Exploring the Benefits for Users of Linked Open Data for Digitized Special Collections

White Paper #1

Transforming Special Collections Metadata into Linked Open Data: Mappings, entity reconciliation, workflows implemented & lessons learned

Myung-Ja Han, Alex Kinnaman, Timothy Cole, Ann Foster, Caroline Szylowicz

Transforming and migrating legacy metadata for special collections to a linked data compatible ontology requires metadata remediation, enhancement, and mapping. Entity reconciliation (adding the links) is a critical component as well. The first part of this white paper summarizes the mappings and workflows developed for our three digital special collections (Motley Collection of Theatre and Costume Design[1], Portraits of Actors[2], and the Kolb-Proust Archive Research[3]), the challenges encountered, and solutions identified for these challenges. The second part of this white paper describes entity reconciliation approaches used to discover links to more information about the entities mentioned in the metadata.

PART I: TRANSFORMATION MAPPINGS & WORKFLOWS

MOTLEY COLLECTION OF THEATRE AND COSTUME DESIGN AND PORTRAITS OF ACTORS

Metadata extraction

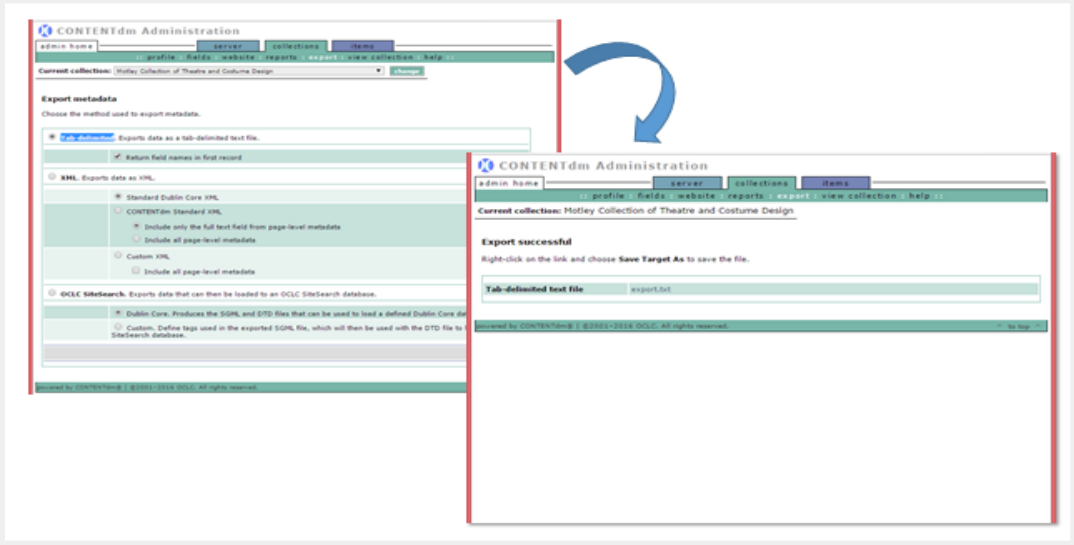

Figure 1: Metadata extraction from the CONTENTdm into a text file.

For the two special digital collections (Motley Collection of Theatre and Costume Design and Portraits of Actors) that were originally housed in the digital asset management system CONTENTdm, we exported each collection's metadata from CONTENTdm into a tab-delimited text file that included all local field

[1] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/motley
[2] http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/actors
[3] http://www.library.illinois.edu/kolbp/

names in its first row, i.e., suitable for import into Microsoft Excel -- see Figure 1. Having the metadata in a spreadsheet format facilitated subsequent metadata remediation, mapping and reconciliation tasks.

Metadata remediation

Although the legacy metadata for these two collections has been actively curated over the years, we identified two areas where further metadata remediation was needed before reconciliation and transformation into linked data could take place:

- metadata cleanup and enhancement
- dealing with metadata conflation

Metadata cleanup and enhancement

We found that pre-transformation metadata cleanup was an important task. While it can be time-consuming and labor-intensive, it is necessary since reconciliation is sensitive to the correctness of names and terms included in the metadata being transformed into linked open data. In cleaning our legacy records, we focused on personal, corporate, place names and subject terms since there are established controlled vocabularies for these entities that support linked data, i.e., vocabularies that associate classes of entities with persistent unique URLs useful for providing information about each individual entity. We examined terms used in our legacy metadata against appropriate linked data authorities and other resources, adding URLs and updating the strings in our metadata with controlled terms and preferred labels as applicable. For persons and organizations, we relied primarily on the following sources:

- Library of Congress (LC) Name Authority Files[4]
- Virtual International Authority File (VIAF)[5]
- Wikipedia[6]
- Worldcat Identities[7]
- Internet Movie Database (IMDb)[8]
- Internet Broadway Database (IBDb)[9]

(Note that the latter two resources listed are not yet linked data conformant.) As documented in Part II of this whitepaper, generally we found that the LC Name Authority Files, Wikipedia, IMDb, and IBDb were the most helpful. IMDb and IBDb were especially helpful given their focus on persons in the field of theater. These two sources and Wikipedia were also helpful for their more extensive contextual information, which helped ensure we had found the right person or organization.

There remained some individuals who we did not find in the sources listed above. This made finding URLs and/or confirming a person's or organization's details difficult. Some names we were able to find in online encyclopedias and smaller databases. Though Theatricalia[10] is not yet linked data conformant and is far

[4] http://id.loc.gov/authorities/names.html
[5] http://viaf.org
[6] https://en.wikipedia.org/
[7] https://www.worldcat.org/identities/
[8] http://www.imdb.com/
[9] https://www.ibdb.com/
[10] https://theatricalia.com/

from complete, we found it useful for confirming cast lists for specific productions, and gathering or confirming tentative birth and death dates. Theatricalia was also useful for cross-checking an individual's identity and his or her involvement with a particular performance. Additional sources for person and organization entities were found via simple Google search, and from following links found on pages for particular performances. J.P. Wearing's book *The London Stage 1930-1939: A Calendar of Productions, Performers, and Personnel*, and the subsequent editions for 1940-1949 and 1950-1959, were also useful for confirming cast and personnel lists of plays. In sum, our work revealed that there were a few names found only in other less common sources, such as:

- Canadian Theatre Encyclopedia[11]
- Encyclopedia Britannica[12]
- Turner Classic Movies[13]
- Goodreads[14]
- Obituaries in various digital newspapers
- Australian Dictionary of Biography[15]
- doollee.com[16]
- Opera Scotland[17]
- Specific textbooks found on Amazon Books[18]

For reconciling subject headings and obtaining appropriate URLs, the Library of Congress Subject Headings,[19] the Art & Architecture Thesaurus,[20] and the Thesaurus For Graphic Materials[21] proved useful and were closer to comprehensive since these sources had been consulted during original cataloging. While doing this work, metadata enhancement was done as well, e.g., information known or found during our review about items that was missing from our legacy metadata was added and capitalization, typos and punctuation errors were also corrected.

