Extending Legacy Metadata with Linked Open Data

Jacob Jett (jjett2@Illinois.edu)
Timothy W. Cole (t-cole3@Illinois.edu)
Deren Kudeki (dkudeki@illinois.edu)
Myung-Ja Han (mhan3@Illinois.edu)
• **Exploring the Benefits for Users of LOD for Digitized Special Collections**
  – 18-month exploratory study
  – Funded by the Andrew W. Mellon Foundation

• Digitized Library Special Collections
  – Many relegated to information silos largely disconnected from the broader Web
  – How can we better connect these special resources to the Web?
    • Can we use Linked Open Data to help us? If so, how hard is it to do?

• Objectives
  – Map legacy metadata schemas to LOD-compliant schemas
  – Actively link to and from DBpedia, VIAF, Wikidata, and related Web resources
• The Motley Collection of Theatre & Costume Design
  – About 5,000 images of costume and set designs, sketches, production notes, and similar objects
  – Represents a variety of objects from the Motely Group’s career (1932-1976)
• Portraits of Actors, 1720-1920
  – Nearly 3,500 pictures of actors, including Sarah Siddons, Edmund Kean, and others
• Kolb-Proust Archive for Research
  – About 8,700 of Professor Philip Kolb’s research notecards on Marcel Proust
    • A chronology of events concerning Marcel’s Proust’s life
    • A bibliography of works mentioned in Marcel’s Proust’s correspondences
• 1-to-1 entity description violations
  – Each metadata record contained information about the image, the production or performance it related to, the associated people, and the associated play
  – Had to disambiguate these different entities from one another in order to produce quality Linked Data

• Mixed levels of authority control
  – Name authority control idiosyncratic and localized
  – Subject authority control used standard vocabularies and thesauri

• Conflated data (e.g., associated people: actors, composers, authors, etc., all in the same field)
Extending Legacy Metadata with Linked Open Data

Motley & Portraits of Actors Collections

Young Aristocrat: Act I, Scene 1
Second German General: Scene 8
First German General: Scene 8
Denisof; List of Singers and
What the CONTENTdm metadata actually describes

Linked Data Adds Context

Costume design by Motley
For a stage production directed by Glen Byam Shaw

Of a play by William Shakespeare

Performed at the Royal Shakespeare Theater

26 October 2017

Extending Legacy Metadata with Linked Open Data
Extending Legacy Metadata with Linked Open Data

```
<schema:VisualArtWork>
  "@type": "VisualArtwork",
  "creator": {
    "@type": "Organization",
    "@id": "http://viaf.org/viaf/121005107",
    "sameAs": "https://en.wikipedia.org/wiki/Motley_Theatre_Design_Group"
  },
  "name": "Caesar",
  "isPartOf":
</schema:VisualArtWork>

<schema:CreativeWork> (StageWork)

```
<schema:CreativeWork>
  "@type": "CreativeWork",
  "additionalType": "scp:StageWork",
  "name": "Iulius Casesar",
  "dateCreated": "1967",
  "locationCreated": {
    "@id": "https://en.wikipedia.org/wiki/Royal_Shakespeare_Theatre",
    "exampleOfWork": {
      "@id": "https://en.wikipedia.org/wiki/Royal_Shakespeare_Theatre",
    }
  }
</schema:CreativeWork>

<schema:Book>

```
<schema:Book>
  "@type": "Book",
  "author": [
    {
      "@type": "Person",
      "@id": "https://viaf.org/viaf/96994048",
      "name": "William Shakespeare",
      "sameAs": [
      ]
    }
  ]
</schema:Book>
```
Digital objects are TEI-encoded notecards, not metadata records
  - But they do contain metadata about named entities
Identifying named entities
  - Need to differentiate when named entities appear in citations vs. only mentioned in passing
Inconsistent structure of markup used to digitize cards insufficient to consistently differentiate the roles of various entities
The Kolb-Proust Archive for Research

Search: *plaignant* in the full text [X]
*Calmette, Gaston* (calmet1) in subject [X]

Results: 1 Item
Sorted by: relevance ▼ Go!

