Leveraging VIAF in Social Networks and Archival Context

Authority Data on the Web
IFLA 2016

Daniel V. Pitti - IATH, University of Virginia
Preface

• Context: Archival Description
• Social Networks and Archival Context (SNAC)
• Current and future use of VIAF
• History Research Tool (demonstration)
Archival Records

• Records are the by-products of people living and working as individuals, in groups (organizations and families)
• Records *document* people living and working
• People and records exist in social-documentary contexts, in relation to one another
Description of Archival Records: Principles

• Archivist describe *Records in Context*: Principle of Provenance
  • Records from one source kept together (external)
  • “Original order” (internal)
  •Preserving the social and documentary context essential
  • Authenticity and integrity: evidence
Description of Archival Records: Practice

• Archival descriptive dominated by using one provenance-based apparatus that combines/intermixes context description and record description: the finding aid
  – Creators are documented in detail
  – Many individuals and groups documented in the records are also referenced in description (as access points or informally)
• Archival description documents interrelations among people and records
• A vast social-document network connecting the past to the present to the future
• Though interrelations have been more implicitly than explicitly documented
• Among other objectives, SNAC aspires to make the interrelations explicit
Overview of SNAC

• From R&D to Cooperative Program
• Funding: NEH (2010-12), IMLS (2011-2014) and Mellon (2012-2017)
• R&D Partners
  – University of Virginia, IATH
  – University of California, Berkeley SI
  – California Digital Library (University of California)
Inaugural Members of the Cooperative

- American Institute of Physics
- American Museum of Natural History
- George Washington University
- Getty Research Institute
- Harvard University
- Library of Congress
- 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 Virginia
- Yale University

Daniel V. Pitti - IATH, University of Virginia
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 and ...
  – To lay the foundation for an international cooperative for centrally maintaining the biographical/historical descriptions
R&D Data Sources

• 2.25M WorldCat archival descriptions (MARC21)
• Nearly 190,000 EAD-encoded finding aids, primarily from US and UK, though some French
• 300,000 British Library authority records
• NARA authority records
• Agency descriptions from Smithsonian Institution Archives/New York State Archives Archives
• And more ...
Methods and Processing

• Extract/Assemble/Migrate EAC-CPF (archival standard) records from existing archival description
  – Extracting both creators and referenced CPF names
• Match EAC-CPF records against one another and against existing authority records (VIAF)
  – Enhance EAC-CPF by normalizing entries, adding alternative entries, titles, “same as” links (VIAF)
• Create a prototype historical resource and access system
  – Social networks in which people lived and worked
  – Integrated access to distributed archival resources
  – Access to other resources by and about (publications, other artifacts)
The Identity Resolution Challenge

• Different names for the same person
• Different people with the same names
• Names are weak identifiers
• A challenge for computers
• A challenge for people
Extraction 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
SNAC Cooperative
Rationale for Cooperative

- Economy of sharing descriptive work: cooperatively curate the data
- Link the people descriptions to one another and to record descriptions: build the social-document network
- An international, Internet-based “linked archival authority system”
- Integrated access to distributed historical resources
- Expand the context and understanding of the records

Daniel V. Pitti - IATH, University of Virginia
Use of VIAF

• Use VIAF to “domesticate” identities gathered in the “wild”
• Curated data
  – An “axis” against which to match (identity resolution)
  – When matched, a source of data to enhance SNAC identities
    • Alternative names
    • Links to resources in WorldCat
    • “Same as” links to other authorities and resources: WorldCat identities; Wikipedia; NACO NAF; other national and domain authority files
• 24.8% of SNAC identities match VIAF
  – Archival domain and bibliographic domain only partially intersect
Going Forward

• Use of VIAF as a reference source for editors of SNAC identities
• Continue use of VIAF as “participant” in Identity Reconciliation
• A major challenge
  – Remaining current with VIAF
  – Identifiers in VIAF not fully persistent
  – Both SNAC and VIAF dynamic, changing

Daniel V. Pitti - IATH, University of Virginia
SNAC Cooperative Identities

**Sparse**
- Human editors: evaluate, verify, add new evidence & create, edit, link...
- Smart people

**Dense**
- EAC-CPF

**Certain**
- Sources: archives, libraries, museums, scholarly research projects...
- MARC21, EAD, TEI, Local formats
- Smart algorithms
- EAC-CPF

**VIAF USE**

Daniel V. Pitti - IATH, University of Virginia
Thank you

Homepage
http://socialarchive.iath.virginia.edu/

History Research Tool
http://socialarchive.iath.virginia.edu/snac/search

Cooperative Rationale
http://socialarchive.iath.virginia.edu/SNAC-C_rationale.pdf