Dealing with metadata conflation

Metadata values in the excel spreadsheet were also changed during remediation processing. We identified that there were several fields that commonly contained more than one value. Such conflated values were separated with a semicolon. <Theater Name> and <Object> were two fields were this was a frequent issue. This is a common practice in the CONTENTdm software, i.e., when there is more than one value for a field, each value is separated by a semicolon, instead of repeating the field. However, we realized that when more than one personal name or subject term was present in a single field, it becomes

[11] http://www.canadiantheatre.com/
[12] https://www.britannica.com/
[13] http://www.tcm.com/
[14] https://www.goodreads.com/
[15] http://adb.anu.edu.au/
[16] http://www.doollee.com/
[17] http://www.operascotland.org
[18] https://www.amazon.com/books-used-books-textbooks/b?ie=UTF8&node=283155
[19] http://id.loc.gov
[20] http://www.getty.edu/research/tools/vocabularies/aat/
[21] http://www.loc.gov/pictures/collection/tgm/

harder to streamline and automate reconciliation workflows. So we de-conflated, i.e., divided these values into repeated fields with the same field names. Other fields had conflated meanings, e.g., <Author/Composer>. These were divided into two fields, <Author> and <Composer> so the mapping to schema.org semantics could be more accurately represented. This largely manual process was facilitated using the information hidden within the text strings. For example, the field <Associated People> included a name with a role in a parenthesis such as Shaw, Glen Byam (director). In addition to <director>, there are other roles that appeared in the field, including <actor>, <producer>, <dancer>, and <translator>. All of these roles have matching semantics in schema.org, so we created new fields to de-conflate and moved each person with a unique role into the appropriate field. New information about Associated People and their roles was confirmed by cross-referencing existing metadata and consulting outside resource. New/explicit roles for Associated People were added as appropriate. Names for which role information could not be determined were kept in the <Associated People> field (mapped to <schema:contributor>).

Metadata mapping to schema.org

The Motley Collection of Theatre and Costume Design and Portraits of Actors collections consist of performing arts related images described using a locally developed metadata schema based in large part on Dublin Core, but with field names specific to our user interface. (See Appendices 1 and 2 for a listing of these field names and their mappings to Dublin Core property names.) Described here is how we mapped our legacy theater metadata to schema.org semantics and the RDF data model.

Metadata for Theater Collections: Creating Relationships Between Item and Play

For the Motley Collection of Theatre and Costume Design our legacy schema consisted of 24 fields that describe (in conflation) the digitized item itself, the printed resource, the collection from which the item originated, and the original stage production for which the item was created. Our initial mapping to schema.org semantics relied on schema's <VisualArtwork> class to describe the item itself and on the <TheaterEvent> class to describe the original play and related information. As we progressed, we realized that the base type <TheaterEvent> (a kind of schema.org Event) was a poor match for the stage production since it was limited to discrete performances of it, i.e., events, rather than providing a way to describe a stage production. This was a complication because the <VisualArtwork>, e.g., a costume drawing, was only indirectly linked to a particular performance. We consulted with the schema.org community and discovered that an online theater ticket sales agency[22] had a similar use case to ours. We both presented our use cases to the schema.org community, leading to a suggestion of a new <CreativeWork> subclass -- <StageWork>. This entity better matches the actual representational situation: A <VisualArtwork> is part of a <StageWork>, i.e., a stage production, for which multiple <TheaterEvent>s are performed.

Our current working mapping therefore employs two subclasses of <CreativeWork>: <VisualArtwork> and <StageWork>. Each <VisualArtwork> has at least two <isPartOf> attributes to describe what the item is a part of, i.e., the Motley Collection (a <CreativeWork>) and one or more <StageWork>s. We use the URL of the item's splashpage (originally, its CONTENTdm reference URL) as the <VisualArtwork>'s identifier. Almost all of the <StageWork>s have descriptions in Wikipedia, so each <StageWork>'s Wikipedia URL is used to identify it.

[22] http://www.globetrottoirs.com

Field Name	**Mapping for schema:VisualArtwork**
Image Title	schema:name (Text)
Object	schema:genre (Text)
Type	schema:artform (Text or URL)
Material/Techniques	schema:artMedium (Text or URL)
Dimensions	schema:height & schema:width (schema:Distance or schema:QuantitativeValue)
Subject I (AAT)	schema:about (schema:Thing)
Subject II (TGMI)	schema:about (schema:Thing)
Subject III (LCSH)	schema:about (schema:Thing)
Rights	schema:copyrightHolder (schema:Organization or schema:Person)
Physical Location	schema:provider (schema:Organization or schema:Person)
Inventory Number	spc:standardNumber (Text or URL)
JPEG 2000 URL	schema:associatedMedia (schema:ImageObject) / contentUrl
Collection Title	schema:isPartOf (schema:Collection)
[Design by]	schema:creator (schema:Organization) [always Motley in this case]
[is part of Stage Production]	schema:isPartOf (schema:CreativeWork, spc:StageWork)
Field Name	**Mapping for schema:CreativeWork (i.e., StageWork)**
Performance Title	schema:name
Theatre	schema:locationCreated (schema:Place)
Opening Performance Date	schema:dateCreated (Date)
Notes	schema:description or schema:mainEntityOfPage
[additional type]	schema:additionalType (URL) [spc:StageWork]
[production of]	schema:exampleOfWork (schema:Book, fabio:Play)
Field Name	**Mapping for schema:Book (i.e., the text of the Play)**
Author/Composer	schema:author (schema:Person)
[additional type]	schema:additionalType (URL) [http://purl.org/spar/fabio/Play]
[Published Work]	schema:name
[publication date]	schema:datePublished (Date)
[part of]	schema:isPartOf (schema:CreativeWorkSeries) [when true]
[adaptation of]	schema:exampleOfWork (schema:Book or schema:CreativeWork) [when true]

Table 1: Mapping from Motley collection's local field names to schema.org.