Preferred Name
- *Calmette, Gaston*
- *Chevassu, Francis*

Date
- **1913** (1)

Type
- *Chronology* (1)

1

Date: 1913 mercredi soir 12 novembre

Proust écrit à Calmette pour faire annoncer Swann, se plaignant: "... mes derniers articles avaient paru vous avoir été à charge, si bien que j'ai renoncé à finir la série [en 1912]. D'autre part, M. Chevassu n'a pas pris ceux que j'avais envoyés cette année (ils étaient trop longs) pour le Supplément. C'est pourquoi je me suis contenté d'envoyer aux Annales et au Temps des extraits qui y paraîtront."

à Gaston Calmette, Cor XII, p. 308, n. 142 [Le mercredi soir 12 novembre 1913]

Record: c69410
Proust écrit à Calmette pour faire paraître ses derniers articles extraits d'Annals et de Temps.

Extrait d'un article de Proust, Marcel; pseud. Horatio, 1913: 

"... mes derniers articles avaient paru vous avoir été à charge, si bien que j'ai renoncé à finir la série [en 1912].

D'autre part, n'a pas pris ceux que j'avais envoyés cette année (ils étaient trop longs) pour le Supplément. C'est pourquoi je me suis contenté d'envoyer aux Annales et au Temps des extraits qui y paraîtront."

Bibliographie:

Proust, Marcel; pseud. Horatio, 1913: Calmette, Gaston, Cor XII, p. 308, n. 142 [Le mercredi soir 12 novembre 1913]
Partial Mapping of TEI document elements

<table>
<thead>
<tr>
<th>TEI</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>div1 @id</td>
<td>schema:DataSet</td>
</tr>
<tr>
<td></td>
<td>schema:author <a href="http://viaf.org/44300868">http://viaf.org/44300868</a></td>
</tr>
<tr>
<td></td>
<td>schema:inLanguage “fr”</td>
</tr>
<tr>
<td>-&gt;head-&gt;date @value</td>
<td>schema:temporalCoverage [schema:DateTime]</td>
</tr>
<tr>
<td>-&gt;div2-&gt;p-&gt;name</td>
<td>schema:mentions [schema:Person]</td>
</tr>
<tr>
<td>-&gt;div2-&gt;note-&gt;name</td>
<td></td>
</tr>
<tr>
<td>-&gt;div2-&gt;p-&gt;title</td>
<td>schema:mentions [schema:CreativeWork]</td>
</tr>
<tr>
<td>-&gt;div2-&gt;note-&gt;title</td>
<td></td>
</tr>
<tr>
<td>-&gt;div2-&gt;(listBibl)-&gt;bibl</td>
<td>schema:citation [schema:CreativeWork]</td>
</tr>
</tbody>
</table>
Mapping Challenges Across all Collections

- Target vocabulary (Schema) still missing some key entities
  - Specifically no way to differentiate the production of a play from the individual performances
  - Solved by locally extending Schema
- Many entities are not currently listed in linked data sources
  - For Kolb-Proust we assigned URIs to every name and then linked the ones listed in authority control databases to those databases
  - Could do this for other collections
Enriching Metadata Links by Providing Linking or Canonical URIs for

- **Persons**
  - E.g., Peter Ustinov, Marcel Proust, etc.
- **Venues**
  - E.g., the Old Victoria Theatre, Alexandra Theatre, etc.
- **Plays/Productions/Performances**
  - E.g., *The Unknown Soldier & His Wife, Romeo & Juliet*, etc.
- **Subject Headings/Terms**
  - E.g., Theater—History, Costume Design, etc.
- **Bibliographic References**
  - E.g., *Figaro, Gaulois, Journal des Debats*, etc.
Focused on VIAF and DBpedia, which also linked to additional resources like,

- Library of Congress Name Authority File (LCNAF)
- OCLC’s WorldCat Identities
- Wikipedia

Also manually connected to non-linked data Web Resources

- Theatricalia
- International Broadway Database
- International Movie Database
- Gallica (BnF)
## Count of Motley Person URI’s found through manual processes

