The Social Networks and Archival Context Project: Status Report

Daniel Pitti*, Ray R. Larson**, Brian Tingle***

*Institute for Advanced Technology in the Humanities, UVA
**University of California, Berkeley - School of Information
***California Digital Library

Thanks to Daniel V. Pitti of the Institute for Advanced Technology in the Humanities, University of Virginia, and Brian Tingle of the California Digital Library for many of the slides here
Funding and People

• Funding and Timeline
  – National Endowment for the Humanities
  – May 2010-April 2012
  – Andrew W. Mellon Foundation
  – May 2012-April 2014

• People
  – 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)
Two Interrelated Project Objectives

• Further the transformation of archival description (separate description of records from description of people documented in them) in order to …

• Enhance access to archival resources, though in fact all cultural heritage resources

• Enhance understanding of resources by providing the social-professional context within which people lived and worked
The Source Data

- EAD-encoded finding aids (guides to archival records)
  - 150K
  - Primarily from U.S. sources, but also U.K. and France
- Archival authority records (360K)
  - National Archives and Records Administration
  - State Archive of New York
  - Smithsonian Institution
  - British Library
  - National Archives (France) & BnF
- WorldCat Archival Descriptions: 2M
Library and Museum Authority Records

• Getty Vocabulary Program: Union List of Artist Names (293K personal and corporate names)

• Virtual International Authority File (16M+ cluster records)
  – Contributed from around the world by national libraries and others
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)
  - Enhance EAC-CPF by normalizing entries, adding alternative entries, titles (VIAF), and historical data (ULAN)
- Create a prototype historical resource and access system
  - Historical data and social-professional networks
  - Links to archive, library, and museum resources (by and about)
Example EAD Record (Hub)

```xml
<EAD>
  <EADHEADER LANGENCODING = "ISO 639">
    <EADID>
      GB 0133 TAB
    </EADID>
    <FILEDESC>
      <TITLESTMT>
        <TITLEPROPER>
          Tabley Muniments
        </TITLEPROPER>
      </TITLESTMT>
      <PUBLICATIONSTMT>
        <PUBLISHER>
          John Rylands University Library of Manchester
        </PUBLISHER>
      </PUBLICATIONSTMT>
    </FILEDESC>
    <ARCHDESC LEVEL = "FONDS" LANGMATERIAL = "English">
      <DID>
        <REPOSITORY>
          University of Manchester, John Rylands University Library of Manchester
        </REPOSITORY>
        <UNITID ENCODINGANALOG = "ISADG3.1.1." COUNTRYCODE = "GB">
          GB 0133 TAB
        </UNITID>
        <UNITitle LABEL = "Title" ENCODINGANALOG = "ISADG3.1.2.">
          Tabley Muniments
        </UNITtitle>
        <UNITDATE LABEL = "Dates of Creation" ENCODINGANALOG = "ISADG3.1.3.">
          19th century
        </UNITDATE>
        <PHYSDESC LABEL = "Extent" ENCODINGANALOG = "ISADG3.1.5.">
          1.24 cu.m
        </PHYSDESC>
        <ORIGINATION LABEL = "Creator" ENCODINGANALOG = "ISADG3.2.1.">
          Warren, family, of Tabley, Cheshire
        </ORIGINATION>
        <FAMNAME SOURCE = "NCARULES">
          Warren, John Byrne Leicester, 1835-1895, 3rd Baron de Tabley, poet
        </FAMNAME>
        <PERSNAME SOURCE = "NCARULES">
          Warren, family, of Tabley, Cheshire
        </PERSNAME>
        <ORIGINATION>
          ... (Parts removed)...
        </ORIGINATION>
      </DID>
    </ARCHDESC>
  </EADHEADER>
</EAD>
```
Example EAD Record (Hub)

<BIOGHIST ENCODINGANALOG = "ISADG3.2.2;">  
<HEAD>  
Administrative/Biographical History  
</HEAD>  
<P>  
The poet John Byrne Leicester Warren, later 3rd and last Baron de Tabley, of Tabley near Knutsford, Cheshire, was born in 1835, the son of the 2nd Baron de Tabley (1811-1887), and his wife, Catherina. His mother was Italian, the daughter of the count de Soglio, and Warren spent much of his early childhood with her in Italy and Greece. He was educated at Eton and Christ Church, Oxford. At Oxford he published a volume of poetry. Originally he published under the pseudonyms George F. Preston (1859-1862) and William Lancaster (1863-1868), but latterly under his own name.  
</P>  
<P>  
His early verse included  
<TITLE>  
Praeterita  
</TITLE>  
(1863),  
<TITLE>  
Eclogues and Monodramas  
</TITLE>  
(1864),  
<TITLE>  
Studies in Verse  
</TITLE>  
(1865),  
<TITLE>  
Philocletes  
</TITLE>  
(1866), and  
<TITLE>  
Orestes  
</TITLE>  
(1868). His early work was Tennysonian in style, but he was later to be influenced by both Browning and Swinburne. In 1873 he produced …. (some data removed)…
Example EAD Record (Hub)