The original play text, the theater(s) associated with the production (<StageWork>), people associated with the production, the text and score of the play (and it's author and composer), are linked to each <StageWork> using properties such as <exampleOfWork>, <locationCreated>, <contributor>, etc. Nesting is recursive as warranted, e.g., when a <StageWork> is an <exampleOfWork> of another <StageWork> which itself is an <exampleOfWork> of a <Book>. Because the proposal for a <StageWork> as a new schema.org class has yet to be adopted, the more generic <CreativeWork> class is used, with <StageWork> in our local namespace included as an <additionalType>.

In developing this mapping, we identified additional candidate properties for schema.org, useful for more fully describing a <VisualArtwork> or <StageWork>. Under <VisualArtwork>, we suggest two properties - <artStyle> and <artPeriod>, consistent with properties of the same name included in Visual Resources Association (VRA) Core[23] and Categories for the Description of Works of Art (CDWA)[24]. We also suggest <standardNumber> as a new property providing an identifying number given a <VisualArtwork> in a local context. For <StageWork>s, because descriptions include several personal names, each having a specific role, we initially considered proposing each role as a separate property, e.g., director, choreographer, dancer, set designer etc., to clearly describe roles played in the <StageWork>. But we decided to follow the common practice of using <contributor> with a role property. Other than <productionVisual>, a <StageWork> can have any property allowed a <CreativeWork>, including <text>, <name>, <dateCreated>, <locationCreated>, and <exampleOfWork>. By this mapping, two different relationships between resources have been created, between an item and a collection the item belongs to, and between an item and a play for which the item was created. See the complete current mappings in Table 1 above.

Portraits of Actors: Creating Relationships Between Item, Play, and Book

Keeping in mind the lessons learned from the mapping developed for the Motley Collection, we reviewed the legacy field names used for the Portraits of Actors Collection (see appendix 2) and determined that the legacy descriptions conflated descriptions of <Persons> and four distinct subclasses of <CreativeWork>: <VisualArtwork>, <StageWork>, <Book>, and <Collection>. For the <VisualArtwork>, we mapped fields that described the visual image itself, such as ID Number, Title, and Date. Generally the same, four relatively generic topical subject headings (Actors, costumes, Theatrical Managers, Theater--History--19th century--Pictorial works) were assigned to each portrait; these were also associated with the <VisualArtwork>, mapped as <schema:about>. The <isPartOf> property was used to associate each <VisualArtwork> with the Portraits of Actors collection. Local fields that described the production associated with a portrait were mapped to <StageWork> (as for Motley). The legacy subject heading naming the actor depicted in the portrait was classed as a <schema:Person> and associated with <StageWork> through the <contributor> property, and therefore associated only indirectly (i.e., through the <StageWork>) with the <VisualArtwork>, although the actor's name typically was embedded within the <VisualArtork>'s Title. When known <Person> properties such as jobTitle, birthdate, deathdate and sameAs, were included in the RDF graph generated. Any mentioned published work(s) associated with a production (<StageWork>), e.g., the text of a play, was classed as a <Book> with an <additionalType>, 'Play' as the default value; unlike in Motley, the source text for the play was not mentioned. Please see the complete mapping for the Portraits of Actors collection in Table 2 below.

[23] https://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf
[24] https://getty.edu/research/publications/electronic_publications/cdwa/definitions.pdf

Field name	**schema.org mapping Thing > Creative work > VisualArtwork**
ID Number	scp:standardNumber (Text or URL)
Title	schema:name (Text)
Date	schema:dateCreated (CreativeWork)
Role	schema:character (schema:Person)
Subject (LCSH)	schema:about (schema:Thing)
Type	schema:artform (Text or URL)
Dimensions	schema:height and schema:width
Technique	schema:artMedium (Text or URL)
Creator	schema:creator (schema:Person or schema:Organization)
Publisher	schema:publisher (schema:Organization)
Description	schema:description (Text)
Rights	schema:license and use URL. (The statement should be stored in somewhere, such as Project webpage.)
[copyright]	schema:copyrightHolder (schema:Organization and <rdf:about="http://viaf.org/viaf/123824539"> for UIUC Library as a default value.)
Collection	schema:isPartOf (schema:Collection)
Repository	schema:provider (schema:Organization)
[photo]	schema:associatedMedia (schema:ImageObject) / contentUrl
Field name	**schema.org mapping Thing > Creative work > StageWork**
Play	schema:name (Text)
Subject (Actor portrayed)	schema:contributor (schema:Person) / role, birthdate, deathdate, etc.
Field name	**schema.org mapping Thing > Creative work > Book**
[published work]	schema:name (Text)
[additional type]	schema:additionalType (URL) [http://purl.org/spar/fabio/Play]
Field name	**schema.org mapping Thing > Creative work > Collection**
Collection	schema:name
Physical collection	schema:isPartOf (schema:Collection) [asserting that Portraits Collection is part of the physical collection]

Table 2: Mapping from Portraits of Actors collection's local field names to schema.org.

Summary of Motley and POA Metadata Remediation & Mapping Processes

To sum up, figure 2 depicts the metadata processing workflow used to remediate and transform our legacy, non-MARC metadata into schema.org compliant RDF. As described above, metadata was exported into a spreadsheet format from our prior content management system (CONTENTdm), analyzed, de-conflated, remediated with links added (see discussion of reconciliation, below), and then re-serialized as JSON-LD,

one file per image. These JSON-LD files were then used to generate HTML for display with the source JSON-LD retained within a <script> tag (id='rdf') in the generated HTML. The Python scripts used to generate RDF (JSON-LD) from the spreadsheets of remediated Motley and Portraits metadata are available freely from our GitHub repository.[25]

