: two or more people with the same name; two or more names for the same person
- Create a prototype historical resource and access system
  - Historical data and social-professional networks
  - Links to archive, library, and museum resources (by and about)

EAD Source Data

- Encoded Archival Description
  - Intermixes description of creators of records and, at the discretion of the archivists, names associated with the content of the records
  - Detailed description of creators of records
- Widely varying quality
  - In the number of names identified and encoded
  - In the formation of the names (direct or inverted, capitalization, punctuation, and so on)
  - In the categorization of names (personal, corporate, or family
- Many names given but not identified as such
- Most important of these in biographies/histories and in correspondence description
- Extraction has focused on the “low hanging fruit,” that is the names tagged as names
- Attention shifting to names not identified as such

Archival Records

- Records are the by-products of people living and working as individuals, in organized groups, in families
- Records document people living and working
- People exist in social-professional contexts, in relation to others
- Records document these relations
- All records created by the same entity are described together (a fonds or collection)
  - Creators documented in detail
  - Many of the people documented in the record referenced in description
- Archival descriptions document interrelations among people and records (documents)

Source: J. Robert Oppenheimer Papers (LoC)

```
<origination>
        <persname source="lcnaf">Oppenheimer, J. Robert, 1904-1967</persname>
</origination>

<controlaccess>
        <persname source="lcnaf" encodinganalog="100" role="creator">Oppenheimer, J.
         Robert, 1904-1967</persname>
        <persname source="lcnaf" encodinganalog="600" role="subject">Bethe, Hans
         Albrecht, 1906- --Correspondence</persname> <!-- […] -->
        <persname source="lcnaf" encodinganalog="600" role="subject">Born, Max,
         1882-1970 --Correspondence</persname>
        <persname source="lcnaf" encodinganalog="600" role="subject">Boyd, Julian P.
         (Julian Parks), 1903- --Correspondence</persname>
        <persname source="lcnaf" encodinganalog="600" role="subject">Bush, Vannevar,
         1890-1974 --Correspondence</persname>
        <persname source="lcnaf" encodinganalog="600" role="subject">Casals, Pablo,
         1876-1973 --Correspondence</persname> <!-- […] -->
        <corpname source="lcnaf" encodinganalog="610" role="subject">Institute for
         Advanced Study (Princeton, N.J.)</corpname>
        <corpname source="lcnaf" encodinganalog="610" role="subject">Los Alamos
         Scientific Laboratory</corpname> <!-- […] -->
</controlaccess>
```

Source: Leonard Bernstein Collection (LoC)

```
<c02>
  <did>
    <container type="box">1</container>
    <unittitle>Aaltonen, Erkki <unitdate era="ce" calendar="gregorian">1981</unitdate>
    </unittitle>
    <physdesc>
      <extent>1</extent>
    </physdesc>
  </did>
</c02>
<c02>
  <did>
    <unittitle>Abbado, Claudio <unitdate era="ce" calendar="gregorian">1963-90</unitdate>
    </unittitle>
    <physdesc>
      <extent>5</extent>
    </physdesc>
  </did>
</c02>
[…]
```

```
<bioghist>
   <head>Biographical Sketch</head>
   <p>José Marcos Mugarrieta, prior to his term as Mexican consul in San Francisco 1857-1863, served in the Mexican army from 1837. He saw action in numerous battles and campaigns – Jamaica, under General Canalizo in 1841; Campeche, 1842-1843; Merida, 1843; Veracruz, 1845; Mexico City, 1846; Angostura and Cerro-gordo, 1847; Guanajuato, 1848, and Sierra-Gorda under Bustamante, 1848-1849; and Matamoros, 1849-1850. […]
</p>
   <p>In April 1857 Mugarrieta received an appointment from the Comonfort government for the consulship in San Francisco. He did not actually begin his new duties until September 1, 1859, due to illness and to the political situation in Mexico. […]</p>
</bioghist>
```

```
<bioghist>
  <head>Chronology</head>
  <chronlist>
    <chronitem>
      <date>1900</date>
      <event>Born on Jan. 20 in Hastings, Minnesota.</event>
    </chronitem>
    <chronitem>
      <date>1922</date>
      <event>Received baccalaureate from Princeton University, major in philosophy.
        </event>
    </chronitem>
    […]
    <chronitem>
      <date>1965</date>
      <event>Died on April 4.</event>
    </chronitem>
  </chronlist>
 </bioghist>
```

EAC-CPF

- Encoded Archival Context-Corporate bodies, Persons, Families
- An international communication standard for archival authority control
- Based on International Council for Archives, International Standard Archival Authority Records-Corporate bodies, persons, families (ISAAR(CPF))
- SAA Standards Committee, Technical Subcommittee on Encoded Archival Context
- Co-chairs
  - Katherine Wisser, Simmons College
  - Anila Angjeli, Bibliothèque nationale de France