<table>
<thead>
<tr>
<th>Total persons identified in Motley metadata = 984</th>
<th>Count of URIs Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links have been found for 624 names</td>
<td></td>
</tr>
<tr>
<td>having Wikipedia / DBPedia links</td>
<td>311 (32%)</td>
</tr>
<tr>
<td>having VIAF links</td>
<td>218 (22%)</td>
</tr>
<tr>
<td>found by searching viaf.org directly</td>
<td>87**</td>
</tr>
<tr>
<td>found by searching LC Name Authority File</td>
<td>196**</td>
</tr>
<tr>
<td>found by searching WorldCat Identities</td>
<td>93**</td>
</tr>
<tr>
<td>*combined with automatic results</td>
<td>*582 (59%)</td>
</tr>
<tr>
<td>having Theatricalia links</td>
<td>475 (48%)</td>
</tr>
<tr>
<td>having IMDb links</td>
<td>353 (36%)</td>
</tr>
<tr>
<td>having IBDb links</td>
<td>42 (4%)</td>
</tr>
<tr>
<td>having more than 1 link</td>
<td>446 (45%)</td>
</tr>
</tbody>
</table>

*VIAF links for 476 persons (364 not found by manual search) were found using VIAF Auto Suggest
**Represents some overlapping results
Count of Motley Theater and Play/Performance URI’s found through manual searching

<table>
<thead>
<tr>
<th>Total theaters identified in Motley metadata = 59</th>
<th>Count of URIs Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links were found for 52 theaters</td>
<td></td>
</tr>
<tr>
<td>having Wikipedia / DBPedia links</td>
<td>49 (83%)</td>
</tr>
<tr>
<td>having VIAF links</td>
<td>45 (76%)</td>
</tr>
<tr>
<td>having home page links</td>
<td>36 (61%)</td>
</tr>
<tr>
<td>having other links</td>
<td>16 (27%)</td>
</tr>
<tr>
<td>having more than 1 link</td>
<td>47 (80%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total plays / performances identified in Motley metadata = 127</th>
<th>Count of URIs Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links were found for 105 plays / performances</td>
<td></td>
</tr>
<tr>
<td>having Wikipedia / DBPedia links</td>
<td>95 (75%)</td>
</tr>
<tr>
<td>having Theatricalia links</td>
<td>45 (35%)</td>
</tr>
<tr>
<td>having other links</td>
<td>10 (8%)</td>
</tr>
<tr>
<td>having more than 1 link</td>
<td>44 (35%)</td>
</tr>
</tbody>
</table>
### Late Breaking Figures for Kolb-Proust Archive Entities

**Total number of names found in the Kolb-Proust dataset = 5,727**

Links were found for 1,953 people

<table>
<thead>
<tr>
<th>Count of URIs Found</th>
<th>having VIAF links</th>
<th>having French Wikipedia links</th>
<th>having English Wikipedia links</th>
<th>having other links</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1,678 (29%)</td>
<td>1,236 (22%)</td>
<td>999 (17%)</td>
<td>264 (5%)</td>
</tr>
</tbody>
</table>

**Total number of notecards in the Kolb-Proust dataset = 8,716**

<table>
<thead>
<tr>
<th>Count of URIs Found</th>
<th>Citations found on notecards</th>
<th>Links founds for citations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13,923 (~1.6 citations/card)</td>
<td>4,812 (35%)</td>
</tr>
</tbody>
</table>
Observations

- Additional name authority sources needed for special collections
  - Many current sources are focused on authors
- When searching manually for entity links we found:
  - Easiest to start in WorldCat Identities; Google Web Search next best
  - Googling with full names and birth dates usually insufficient, needed to include additional keyword instead for best results
- Manual clean up of metadata and manual search helps recall:
  - Different name spellings/maiden names/nicknames, slightly different birth/death dates, and looking for contextual clues
Tasks & Work Time Spent for Motley Linking

- **Tasks**
  - Cleanup
  - Enhancement
  - Reconciliation
  - Link collection
- **Timeline**
  - 6 months graduate hourly work at 10-12hrs/week
  - Total = ~240 hours
What’s the benefit of all this?