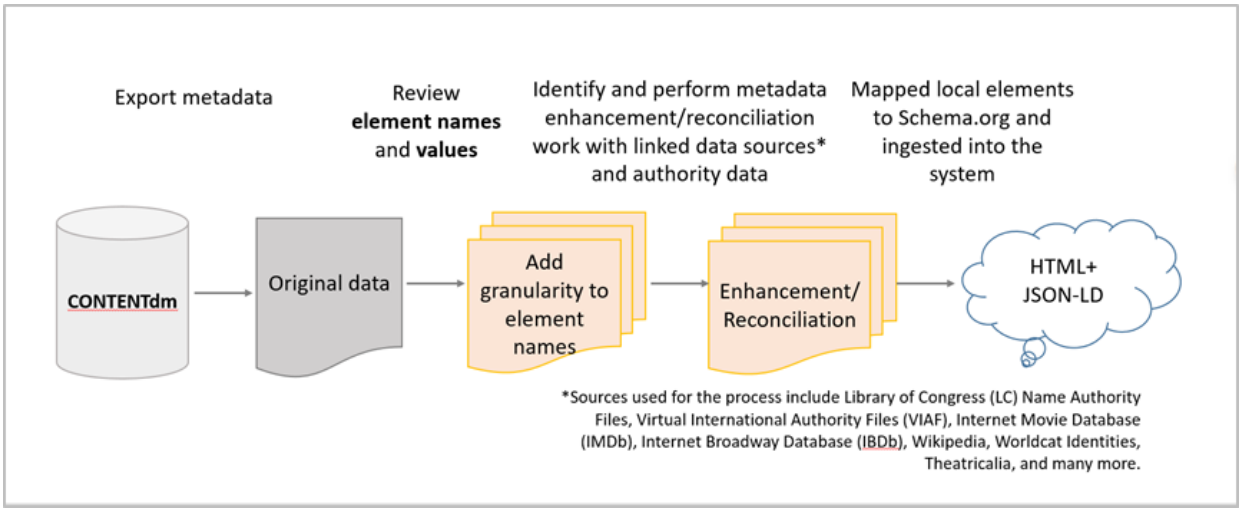

Figure 2: High-level schematic of metadata extraction from the CONTENTdm into JSON-LD and HTML.

KOLB-PROUST ARCHIVES FOR RESEARCH

The third collection, the Kolb-Proust Archive for Research (KPA)[26], presented its own unique challenges in terms of metadata. Unlike our collections of digitized image resources, the KPA collection consists of the transcribed textual research notecards of Professor Philip Kolb. These materials were developed by Kolb as he assembled a comprehensive, twenty-one volume edition of Marcel Proust's correspondence. About 10,000 of these note cards were subsequently encoded using TEI P5 version 2.0.0 to create the KPA. A local name database for all names (5,000+) that appear on these notecards is maintained as part of the Archive. Both the digitized cards and the names database records were transformed into RDF.

Transforming names database into linked open data

Full Name	KeyCode	Info
Daudet, Léon	daudet1	1868-1942, fils aîné d'Alphonse Daudet
Daudet, Marthe Allard, Mme Léon; pseud. Pampille	daudet6	1878-1960, cousine et 2ème femme de Léon Daudet,mariée en 1903
Daudet, Philippe	daudet10	?-1923, fils de Léon Daudet
Daudet, Claire-Antoinette	daudet11	1918- ; fille de Marthe (née Allard) et Léon Daudet (LJP)

Table 3: Kolb-Proust Archive for Research manages its own name database in a three column SQL database

[25] https://github.com/CIRSS/lod-project/tree/master/theater-collections/backend/portraits-of-actors, https://github.com/CIRSS/lod-project/tree/master/theater-collections/backend/motley
[26] http://www.library.illinois.edu/kolbp/

The KPA names database serves as a local name authority file, with authorized forms of all names and a range of additional information about each individual, e.g., dates of birth, wedding, divorce, family relationships (spouses, parents, children, etc.) and information about professions (illustrated in Table 3).

As with Motley and Portraits of Actors, the name metadata needed remediation before descriptions of the Person entities represented in the KPA names database could be reconciled (i.e., linked to external linked data services) and transformed into RDF. We began by extracting the legacy name records from the SQL database into a spreadsheet format, one name per row. We then reviewed the information found in the third column (labeled simply *Info*) and identified a set of relationships as well as specific conflated properties commonly present in this column that might facilitate reconciliation and help us to enhance the linked data RDF we created to describe the Person entities represented in the KPA name database.

Again we chose to encode this information using linked open data-compliant schema.org semantics. One of schema.org's classes, <Person>, supports a set of properties that work well for representing the relationships and the other kinds of information we found in the KPA names database. We selected the following 12 specific <Person> properties to encode our linked data RDF descriptions of individuals mentioned on Kolb's notecards:

- schema:familyName
- schema:givenName
- schema:birthDate
- schema:deathDate
- schema:gender
- schema:nationality
- schema:spouse
- schema:children
- schema:parent
- schema:sibling
- schema:relatedTo
- schema:jobTitle.

Columns were created in our names spreadsheet for each of these properties. We were able to populate several of these columns programmatically by parsing the metadata in our original 3 columns (as exported from the KPA names database). The following properties were populated using this approach:

- schema:gender was largely populated using titles found in name strings, e.g., Mme (Madame) and Mlle (Mademoiselle) indicating female, and M. (Monsieur) indicating male;
- schema:familyName and schema:givenName were populated by relying on punctuation;
- schema:birthDate and schema:deathDate were populated using dates appearing after names.

After this initial automated pass, a graduate research assistant was tasked to clean up, vet and further populate our spreadsheet, focusing especially on family name associations and the discovery of links to VIAF and Wikipedia (both English and French). For the most part the legacy metadata was not sufficiently consistent in format nor explicit enough in content to allow automated processing. For example the entries "Proust, Dr Adrien -- 1834-1903, père de Marcel Proust" and "Proust, Jeanne Clémence Weil,

Mme Adrien -- 1849-1905, mère de Marcel Proust" explicitly mention the father and mother relationships to M. Proust in the <Info> column, but only implicitly, describe their spousal relationship. Manual intervention maximized transformable information as illustrated in the graph shown in figure 3.

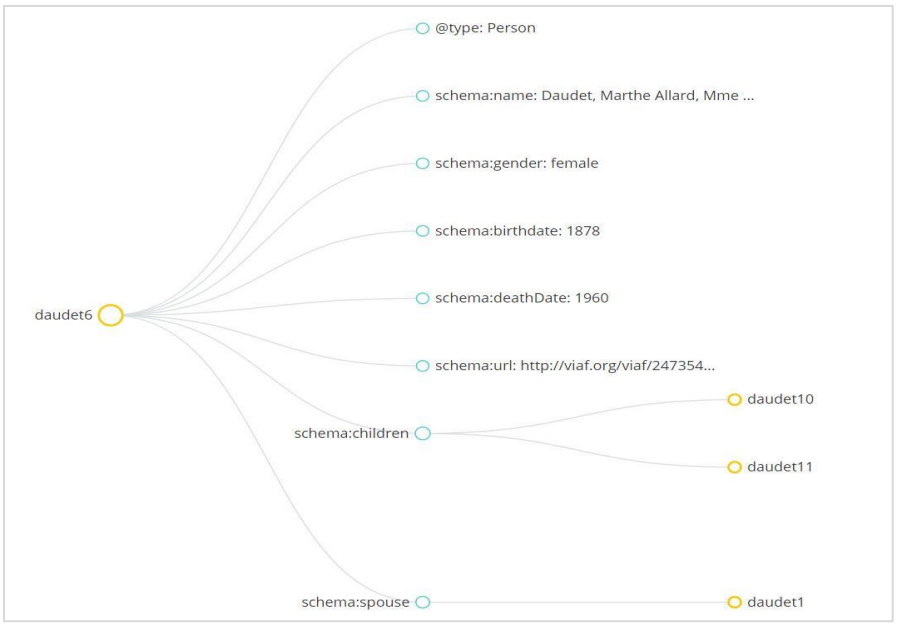

Figure 3: Metadata cleanup of the KPA names database resulted in more fulsome relationship information.