<SCOPECONTENT ENCODINGANALOG = "ISADG3.3.1.">
  <HEAD>
  Scope and Content
  </HEAD>
  <P>
  The collection consists mainly of the personal papers of the 3rd Baron de Tabley. The papers reflect his interests in literature, politics, botany and numismatics and include correspondence with numerous prominent later Victorian figures. Attention should also be drawn to de Tabley’s extensive and important collection of armorial bookplates.
  </P>
  <P>
  Correspondents include Sir Mountstuart Grant Duff, Edmund Gosse, Lord Houghton, A.C.Benson, and Robert Bridges. There are volumes of Tabley’s essays and verse, as well as a considerable number of notebooks and loose manuscripts of verse and other writings. There are various bundles and boxes relating to “Coins”, “Botany”, “Poetry”, “Literary”, and bookplates.
  </P>
</SCOPECONTENT>

<ADD>
  <OTHERFINDAID ENCODINGANALOG = "ISADG3.4.6.">
    <P>
    Preliminary survey list.
    </P>
  </OTHERFINDAID>
  <RELATEDMATERIAL ENCODINGANALOG = "ISADG3.5.3.">
    <P>
    There is correspondence with the 3rd Baron de Tabley among the Edward Freeman Papers, held at JRULM. The Library also has custody of the important Tabley Book Collection.
    </P>
  </RELATEDMATERIAL>
  <SEPARATEDMATERIAL>
    <P>
    The family and estate papers of the Leicester-Warren Family of Tabley are held by Cheshire Record Office. Some of these papers were originally in the custody of the John Rylands University Library of Manchester.
    </P>
  </SEPARATEDMATERIAL>
</ADD>
Example EAD Record (Hub)

```xml
<CONTROLACCESS>
  <HEAD>
  Index terms
  </HEAD>
  <GEOGNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "a">Tabley Inferior</EMPH>
    <EMPH ALTRENDER = "a-">Cheshire SJ7378</EMPH>
  </GEOGNAME>
  <PERSNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "surname">Benson</EMPH>
    <EMPH ALTRENDER = "forename">Arthur Christopher</EMPH>
    <EMPH ALTRENDER = "dates">1862-1923</EMPH>
  </PERSNAME>
  <PERSNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "surname">Bridges</EMPH>
    <EMPH ALTRENDER = "forename">Robert Seymour</EMPH>
    <EMPH ALTRENDER = "dates">1844-1930</EMPH>
  </PERSNAME>
  <PERSNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "surname">Duff</EMPH>
    <EMPH ALTRENDER = "title">Sir</EMPH>
    <EMPH ALTRENDER = "forename">Mountstuart Elphinstone Grant</EMPH>
    <EMPH ALTRENDER = "dates">1829-1906</EMPH>
  </PERSNAME>
  <PERSNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "surname">Gosse</EMPH>
    <EMPH ALTRENDER = "title">Sir</EMPH>
    <EMPH ALTRENDER = "forename">Edmund William</EMPH>
    <EMPH ALTRENDER = "dates">1849-1928</EMPH>
  </PERSNAME>
  <PERSNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "surname">Milnes</EMPH>
    <EMPH ALTRENDER = "forename">Richard Monckton</EMPH>
    <EMPH ALTRENDER = "dates">1809-1885</EMPH>
    <EMPH ALTRENDER = "epithet">1st Baron Houghton</EMPH>
  </PERSNAME>
  <PERSNAME SOURCE = "NCARULES">
    <EMPH ALTRENDER = "surname">Gosse</EMPH>
    <EMPH ALTRENDER = "title">Sir</EMPH>
    <EMPH ALTRENDER = "forename">Edmund William</EMPH>
    <EMPH ALTRENDER = "dates">1849-1928</EMPH>
  </PERSNAME>
  <SUBJECT SOURCE = "LCSH">
    <EMPH ALTRENDER = "a">Bookplates</EMPH>
  </SUBJECT>
  <SUBJECT SOURCE = "LCSH">
    <EMPH ALTRENDER = "a">Botany</EMPH>
  </SUBJECT>
  <SUBJECT SOURCE = "LCSH">
    <EMPH ALTRENDER = "a">Numismatics</EMPH>
  </SUBJECT>
  <SUBJECT SOURCE = "LCSH">
    <EMPH ALTRENDER = "a">Poetry</EMPH>
    <EMPH ALTRENDER = "y">Modern</EMPH>
    <EMPH ALTRENDER = "y">19th century</EMPH>
  </SUBJECT>
</CONTROLACCESS>
</ARCHDESC>
</EAD>
```
2010-2012 Extraction Results

- Source data: 30,000 finding aids
- EAC-CPF records extracted
  - LoC: 43,702 from 1,159 finding aids
  - OAC: 91,811 from ~15,400
  - NWDA: 22,609 from 5,160
  - VH: 15,175 from 8,390
  - Total 173,297
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)
  – Enhance EAC-CPF by normalizing entries, adding alternative entries, titles (VIAF), and historical data (ULAN)

• Create a prototype historical resource and access system
  – Historical data and social-professional networks
  – Links to archive, library, and museum resources (by and about)
The Problem

• Proliferation of the forms of names
  – Different names for the same person
  – Different people with the same names

• Examples
  – from Books in Print (semi-controlled but not consistent)
  – ERIC author index (not controlled)
Goethe, Johann

---Kindred by Choice. Waidson, H. M., tr. from Ger.