Library and Archive Authority Control

- Library (or bibliographic) authority control is almost exclusively about the control of names
- Archival authority control involves biographical-historical description of the CPF entity
  - Descriptions based on controlled vocabularies or values, for example, occupations, place of birth and death
  - But also biographical-historical description
    - Prose
    - Chronological list
- Archival authority control provides **context** for understanding records, the context of their creation, the provenance

```xml
<identity>
        <entityType>person</entityType>
        <nameEntry  xml:lang="en-Latn">
                <part>Oppenheimer, J. Robert, 1904-1967.</part>
                <authorizedForm>AACR2</authorizedForm>
        </nameEntry>
        <nameEntry localType="VIAF:MainHeading">
                <part>Oppenheimer, J. Robert (Julius Robert), 1904-1967</part>
                <alternativeForm>VIAF</alternativeForm>
        </nameEntry>
        <nameEntry localType="VIAF:MainHeading">
                <part>Oppenheimer, Julius Robert, 1904-1967</part>
                <alternativeForm>VIAF</alternativeForm>
        </nameEntry>
                <nameEntry localType="VIAF:x400">
                <part>Oppenheimer, Robert</part>
                <alternativeForm>VIAF</alternativeForm>
        </nameEntry>
        <nameEntry localType="VIAF:x400">
                <part>Ou-pẽn-hai-mo, 1904-1967</part>
                <alternativeForm>VIAF</alternativeForm>
        </nameEntry>
</identity>
```

```
<existDates>
        <dateRange>
                <fromDate standardDate=“1904-04-22”>1904, Apr. 22</fromDate>
                <toDate standardDate=“1967-02-18”>1967, Feb. 18</toDate>
        </dateRange>
</existDates>
<!-- ... -->
<localDescription localType="subject">
        <term>Science--Societies, etc.</term>
</localDescription>
<localDescription localType="VIAF:nationality">
        <placeEntry countryCode="US"/>
</localDescription>
<localDescription localType="VIAF:gender">
        <term>Male</term>
</localDescription>
<languageUsed>
        <language languageCode="eng"/>
</languageUsed>
<occupation>
        <term>Physicists.</term>
</occupation>
<!-- ... -->
```

```
<chronList>
        <chronItem>
                <date>1904, Apr. 22</date>
                <placeEntry>New York, N.Y.</placeEntry>
                <event>Born, New York, N.Y.</event>
        </chronItem> <!-- ... -->
        <chronItem>
                <date>1943-1945</date>
                <placeEntry>Los Alamos, N. Mex.</placeEntry>
                <event>Director, Los Alamos Scientific Laboratory, Los Alamos, N. Mex.</event>
        </chronItem> <!-- ... -->
        <chronItem>
                <date>1954</date>
                <event>(1) Denied security clearance […] (2) Published Science and the
                        Common Understanding […]
                 </event>
        </chronItem> <!-- ... -->
        <chronItem>
                <date>1967, Feb. 18</date>
                <placeEntry>Princeton, N.J.</placeEntry>
                <event>Died, Princeton, N.J.</event>
        </chronItem>
</chronList>
```

```
<cpfRelation xmlns:xlink="http://www.w3.org/1999/xlink"
    xlink:type="simple"
    xlink:role="http://RDVocab.info/uri/schema/FRBRentitiesRDA/Person"
    xlink:arcrole="correspondedWith">
    <relationEntry>Bush, Vannevar, 1890-1974.</relationEntry>
    <descriptiveNote>
        <p>recordId: DLC.ms998007.r007</p>
    </descriptiveNote>
</cpfRelation>
```

```
<resourceRelation xmlns:xlink="http://www.w3.org/1999/xlink" xlink:arcrole="creatorOf"
        xlink:role="archivalRecords” xlink:type="simple”
        xlink:href="http://hdl.loc.gov/loc.mss/eadmss.ms998007">
        <relationEntry>J. Robert Oppenheimer Papers, 1799-1980 (bulk 1947-1967)</relationEntry>
        <objectXMLWrap>
        <did xmlns="urn:isbn:1-931666-22-9” >
                <unittitle>Papers <unitdate  normal="1799/1980” era="ce” calendar="gregorian">1799-
1980
                 </unitdate><unitdate label="Bulk Dates" type="bulk" normal="1947/1967”
                era="ce” calendar="gregorian">(bulk 1947-1967)</unitdate></unittitle>
                <unitid countrycode="US" repositorycode="US-DLC">MSS35188</unitid>
                <origination label="Creator">
                        <persname>Oppenheimer, J. Robert, 1904-1967</persname>
                </origination> <!-- ... -->
                <repository><corpname>Manuscript Division. Library of Congress</corpname>
                </repository>
                <abstract>Physicist and director
                of the Institute for Advanced Study, Princeton, New Jersey. [...] Topics include theoretical
                physics, development of the atomic bomb, the relationship between government and
                science, nuclear energy, security, and national loyalty. </abstract>
        </did>
        </objectXMLWrap>
</resourceRelation>
```