Publishing notecards as linked open data

```
<div0 id="c20090" type="card">
   <head>
     <date value="18990000">1899</date>
   </head>
   <div1 type="subdiv">
     <bibl>Proust. <title type="es">La Peste de Vienne et le danger que peut
faire courir à l'Europe la peste du Turkestan.</title> <title>Comptes-rendus
des séances de l'Académie des sciences morales et politiques</title>, 59e
année, p. 4
   </bibl>
   <bibl>Cf. 1897: Proust. <title type="es">La conférence sanitaire
internationale de Venise de 1897</title>, <title level="j">Revue
d'Hygiène</title>, vol. XIX, p. 7
   </bibl>
   <bibl>Cf. 1897: Proust. <title type="es">La défense de l'Europe contre la
peste.</title> <title>Comptes-rendus des séances de l'Académie des sciences
morales et politiques</title>, 57e année, p. 4
   </bibl>
  </div1>
</div0>
```

Figure 4: A sample TEI data

The original research notes contain large amounts of information and citations collected from French historical newspapers, books and magazines. The original bibliographic citations were generally found to be complete and to follow a consistent style but their elements were not explicitly encoded, beyond the

<title> and <name> element. As shown in figure 4 (above), when digitized, each physical card was recorded separately in a <div0> element, and each <div0> included a <date> element to record the temporal scope of the information described on the card. Each <div1> may have one or more <bibl> elements; these elements provide bibliographic references relevant to the information on the card. Detailed publication information is encoded with the <title> and sometimes with <name> including a proper role, such as <author>. Some cards start with a <p> for textual notes and add multiple <div1>s within a <listBibl> element. However, the content included in <div1> and <bibl> elements is consistent. Based on the TEI document structure and contents, we developed a mapping from the TEI semantics as used to schema.org sematics as shown below in Table 4.

TEI	**Schema.org**
div1 @id	schema:Dataset
	schema:author <http://viaf.org/44300868>
	schema:inLanguage "fr"
->head->date @value	schema:temporalCoverage [schema:DateTime]
->div2->p->name ->div2->note->name	schema:mentions [schema:Person]
->div2->p->title ->div2->note->title	schema:mentions [schema:CreativeWork]
->div2->(listBibl)->bibl	schema:citation [schema:CreativeWork]

Table 4: Mapping from TEI to Schema.org for Kolb-Proust Archive for Research

Once metadata remediation is complete, the enhanced and enriched spreadsheet is transformed into json-ld files (one per resource) using python scripts. These scripts are available from our GitHub repo.[27]

Observations on Metadata Mapping and Transformation to Linked Data

Preparing special digital collections metadata for linked data conversion requires several processes including metadata cleanup, enhancement, and reconciliation. Although metadata for these collections were created by subject specialists and had undergone several iterations of metadata cleanup, we found that linked data imposed a new set of challenges:

1. Working with unique local field names

While metadata describing items in digitized special collections often warrant descriptions with unique field names, when a local field name contains more than one meaning or multiple values with different roles in parentheses, those values AND the field name are better to be separated for the semantic mapping. As mentioned in the Motley collection mapping examples, the field name <Author/Composer> includes two distinct roles that schema.org can accommodate with two different properties. Another example is the local field <Associated People>, which included values with name and role for which each role is its own property in schema.org, such as director, producer, and etc. Although these local field

[27] https://github.com/CIRSS/lod-project/tree/master/kolb-proust/porcess-cards

names work perfectly fine in an isolated digital collections user interface, collection owners and metadata professionals need to review this approach when contemplating conversion to linked data.

2. Working with metadata values

Metadata value cleanup and enhancement processes that ensure the values used in the metadata will match terms found in controlled vocabularies and other authority services is the first step in moving toward linked data, because the reconciliation result depends on this conformity. However, metadata cleanup is not as easy as it seems. Many of the names used in special collections may not be well-known individuals whose names have established name authority entries. Also some names have been changed over time, so tracing these names and decisions on which variant is likely the preferred version for display and reconciliation became a challenge. This is in addition to outright reconciliation failures (discussed further below). For our theater-focused collections, we found that a significant number of performer names are not found in recognized library authority files. A similar problem was encountered with names of individuals in Proust's social network. Generally URLs did not already exist for these names. This reinforces the need for more discussions within the library community about establishing and developing a workflow for a local authority files tied into link data services that are broadly discoverable. While we mint URLs and make available brief RDF descriptions for entities mentioned in our metadata that we could not reconcile, we have no illusions given the current state of linked data service discovery, that these URLs will found and used (and potentially reconciled at a later data) by others. More work on this is clearly needed.

3. Importance of manual process

For this project, every step of metadata work - cleanup, enhancement, and reconciliation - required at least some initial manual steps, if only to understand what needed to be done. While, as described elsewhere in this whitepaper, more automation is feasible and essential for doing this work at scale, we also have learned that the batch process or automatic process can remediate metadata quality only so much, and that as an unintended consequence it can also cause unforeseen mistakes, i.e., wrong matches or adding wrong values into the wrong element. Although the metadata we have worked with for this project were created by subject specialists and were of fairly good quality, a new set of workflows for further data cleanup and enhancement was required for a successful transformation of the metadata to linked data. Also for the reconciliation work, when there are multiple entries with the same name, we painfully learned that machine could not disambiguate and identify the exact match. Rather it usually picked up the first entry. For this reason, about 240 hours of graduate student work was required for Motley and Portraits metadata cleanup and enhancement before transformation to linked data could proceed. After a similar amount of hours were spent to remediate he Kolb-Proust Archive for Research's metadata we found that though sufficient progress had been made to allow an initial transformation of KPA metadata into linked open data, we were only a little more than halfway done with all that we wanted to do -- meaning that the linked data graphs created for the KPA are uneven.

