Enhancing an OAI-PMH Service Using Linked Data

The case of the Sheet Music Consortium

Stephen Davison, University of California, Los Angeles
New York: Howley, Haviland, Dresser, 1903

Race relations

Performance and performers

Graphic art

Musical composition

“When it’s moonlight on the Levee, Caroline”

“When I hear the banjos ringing”

Has: composer, lyricist, graphic artist, publisher, performers
men
women
Society and Culture--Sentimental song
Songs with piano
Songs
Landscapes
Legacies of Racism and Discrimination--Afro-Americans
Entertainment
Legacies of Racism and Discrimination--Stereotypes--Afro-Americans
Singers
Couples
Afro-Americans
rivers
Society and Culture--Couples
Performers--Men--Kenny
Kenny

Subject headings assigned by Duke University
The Nature of Sheet Music

- Cultural documents
- Multidimensional (variety of purposes)
- Various communities of interest
- Ephemeral in nature
- Printed components mixed, remixed upon reissue
- Variety of descriptive methods and levels
  - Special collections: Finding aids
  - Libraries: Library catalogs
  - Collectors: often interested in graphical components
- All this results in a challenge for a data aggregation service
The Sheet Music Consortium: history and background

• First version launched in 2002
  o 4 members
  o 7 contributing institutions
• “Next Generation” launched in 2011
  o 2 supporting institutions (UCLA, Indiana U)
  o 31 institutions, 29 collections, 228,000+ records
  o metadata mapped to MODS
  o user-contributed metadata services
• Going forward…
  o leveraging “next generation” infrastructure to support publication of linked data
Keep normalized and user-supplied data separate ...

- ... from the harvested metadata
- New data is not easily written back to contributing institution
- Association of harvested and contributed metadata could be lost upon reharvesting
  - Harvested data maintained in XML format and indexed using Solr
  - User contributed data is stored in a separate database
Welcome

The Sheet Music Consortium provides tools and services that promote access to and use of online sheet music collections by scholars, students, and the general public.

Not sure what to search for? Try Browsing.

News

June 6, 2013

Linked Open Data project

The Sheet Music Consortium is exposing music publisher information extracted from the Consortium's data as linked open data LOD) Read more.

February 15, 2013

Collections and services update

The Sheet Music Consortium now provides cross collection searching to 22 sheet music collections and a total of 348,514 items. If you manage a collection that you would like to be searchable through this site read the Participation page and contact us!

New services to be added soon include the ability to download records, assign additional metadata and save back to the Consortium, search user contributed metadata, and access our RSS feed.

June 7, 2011

New Collections, New Services

http://digital.library.ucla.edu/sheetmusic/
http://www.dlib.indiana.edu/smc/migrator/

Starting today this site becomes the official site of the Sheet Music Consortium. The previous site is still temporarily available for reference purposes at http://digital.library.ucla.edu/smc/ The new site offers improved searching and browsing, access to collections at 16 institutions (up from seven), and guidelines for the creation of descriptive metadata.

In addition there is a Metadata Mapping Tool, developed at Indiana University, making it possible to upload, crosswalk and validate metadata from a variety of sources, including plain text files, Microsoft Excel spreadsheets (xls or xlsx), or dBase files (dbf). If you have data in one of these
Yiddish Songs

Mein Weib Shlept Sich Mit. [English, Hebrew, and Yiddish]

Publisher: J. & J. Kammen
Repository: Johns Hopkins University, Lester S. Levy Collection of Sheet Music

Ich Benk a Heim. Das Voegele.

Publisher: Theodore Lohr
Repository: Johns Hopkins University, Lester S. Levy Collection of Sheet Music

Levine Levine Der Yid Mit Zein Machine. [In Yiddish]

Publisher: Gus Goldstein
Repository: Johns Hopkins University, Lester S. Levy Collection of Sheet Music

The Wedding Dress. [German dialect]

Publisher: Vaik's Theatre Co., 235-37 Bowery
Repository: Johns Hopkins University, Lester S. Levy Collection of Sheet Music

Der Naser Keiver Oder Churban Titanic

Publisher: New York : Hebrew Publishing Co.
Date: 1912
Repository: Mississippi State University, Charles Templeton Sheet Music Collection
Schemas and Workflows used to harvest records for the Sheet Music Consortium.

<table>
<thead>
<tr>
<th>SCHEMA</th>
<th>institutions</th>
<th>records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dublin Core</td>
<td>14</td>
<td>98,317</td>
</tr>
<tr>
<td>Qualified Dublin Core</td>
<td>9</td>
<td>26,236</td>
</tr>
<tr>
<td>MODS</td>
<td>4</td>
<td>103,504</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WORKFLOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct harvesting using the OAI protocol</td>
</tr>
<tr>
<td>Harvesting the metadata via the Static Repository Gateway</td>
</tr>
<tr>
<td>Manual extract of MARC records from an integrated library system and mapping to MODS and ingest</td>
</tr>
</tbody>
</table>
SMC and Name Authority

SMC metadata is harvested from diverse institutions, with varying practices

- inventories & finding aids
- spreadsheets
- bibliographic “records”
- focus on music vs. focus on illustrations
SMC and Name Authority