---Ballinger Pub.
---Van Goethe, J. W. see Goethe, Johann Wolfgang Von.
---Van Goethe, Johann see Goethe, Johann Wolfgang Von.


...etc...
John Muir
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, 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
Merging EAC-CPF Records

LCNAF Repository → VIAF Repository → ULAN Repository

Connect exactly matching records
Connect records using name authority information

Cheshire Search

EAC Repository → Repository of connected EAC Records

Merge → Repository of merged EAC Records
Merging EAC-CPF Records

VIAF Repository

Connect exactly matching records

EAC Repository

Connect records using name authority information

Cheshire Search

Repository of connected EAC Records

Merge

Repository of merged EAC Records
Connect Exact Matches

• The EAC-CPF records provide the names without having to parse texts, etc.
• Allows us to use some simple methods like exact matching
  – Assume identical name entries means the same person/corporate body/family
  – Enter the full names and record IDs into a database and flag IDs with same names for merging
But…

• Exact merging assumes that archives are following LC cataloging practice in their EAD records
  – There are some problems with this assumption
Some failures for merging...

- Different abbreviations:
  - A. & G. Carisch & C.

- And spacing issues:
  - A. C. Peters. (??)

- Completeness and alternate rules
  - Tabb, John B. (John Banister), 1845-1909.
  - Tabb, John Banister, 1845-1909.

- Also differing transliterations for non-Latin scripts
More…

• Variant romanizations (and spacing):
  – M. P. Belaieff.
  – M. P. Belaïeff.
  – M. P. Bieliaev.
  – M.P. Belaïeff.
  – M.P.Belaïeff.

• Initials vs. names:
  – Zabolotskii, N.A.
  – Zabolotskii.
More…

• Inverted order vs. uninverted
  – Taylor, Zachary, 1784-1850.
  – Zachary Taylor.

• Various combinations:
  – Tchaikovsky, Peter I.
  – Tchaikovsky, Pëtr II.
  – Tchaikovsky, Piotr Ilyich.
  – Tchaikovsky, Pyotr II.
  – Tchaikovsky, Pyotr Ilyich.
Merging EAC-CPF Records

1. EAC Repository
2. Connect exactly matching records
3. VIAF Repository
4. Connect records using name authority information
5. Cheshire Search
6. Repository of connected EAC Records
7. Merge
8. Repository of merged EAC Records
Search Authority Files

• For each name, formulate a search of the VIAF database using the Cheshire system (SGML/XML retrieval system with probabilistic and Boolean matching)
  – Search both the “authoritative” and “non-authoritative” forms
  – Consider any name matching a non-authoritative form to be a candidate match for the authoritative form
  – Flag EAC records that match the same authority record as potential matches
**Name:** Einstein Albert

**Shingle sequence:** ein, ins, nst, ste, tei, ein …, ert

Probability that the sequence (ins, nst, ste) follows ein is very high for the name einstein

**Shingle Language Model for names**

Krishna Janakiraman and Sean Marimpietri - Biograph
**Shingle Language Model for names**

Krishna Janakiraman and Sean Marimpietri - Biograph
Merging EAC-CPF Records

VIAF Repository

Cheshire Search

EAC Repository

Connect exactly matching records

Connect records using name authority information

Repository of connected EAC Records

Merge

Repository of merged EAC Records
Merge Flagged Records

• For all of the exact matches and authority matches
  – Use the Authoritative form of the name
  – Combine data from each match into a single EAC-CPF record
  – Retain all source record IDs and information

• Finally, output the merged EAC-CPF records
Inputs to SNAC merging

- LoC: 43,702 EAC-CPF records derived from 1159 finding aids
- OAC: 91,811 EAC-CPF records derived from ~15,400 finding aids
- NWDA: 22,609 EAC-CPF records derived from 5,568 finding aids
- VH: 15,184 EAC-CPF records
- Result: 128,783 “unique” names
Another view of the numbers…

- 93033 Person names merged from 114639 Person records
- 30161 Institutions merged from 41177 Institution records
- 1669 Families merged from 2263 Family records
Merging Conclusions

• There will not be a single merging method, but a staged set of approaches that will allow us to go from the simplest exact matches, to (we hope) reliably identifying various variant forms of a name, etc. when corroborated by contextual (date, etc.) information.
Next

• Process the 2M WorldCat archival descriptions
• Process the 150,000 finding aids
• Convert several hundred thousand archival authority records into EAC-CPF and match/merge process
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)
  – Enhance EAC-CPF by normalizing entries, adding alternative entries, titles (VIAF), and historical data (ULAN)

• Create a prototype historical resource and access system
  – Historical data and social-professional networks
  – Links to archive, library, and museum resources (by and about)
For More Information

- [http://socialarchive.iath.virginia.edu/xtf/search](http://socialarchive.iath.virginia.edu/xtf/search) (public prototype)