Automated Name Authority Control

Mark Patton
David Reynolds
The Johns Hopkins University
Why do we need automated name metadata remediation?

- Inconsistent name representation
- Metadata harvested from multiple providers
- Hand-crafted data is expensive
- Commercial alternatives are expensive
ANAC background

- 29,000 Levy sheet music records
- 13,764 unique names
- 3.5 million LC name authority records (at the time of the project)
The Lester S. Levy Collection of Sheet Music

The Lester S. Levy Collection of Sheet Music is part of Special Collections at the Milton S. Eisenhower Library of The Johns Hopkins University. It contains over 29,000 pieces of music and focuses on popular American music spanning the period 1780 to 1960. All pieces of the collection are indexed on this site and a search will retrieve a catalog description of the piece. An image of the cover and each page of music will also be retrieved if the music was published before 1923 and is in the public domain.

Biography of Lester S. Levy

History of the Levy Collection

How to Use the Collection

Browse the Collection

Search the Collection

Home Browse Search Subject Headings How To Contact Us
ANAC Architecture

- Levy records stored as individual XML files
- MARC records stored in MySQL
- TCL scripting language
- Ease of implementation
<LSM:Title>Winter Polka, or, Recollections of a Merry Sleigh Ride.</LSM:Title>
<LSM:ComposerLyracistArranger>Composed by Wm. Vincent Wallace.</LSM:ComposerLyracistArranger>
<LSM:Name role="composer">Wm. Vincent Wallace</LSM:Name>
<LSM:Publication><LSM:Location>New York</LSM:Location></LSM:Publication><LSM:Publisher>William Hall & Son, 239 Broadway</LSM:Publisher><LSM:Date>1853</LSM:Date></LSM:Publication>
<LSM:FormOfComposition>sectional</LSM:FormOfComposition>
<LSM:Instrumentation>piano</LSM:Instrumentation>
<LSM:Dedicatee>Dedicated to Maurice Strakosch.</LSM:Dedicatee>
<LSM:EngraverLithographerArtist>Lith. of Sarony & Major, N. York</LSM:EngraverLithographerArtist>
<LSM:PlateNo>2274</LSM:PlateNo>
<LSM:Subject>Sleds & sleighs</LSM:Subject>
<LSM:Subject>Children playing in snow</LSM:Subject>
<LSM:Subject>Children misbehaving</LSM:Subject>
Problems with Levy data

- XML included some .html-like presentation information
- Names had to be extracted
- ANAC name extractor introduced error
- Date and location elements with bad data
Problems with LC data

- Matching on family name slow
- Not all Levy names represented in database
- MARC record format cumbersome
Ground truth generation

- Catalogers checked 2,841 random names from Levy against LC authority file.
- Used evidence such as name, date, notes, other publications.
- Took approximately 7 minutes per name.
- 28% did not have matching LC record.
ANAC

- Rank LC records by confidence
- Limit match possibilities to same family name
- Bayesian classifier calculates confidence based on evidence
- Names below a minimum confidence declared no match
- Train on ground truth data
Data: Levy records

- Given name
- Middle name
- Family name
- Modifiers
- Date
- Location
Data: LC records

- Given names
- Middle names
- Family name
- Modifiers
- Birth & death dates
- Context
Evidence

- Name equality and consistency
- Musical terms in LC record
- Publication date consistent with birth/death
- Publication place consistent with LC record
- New evidence can be added easily
<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.58</td>
<td>0.00</td>
</tr>
<tr>
<td>Accuracy (LC record exists)</td>
<td>0.77</td>
<td>0.00</td>
</tr>
<tr>
<td>Accuracy (LC record does not exist)</td>
<td>0.12</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Observations

- Matching very dependent on contextual data
- Machine matching much faster than manual
- Performance reasonable even with dirty metadata
- Machine matching could enhance manual work
Conclusions

- Combination of machine processing and human intervention produced best results.
- Approach could be tweaked by comparing names to multiple authority files or domain specific databases.
- ANAC not a generalizable tool, but others are out there.
Related Software

- Weka  http://www.cs.waikato.ac.nz/ml/weka
- GATE  http://gate.ac.uk/
- UIMA  http://www.research.ibm.com/UIMA/
- LingPipe  http://www.alias-i.com/lingpipe/
Relevant links


Discussion Questions

- How important is consistent name entry? Would it be more important for some communities than others?
- What types of domain-specific information might be available in OAI metadata that would help cluster names?
- What successes and/or failures have you had with automated name-authority control?