- Resources not always available for authority work at the point of description or aggregation

- Some important elements (e.g., Publisher) not traditionally subject to authority control

San Francisco: M Gray, 1879
Challenges of Aggregated Metadata

- Aggregating sheet music records by “work” (as identified by composer & title)

- Variations in practices by contributing institutions

- Example:
  - Harry Puck (composer)
  - Puck, Harry, 1890-1964
  - Puck, Harry [composer]
Challenges of Aggregated Metadata

- Sheet music “titles” difficult to define
  - First line of text
  - First line of the chorus
  - The same song may be published under multiple titles
    - California and you
    - California (and You)
    - Oh! you old Pacific Coast
  - A variety of distinct songs may have the same title
Options for publishing linked data

- **Works**
  - identified by title, composer, lyricist
  - Hard to identify reliably
- **Creators**
  - authority files exist, e.g. LCNAF
- **Subjects**
  - authority files exist, e.g. LCSH, TGM
- **Publishers**
  - generally not represented in existing authority files…
    some are represented in LCNAF, but usually because they have “authored” works (e.g. catalogs)
Roles of composers, lyricists, publishers & performers more interrelated than in many other forms of publication

On published items publisher names and locations change frequently

LOD provides us with a means of enriching bibliographic information and creating actionable metadata
Strategy for normalizing data

1. Extracted data (names, titles, publishers) from MODS records
2. Rank ordered word frequency using Voyeur/Voyant tools
3. Chose to work on group of dozen most important publishers
4. Used word frequency data to establish name and title groups
5. Used both internal and external information to establish when publishers really changed identity or ownership
6. Used Google Refine to normalize forms of name. Based choice of "preferred form of name" on frequency
7. Wrote these preferred forms back into the repository as "user supplied metadata" (i.e. separate from the harvested data)
8. Published publisher information on the web as HTML and LOD (RDF/XML) (plan also to publish RDFa)
9. Established unique ID’s, permanent URLs and link resolution for each publisher
Process for harvesting new data into the aggregated collection

OAI data harvest (new or updated) → Map harvested data to MODS → Parse data → Search local file for normalized form of publisher name

Found? (YES/NO)

Published metadata to SMC site

Add normalized forms of names to contributed data layer → Establish local publisher name record

New publisher? (YES/NO)

Add variant form of name to local file → Analysis and review of aggregated publisher information
# Summary of publisher information generated from SMC data

<table>
<thead>
<tr>
<th>PUBLISHER NAME</th>
<th>PUBLISHER ADDRESS</th>
<th>DATES OF PUBLICATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalmar &amp; Puck</td>
<td></td>
<td>1905</td>
</tr>
<tr>
<td>Kalmar &amp; Puck</td>
<td>152 West 45th Street, New York</td>
<td>1913-1915</td>
</tr>
<tr>
<td>Kalmar &amp; Puck</td>
<td>New York</td>
<td>1913-1916</td>
</tr>
<tr>
<td>Bert Kalmar &amp; Harry Puck</td>
<td>New York</td>
<td>1914-1915</td>
</tr>
<tr>
<td>Maurice Abrahams Music Co.</td>
<td>New York</td>
<td>1913-1915</td>
</tr>
<tr>
<td>Maurice Abrahams Music Co.</td>
<td>1570 Broadway, New York</td>
<td>1913-1916</td>
</tr>
<tr>
<td>Kalmar Puck &amp; Abrahams</td>
<td>New York</td>
<td>1915-1918</td>
</tr>
<tr>
<td>Kalmar Puck &amp; Abrahams</td>
<td>1570 Broadway</td>
<td>1917</td>
</tr>
<tr>
<td>Kalmar Puck &amp; Abrahams</td>
<td>Strand Theatre Building at 47th St</td>
<td>1917-1918</td>
</tr>
<tr>
<td>Maurice Abrahams, Inc.</td>
<td>1591 Broadway, New York</td>
<td>1923</td>
</tr>
<tr>
<td>Maurice Abrahams, Inc.</td>
<td></td>
<td>1923-1926</td>
</tr>
<tr>
<td>Kalmar &amp; Ruby Music Corp.</td>
<td>6301 Sunset Boulevard, Hollywood</td>
<td>1937-1939</td>
</tr>
</tbody>
</table>
## Timeline for Oliver Ditson, Music Publisher