How do we use it?
Google Structured Data Tool

Extending Legacy Metadata with Linked Open Data

26 October 2017
Context Enhancement

Extending Legacy Metadata with Linked Open Data

PROPERTIES OF THE SKETCH, A schema:VisualArtwork

PERFORMANCE AND VENUE ENTITIES REFERENCED

PERSONS REFERENCED

26 October 2017

21
Preliminary Findings & Conclusions

- Care needs to be taken when mapping legacy metadata to LOD-compliant vocabularies
  - May need to extend with additional entities and properties
  - However can sometimes be rewarded with additional linking properties (e.g., schema:mentions and schema:citation)

- User experiences enriched by adding contextual information
  - Through dynamically added sidebars and clickable links
    - Leverage existing Semantic Web sources
    - Provide opportunities for users to escape the siloed environments of traditional digital libraries
  - However, it is resource-intensive to manually add links, etc. to legacy metadata

- Additional Opportunities for leveraging the Semantic Web remain to be explored
The Alexandra Theatre was a theatre located in the Stoke Newington district of London. It was located at 65 and 67 Stoke Newington Road where the present-day Alexandra Court now stands.

History

Opened on December 27, 1897 as The Alexandra Theatre and Opera House, it was designed by theatre architect Frank Matcham for F. W. Purcell. Upon opening, the theatre had a capacity of 2,025, spread across pit, traditional circle and gallery seating. In 1904, city records list the theatre's capacity as 1,710, along with an assessed value of £1,250.[1] The theatre's first performance was the December 27, 1897 staging of Dick Whittington, an adaption of the pantomime Dick Whittington and His Cat.[2]

The theatre was operated by F. W. Purcell until 1905 when he sold it to new owners. The theatre's new owners changed its name to the Palace.
Data Visualizations
Future Work – Allowing User Annotations

The Daudet’s certainly look like an important family in Proust’s network.

Interesting how these two-times removed families still remain in the center of the network.

This small family seems to have a lot of co-occurrences, are they relatives?
Future Work – Knowledge Cards on Search Results Pages

**Rose**
Performance Title: Shoemaker’s Holiday
Theater: Young Vic
Opening Performance Date: 1947

**Apprentice**
Performance Title: Shoemaker’s Holiday
Theater: Young Vic
Opening Performance Date: 1947

**Thomas Dekker** (Author)
Thomas Dekker (c. 1572 – 25 August 1632) was an English Elizabethan dramatist and pamphleteer; a versatile and prolific writer, whose career spanned several decades and brought him into contact with many of the period’s most famous dramatists. (more from Wikipedia)

**Related Performances**
Shoemaker’s Holiday
Ellist of Edmonton

**Related People**
- George Devine
- Anna Stenham
- Melody Bragg
- [Missing Image]

**External Links**
- DBpedia
- DNB Record
- VIAF Record
- Theatricalia
- ISF Record
- Other

1 2 3 4 ... 12 Next>
Thank you for listening!

Questions?
<table>
<thead>
<tr>
<th>Field Name</th>
<th>Mapping to schema.org – schema:isPartOf (schema:CreativeWork)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>schema:additionalType (URL) [spc:StageWork]</td>
</tr>
<tr>
<td>Performance Title</td>
<td>schema:name (Text)</td>
</tr>
<tr>
<td>Theater</td>
<td>schema:locationCreated (schema:Place)</td>
</tr>
<tr>
<td>Opening Performance Date</td>
<td>schema:dateCreated (Date)</td>
</tr>
<tr>
<td>Notes</td>
<td>schema:description (Text)</td>
</tr>
<tr>
<td>Field Name</td>
<td>Mapping to schema.org – schema:exampleOfWork (schema:Book)</td>
</tr>
<tr>
<td>----------------------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>Performance Title</td>
<td>schema:name (Text)</td>
</tr>
<tr>
<td>[publication date]</td>
<td>schema:datePublished (Date)</td>
</tr>
<tr>
<td>[part of]</td>
<td>schema:isPartOf (schema:CreativeWorkSeries) [when true]</td>
</tr>
<tr>
<td>Author/Composer</td>
<td>schema:author (schema:Person)</td>
</tr>
</tbody>
</table>

Add @id (URI), schema:sameAs (URI), schema:birthdate (Date), etc. as found from LOD resources and services
Why Schema

- Industry-wide use by Web search engines
- Some promising schema’s (e.g., Bibframe 2.0, etc.) where still under development at the time of the project’s beginning
- Some existing schema’s were considered to “heavy-weight” for the project’s data needs and goals (e.g., FRBR\textsubscript{OO}, CIDOC-CRM, etc.)
- Some existing schema’s did not have wide-spread adoption (e.g., the SPAR family of ontologies)
- Were able to reuse previous library-oriented work (at UIUC and OCLC) with Schema.org