[CONTINUED]

PART II: ENTITY RECONCILIATION

ENTITIES FOUND IN DIGITIZED SPECIAL COLLECTION METADATA

The goal of entity reconciliation is to discover links (URLs) to pre-existing descriptions (preferably RDF descriptions) of the entities (names, topics, etc.) encountered in legacy metadata. Such linkages facilitate the implementation and enhance the functionality of interface(s) and allow the building of relationships with relevant external web resources. Links to more information about entities associated with a special collection resource can provide useful context and improve the connectedness of that resource. For our project, the biggest focus of our reconciliation work was Person entities. However, often the names included in our metadata were not those of book authors (which are easy to find in library name authorities), rather they were the names of actors, set designers, directors, or family friends with limited public profile. For this reason, identifying linked open data sources for our reconciliation work was challenging and required looking beyond traditional library authority services. This part of the whitepaper describes the processes employed for entity reconciliation work and documents our results, surfacing challenges that are likely to be encountered by many digitized special collections. The observations included at the end of this part of the whitepaper suggest approaches of possible interest to other curators and collections owners for their own entity reconciliation work.

MOTLEY COLLECTION OF THEATRE AND COSTUME DESIGN

For the Motley collection, our goal was to reconcile entities mentioned in item-level metadata and to clean the data as part of the metadata enhancement task. One of our metadata enhancement goals was to confirm individuals' roles in a specific performance, examples being "actor," "dancer," or "director." Finding identifiers led to external online resources that helped us establish the specific role of an individual in connection with a specific production (<StageWork>). The process began by searching for matches in VIAF[28], Worldcat Identities[29], Wikipedia[30], and Library of Congress Name Authority File[31]. Often these sources link to each other so the process moved quickly. This said we did note variations in the search services/APIs. We found that recall and precision for some names were better using one option, even when all had entries for a name, but there was no clear cut 'best' service/API for all of our names. The International Movie Database (IMDb)[32] and the International Broadway Database (IBDb)[33] were then searched for names not yet found. Finally, a simple Google search of the individual was conducted, usually as a combination of their name as it appears in the metadata and, the term "actor" or "actress" or "singer," usually without any birth and death dates included (which surprisingly reduced recall in many cases).

An additional resource, Theatricalia[34], an online collection of people, theaters, and specific performances, was also consulted and proved frequently beneficial, less as a source of linked data, but rather as another method for confirming individuals' involvement in specific performances. Finding information about a specific production (rather than simply a generic web page for a play) proved to be difficult; Theatricalia

[28] https://viaf.org/
[29] https://worldcat.org/identities/
[30] https://www.wikipedia.org/
[31] http://id.loc.gov/authorities/names.html
[32] http://www.imdb.com/
[33] https://www.ibdb.com/
[34] https://theatricalia.com/

aided in this effort and in confirming cast lists. Performance / production cast lists on Theatricalia were cross-checked against productions described in Motley metadata. When the majority of the cast members, director, and known theater all matched, the link was included. Additional sources that confirmed an individual's identity included obituaries, profile pages in various collections, and digital encyclopedias. These sources again did not often yield linked data URLs, but did disambiguate and confirm identity. As shown in Table 5 and Figure 5, both manual and automated searching was done. The automated searching relied on the VIAF Auto Suggest API[35] using updated versions of scripts previously developed here and available from our GitHub Repository[36] (same scripts used for Portraits and KPA name reconciliation).

Person entities

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

* VIAF links for 476 persons (364 not found by manual search) were found using VIAH Auto Suggest
** Represents some overlapping results

Table 5: Results of the Motley Collection personal name reconciliation work

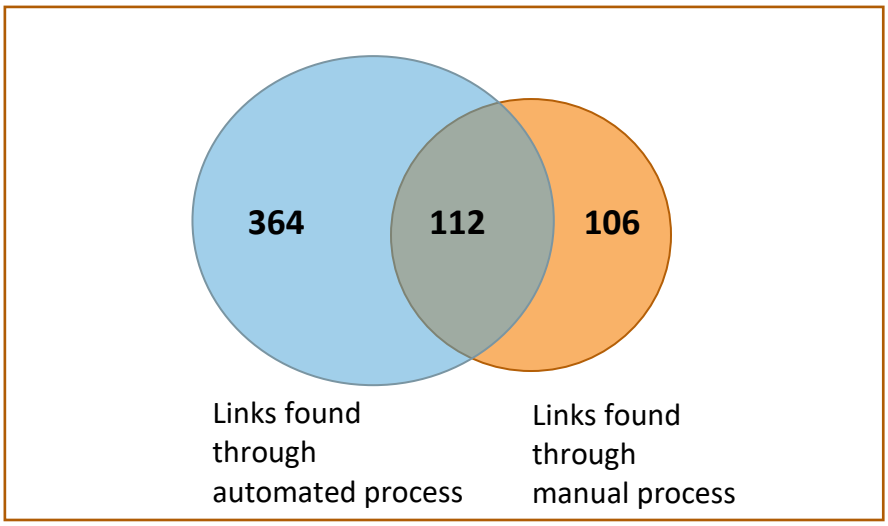

Figure 5: VIAF links found from automated and manual processes

[35] https://platform.worldcat.org/api-explorer/apis/VIAF/AuthorityCluster/AutoSuggest
[36] https://github.com/dkudeki/Fix-Authorities/tree/simple-name-reconciliation

Reconciliation work was performed manually by searching the sources mentioned above. Of the 984 names referenced in Motley metadata, we manually found 218 (22%) names in VIAF links and 311 (32%) names in Wikipedia. In addition, the manual process found 475 (48%) names in Theatricalia and 353 (36%) names in IMDb. Often information associated with a single name can be found on multiple Websites; we found this to be the case for 446 (45%) of the Motley names. The VIAF Auto Suggest API helped us find 476 Motley names, but note that the VIAF Auto Suggest API failed to find VIAF graphs for 106 of the names found in VIAF through manual searching of VIAF and other resources with links to VIAF. Combining manual and automated searches, we found VIAF URLs for 582 (59%) of the Motley names. Obviously automated searching scales better and on the whole had better recall, but it did miss almost 20% of the VIAF links for <Persons> that we were able to find through manual searching.

Theater entities

A similar manual process was used to discover links for the theaters (venues) mentioned in Motley Collection metadata. IMDb was not included as a search target, but URLs from Google searches for theater home pages were included. Theaters required slightly more attention as names often changed based on who owned the theater or if there were memorials to a distinguished individual or family. An example is a theater listed in Motley as the Martin Beck Theatre (New York, N.Y.) that changed its name to the AL Hirschfeld Theatre in 2003. Theaters and their companies often were both listed under the theater name despite one being an institution and the other being an organization, such as the Shakespeare Memorial Company and the Shakespeare Memorial Theatre. Another challenge is the name of the theater changing based on the gender of the monarch in England, such as “His Majesty’s Theatre” and “Her Majesty’s Theatre,” which is the same theater. Other anomalies were the venues in metadata given as a city or building name, such as “Leningrad,” “Kronborg Castle,” or “Madison Square Gardens,” which were still included during data collection but are clearly not theatres. The Motley metadata for "Theatres" also included MGM and Columbia Pictures as values, which are film production studios, for which data was again included for the sake of thoroughness but not as a theater. Table 6 shows the reconciliation results for theatre names.