Year Two Results-Extraction

- Library of Congress: 43,702 EAC-CPF from 1,546 finding aids
  - corporateBody: 7,243
  - person: 36,012
  - family: 447
- Northwest Digital Archive: 24,949 from 5,160
  - corporateBody: 10,303
  - person: 13,294
  - family: 1,352
- Online Archive of California: 91,811 from ~15,400
  - corporateBody: 24,860
  - person: 66,329
  - family: 622

Year Two Results-Extraction

- Virginia Heritage: 15,175 from 8,390
  - corporateBody: 4,783
  - person: 9,919
  - family: 473
- Total: 175,637 EAC-CPF from 30,496
  - corporateBody: 47,189
  - person: 125,554
  - family: 2,894

Year Two Matching and Merging Results

- Total: 128,783 EAC-CPF from 175,637
  - – corporateBody: 31,282 from 47,189
  - – person: 95,583 from 125,554
  - – family: 1,918 from 2,894

Early Observations-Extraction

- Depth of analysis and quality of description of CPF entities varies widely in EAD-encoded finding aids
  - LoC a lot of names under authority control
  - OAC and NWDA have less names and control varies
  - VH still less names, more variance
- To be fair, the finding aids were created **without** SNAC processing in mind!

Next on Extraction

- Refine extraction processing, incorporating some NLP-like processing, for example
  - Verifying type of name: C or P or F
  - Massaging poorly formed names into better formed names
  - Identifying names in strings that are names-plus (but name not identified as such)
  - Provide context information to enhance matching, for example, date or dates of correspondence, or occupation of creator of records for referenced names

SNAC 2012-2014

- SNAC II: Mellon 2012-2014
  - 150,000 EAD-encoded finding aids
  - Most from U.S., but also U.K. and France
  - 1-2M WorldCat MARC archival descriptions
  - British Library: 300K names from mss. Collections
  - Smithsonian Institution: entire agency history; expeditions; and correspondents of Joseph Henry
  - National Archives and Records Administration (80K authority records
  - 16M VIAF clusters
  - And more ...

For more information on SNAC

- http://socialarchive.iath.virginia.edu/ (Project website)
- http://socialarchive.iath.virginia.edu/xtf/search (Public prototype)

National Archival Authorities Cooperative

- Building a National Archival Authorities Infrastructure
  - IMLS funded two-year project, October 2011-September 2013
  - EAC-CPF SAA workshops: 140 scholarships
  - National Archival Authorities Cooperative planning
  - Transforming SNAC into a sustainable national cooperative program

Benefits for Archivists

- Archival authority control at last!
- Best done cooperatively
- Consistent use of same form of name across descriptions
- This is can only effectively be accomplished by maintaining a single, shared authority file
- Economic benefits to cooperating: the creator in one description is the correspondent in another: people exist in social contexts, records document these contexts
- Working cooperatively will ensure identifying the interrelations of different collections
- Cooperative authorities will enable integrated access to distributed records: all of the records relevant to one person, corporate body, or family
- A shared national authority file would be a substantial historical resource, quite apart from the access enabled by it

Benefits for Users

- For scholars
- Integrated access to distributed archival resources
- Contextual data for not only the records of one creator, but other related records
- Access to the socio-historical networks in which people lived and worked
- A biographical-historical resource
- Time for an anecdote

But Not Only Scholars

- Use in K-12 education
- Time for an anecdote
- Life-time learners
  - Historical curiosity
  - Genealogy

Building the Infrastructure

- Institute for Museum and Library Services
- Funding two activities
  - 140 scholarships to seven regional workshops on EAC-CPF (Administered by Simmons College)
  - Series of meetings to develop a blueprint for a sustainable National Archival Authorities Cooperative (NAAC)
- Transforming SNAC into NAAC, project into program

NAAC

- Series of three meetings leading to the development of a blueprint
- All hosted by the National Archives and Records Administration
- Soliciting community input on the business, governance, and technological requirements
- First meeting broad, consensus building and idea gathering, followed by two meetings of three teams to address the requirements

First Meeting

- May 21-22
- Around 90 people
- Archivists, librarians, scholars (40 or so)
- Representatives of the federal repositories (40 or so)
- Funders (10 or so)
- Other stakeholders (OCLC and Getty Vocabulary Program)
- One and one-half day meeting

Federal Repositories

- National Archives and Records Administration
  - Including two presidential libraries
- Library of Congress
- Smithsonian Institution
- National Library of Medicine
- National Agricultural Library
- National Park Service

Conclusion

- This may well be a groundbreaking moment for the national archival profession
- An opportunity to do something really important, really useful
- To accomplish together what none can accomplish alone
- I hope (or is it now hopefully?)