<table>
<thead>
<tr>
<th>DATE</th>
<th>PUBLISHER</th>
<th>EVENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1835</td>
<td>Oliver Ditson, Boston</td>
<td>firm founded by Oliver Ditson</td>
</tr>
<tr>
<td>1867</td>
<td>Oliver Ditson, Boston</td>
<td>acquired Firth, Son &amp; Co., New York</td>
</tr>
<tr>
<td>1867</td>
<td>Charles H. Ditson, New York</td>
<td>firm founded by Oliver’s son</td>
</tr>
<tr>
<td>1873</td>
<td>Oliver Ditson, Boston</td>
<td>acquired Miller &amp; Beacham, Baltimore</td>
</tr>
<tr>
<td>1875</td>
<td>Oliver Ditson, Boston</td>
<td>acquired Wm. Hall &amp; Son, New York acquired Lee &amp; Walker, Philadelphia</td>
</tr>
<tr>
<td>1875</td>
<td>James E. Ditson, Philadelphia</td>
<td>firm founded by Oliver’s son</td>
</tr>
<tr>
<td>1879</td>
<td>Oliver Ditson, Boston</td>
<td>acquired G. André, Philadelphia</td>
</tr>
<tr>
<td>1883</td>
<td>Theodore Presser, Philadelphia</td>
<td>firm founded by Theodore Presser</td>
</tr>
<tr>
<td>1890</td>
<td>Oliver Ditson, Boston</td>
<td>acquired F.A. North &amp; Co., Philadelphia</td>
</tr>
<tr>
<td>1931</td>
<td>Theodore Presser, Philadelphia</td>
<td>acquired Oliver Ditson</td>
</tr>
</tbody>
</table>
Publisher LOD project objectives

● Add a layer of information to the aggregation that leverages existing information through a mixture of machine and human analysis
  ○ Map relationships between names
  ○ Additional derived information
    ■ Addresses and dates

● Publish publisher info in a variety of ways:
  ○ HTML
    ■ Visualization tools, mapping, timelines
  ○ RDF
  ○ RDFa
Archival Resource Keys (ARK) for publishers

<table>
<thead>
<tr>
<th>PUBLISHER</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kalmar &amp; Puck</td>
<td>ark:/21198/r23x84k8</td>
</tr>
<tr>
<td>Maurice Abrahams Music Co.</td>
<td>ark:/21198/r27p8w9m</td>
</tr>
<tr>
<td>Kalmar Puck &amp; Abrahams</td>
<td>ark:/21198/r2cc0xm5</td>
</tr>
<tr>
<td>Kalmar &amp; Ruby Music Corp</td>
<td>ark:/21198/r2057cvv</td>
</tr>
</tbody>
</table>

The Name-to-Thing (N2T) Resolver:

- Permanent URLs
  - e.g. [http://n2t.net/ark:/21198/r2cc0xm5](http://n2t.net/ark:/21198/r2cc0xm5)
- Institutional commitment: 21198 = UCLA
- Maintained by the UC Curation Center
<skos:prefLabel>Kalmar Puck &amp; Abrahams</skos:prefLabel>

<skos:altLabel>Kalmar, Puck &amp; Abrahams</skos:altLabel>
<skos:altLabel>Kalmar, Puck &amp; Abrahams Consolidated Inc.</skos:altLabel>
<skos:altLabel>Kalmar, Puck &amp; Abrahams Consol't'd, Inc.</skos:altLabel>

<rdfs:seeAlso rdf:resource="http://n2t.net/ark:/21198/r27p8w9m/"/>
<!--Maurice Abrahams Music Co.-->

<rdfs:seeAlso rdf:resource="http://n2t.net/ark:/21198/r23x84k8/"/>
<!--Kalmar & Puck-->

<rdfs:seeAlso rdf:resource="http://n2t.net/ark:/21198/r2057cvv/"/>
<!--Kalmar & Ruby Music Corp-->
MADS/RDF (Metadata Authority Description Schema in RDF) vocabulary

- A data model for authority and vocabulary data
- MADS/RDF is a knowledge organization system (KOS) designed for use with controlled values for names (personal, corporate, geographic, etc.), thesauri, taxonomies, subject heading systems, and other controlled value lists
- Fully mapped to SKOS vocabulary
- Designed specifically to support authority data as used by and needed in the library community
- Designed to support the description of cultural and bibliographic resources
<madsrdf:Address>
    <rdf:Description>
        <madsrdf:streetAddress>Strand Theatre Building at 47th Street</madsrdf:streetAddress>
        <time:year>1917</time:year>
        <time:year>1918</time:year>
    </rdf:Description>
</madsrdf:Address>
Conservatoire François Mitterand, Mauritius – SMC’s newest member

- Small collection of sheet music
- Looking for advice
- Wants to publish digital surrogates on the web
- Our strategy:
  - Create descriptive metadata in a local DB
  - Map to MODS using SMC’s online tool
  - Upload metadata to SMC’s Static Repository
  - Ingest to MSC using Static Repository Gateway
  - Metadata added to the Web of Data through SMC
Conclusions

Have demonstrated a strategy for mitigating some of the problems in aggregated metadata and publishing normalized data on the web as linked data.

Over time normalized linked data may take on the role that authority records do in OPACs, and may its way into formal authority vocabularies.

Publishers are just a start… now we need to republish other normalized elements to the “web of data.”

OAI is still a useful tool for harvesting data.

With mapping tools and static repositories even the smallest of players can contribute.

A possible model for other bibliographic projects.
With special thanks to my collaborators and co-authors:

Yukari Sugiyama
East Asia Library, Yale University

Elizabeth McAulay
UCLA Digital Library Program

Claudia Horning
UCLA Cataloging & Metadata Center

Stephen Davison
sdavison@library.ucla.edu