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

Table 6: Results of the Motley Collection theater name reconciliation work

Productions (<StageWork>s) and Plays

The discovery process for collecting URLs for productions and plays was challenging. In particular it proved infeasible to de-conflate these two classes of entities. There was simply not enough time to sort through all the sources that would have been useful for disambiguating individual stage productions, e.g., the playbills, scripts, and similar memorabilia. Instead we typically found it better to make do with Wikipedia links for the general play (rather than the specific production). We also searched Theatricalia which more often yielded links for a specific performance, but was not linked data. Results are shown in table 7.

Total plays/productions in Motley metadata = 127 Links were found for 105 plays/productions	**Count of URIs Found**
having Wikipedia/DBPedia links	95 (75%)
having Theatricalia links	45 (35%)
having other links	10 (8%)
having more than 1 link	44 (35%)

Table 7: Results of the Motley Collection play/production name reconciliation work

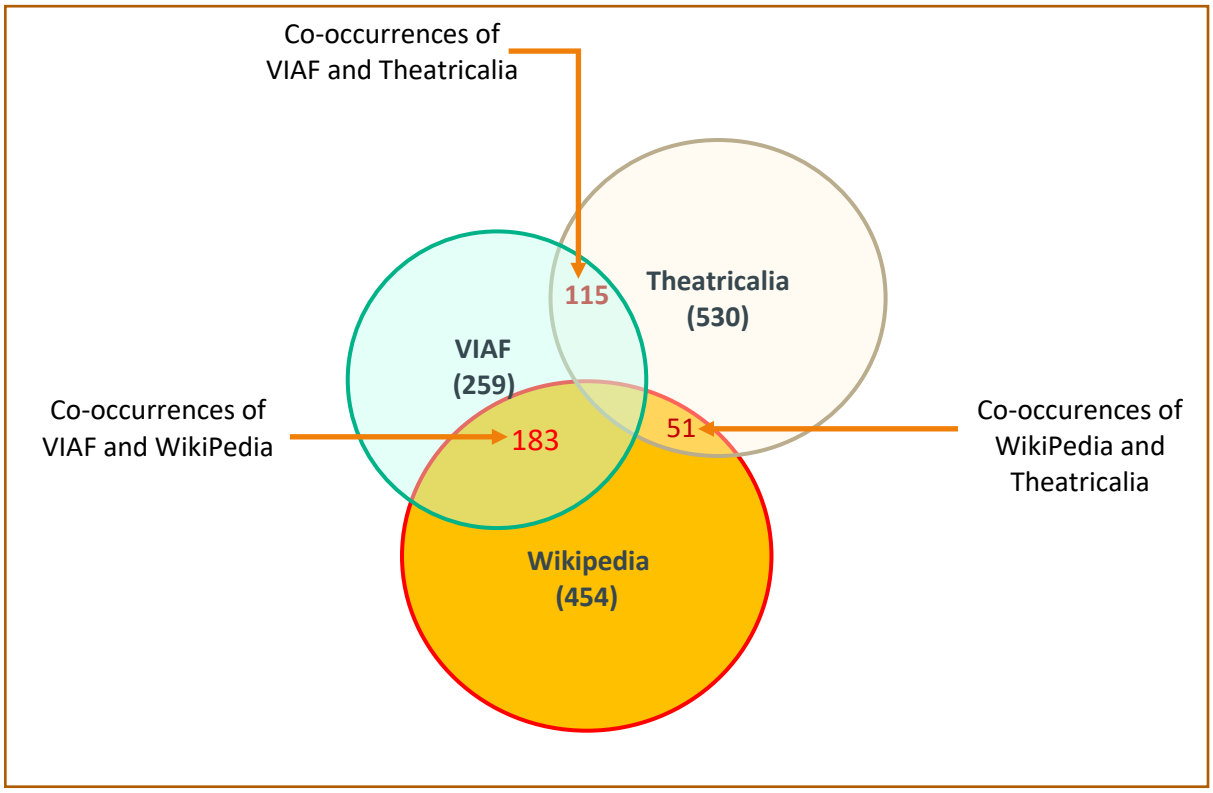

Figure 6: Co-occurrences and unique matches between sources of the Motley Collection name reconciliation work

Combining all reconciliation results for people, theater, and play/production names together, the most unique matches for the Motley Collection metadata were found in Theatricalia (530). Wikipedia had the most matches (688) overall (i.e., when including overlap with VIAF and Theatricalia) as shown in figure 6.

PORTRAITS OF ACTORS

Total names found in the Portraits of Actors = 865	**Count of URLs Found**
having VIAF links	441 (51%)
having Wikipedia links	301 (35%)
having VIAF and Wikipedia links	301 (35%)

Table 8: Results of the Portrait Collection name reconciliation work

Reconciliation work for names in the Portraits of Actors (Portraits) Collection started with automated searches of VIAF and Wikipedia. A similar manual process as used for Motley also was done. If a VIAF

source was found, links to other sources were followed to confirm identification. Sometimes WorldCat Identity was the sole link in VIAF, but then WorldCat Identity would often link to Library of Congress Name Authority File and/or Wikipedia. The individuals in the Portraits collection are on average older than Motley, some having birth dates in the 1600's and 1700's. For this reason, we discovered fewer URLs and a slight increase in name ambiguity. For example, Matthew Buchinger with the German spelling is Matthias Buchinger, and misspelled Mathew Buckinger. He has two VIAF entries and two WorldCat Identity entries, one each for the English and German spelling, and a single Wikipedia entry. He also has the added confusion of having different death dates, with VIAF marking his death in 1739 and Wikipedia in 1740. The reason all of these identities can be confirmed is because all of the sources describe him as being a German man born without arms and legs. This description combined with similar names and similar birth and death years is distinct enough to reasonably assume that they refer to the same entity.

Furthermore, some individuals do not have their own individual URL but could be found as part of a company, minstrel group, or troupe of which they were a member. 'Dan Bryant, 1833-1875,' for example, has VIAF, WorldCat, and LOC entries, but the only Wikipedia article mentioning him is about a group called Bryant's Minstrels rather than just Dan Bryant himself. Similarly, George Swayne Buckley is mentioned in the Wikipedia article for Buckley's Serenaders, with a slight spelling variation in his name, but consistent active career years. For these instances of groups being the only linked resource pertaining to an individual in Portraits, a new column, MemberOf was added (MemberOf is a property of Schema.org <Person>s). Table 8 shows name reconciliation results for the Portraits collection.

KOLB-PROUST ARCHIVE FOR RESEARCH

Total names found in the Kolb-Proust dataset = 5,727 Links were found for 1,953 people	**Count of URIs Found**
having VIAF links	1,678 (29%)
having French Wikipedia links	1,236 (22%)
having English Wikipedia links	999 (17%)
having other links	264 (5%)

Table 9: Results of the Kolb-Proust Archive Research name reconciliation work

<Person>s mentioned in the Kolb-Proust Archive for Research (KPA) metadata include family members, friends, journalists, and others with whom Marcel Proust corresponded or were mentioned in diaries and letters, not public figures. For this reason, we anticipated that reconciliation using VIAF and even Wikipedia would be challenging. The automated reconciliation process confirmed this to be the case. Of 5,727 unique names found in KPA metadata, only 1,678 (29%) matched in VIAF. Again manual searching (though not as exhaustive as was done for Motely) helped, but still only 999 (17%) matches were found in English Wikipedia, and only 1,236 (22%) matches found in French Wikipedia. Reconciliation results for KPA legacy metadata are shown in table 9. Time and resource limits were reached before we had exhausted all search options for finding matches for KPA names, but any improvement in these numbers would have been at best incremental.

As noted in the first part of this whitepaper, he Kolb-Proust collection also includes many bibliographic citations to articles in journals and newspapers and to book and short fiction titles. Among 13,923 citations found from the 8,716 research notecards processed, we were able through semi-automated processes to

link 4,812 (35%) to digitized full text (Table 10). To find these citations we searched at the level of publication title in the Bibliothèque nationale de France's Digital Library (BnF-Gallica). Since citations in the KPA metadata reference specific newspapers and journal issues by date and/or volume and issue, we reviewed the data structures used in BnF-Gallica to see whether we could generate from the title level a link to a specific issue mentioned in a citation. This proved challenging because link patterns in BnF-Gallica tend to be title specific. Experimentation with frequently cited publications (e.g., *Le Figaro*) established that this was feasible given enough time and resources; however, these links were never implemented in the production interface. As best we were able to tell, linking to the page level does not appear to be feasible given current practices used by BnF-Gallica for minting URLs.

Total number of notecards in the Kolb-Proust dataset = 8,716	**Count of URIs Found**
Citations found on notecards	13,923 (~1.6 citations/card)
Links founds for citations	4,812 (35%)

Table 10: Results of the Kolb-Proust Archive Research citation reconciliation work

```
{
"@context": "http://schema.org",
"id":
"http://catalogdata.library.illinois.edu/lod/entities/Persons/kp
/dreyfu1",
"sameAs": ["https://viaf.org/viaf/97781547",
"https://fr.wikipedia.org/wiki/Robert_Dreyfus_(écrivain)"],
"name": "Dreyfus, Robert",
"type": "Person",
"familyName": "Dreyfus",
"givenName": "Robert",
"birthDate": "1873",
"deathDate": "1937",
"gender": "Male",
"description": "journaliste",
"jobTitle": "journalist",
"parent":
"http://catalogdata.library.illinois.edu/lod/entities/Persons/kp
/dreyfu12",
"sibling":
"http://catalogdata.library.illinois.edu/lod/entities/Persons/kp
/dreyfu11"
}
```

Figure 7: Kolb-Proust collection name represented in JSON-LD

For the KPA collection, we generated RDF graphs not only for each of Kolb's notecards, but also separately for each <Person> entity mentioned in KPA legacy metadata (see figure 7). This was appropriate given that we started with a separate local names authority database containing information about individuals beyond what is contained on the original notecards, given that we found external URLs (to which we could simply link for more information) for less than one-third of the <Person>s mentioned in KPA legacy metadata, and given that many names appear on tens or even hundreds of separate notecards (making if more efficient to link to <Person> descriptions rather than repeatedly embedding in each <Dataset> or

<CreativeWork> a subgraph for <Person> entities referenced or mentioned). As a result all <Person>s mentioned in KPA legacy metadata were assigned a URL on our servers. We include <sameAs> triples to link our <Person> graphs to VIAF and/or Wikipedia graphs for the roughly one-third of KPA <Person>s that are included in these resources, but for the remaining two-thirds of the individuals mentioned in the KPA legacy metadata, no other identifier is provided. Our graphs do contain unique information of broader possible interest, and this information is openly available to any interested. But beyond the linkages to our local <Person> descriptions from the KPA notecards (<Dataset>s) that we have transformed into RDF, it is not clear is how best to advertise availability of this information beyond our local context.

OBSERVATIONS

1. HOW BEST TO MANAGE LOCAL AUTHORITY DATA AS LINKED DATA?

Many individuals mentioned in digitized special collection legacy metadata are not authors of books, nor are they public figures whose names are likely to be found in library name authority files or in Wikipedia. In addition, not all sources we had to consult for reconciliation tasks support linked data as of yet; the more domain-specific a resource, the less chance it is linked data conformant. No single resource includes every name. Legacy special collection metadata contains information not currently linked from or even known about by large, centralized linked data compatible authorities. These facts represents a current (albeit hopefully short-term) limitation of linked data (and its potential benefits to users) in the context of library special collections, and highlight the still significant challenges of leveraging the distributed information associated with library special collections for the maximum benefit of users.

2. MANUAL RECONCILIATION WORK IS IMPORTANT FOR SPECIAL COLLECTIONS

Almost by definition, library special collection metadata includes unique (i.e., specialized) information. Legacy special collection metadata often references entities that from a machine processing standpoint may be insufficiently identified, i.e., entity descriptions which while sufficient in the context of the special collection may be imprecise or ambiguous in a broader context. For these reasons, reconciliation workflows need to include some manual processing as well as automated, algorithmically-based processing. In our manual reconciliation process, we found that consulting some resources (e.g., WorldCat Identities, Theatricalia, IMDb, IBDb, etc.) worked better for us than some other resources; however, we anticipate that this will vary according to the domain scope of the special collection.

3. METADATA CLEANUP IMPROVES RECALL

Name reconciliation results can be significantly improved when the metadata is clean and consistent. The automated process can be only effective when the metadata is clean. For our project, we it proved practical to hire a graduate student to help cleanup, enhance, remediate, and find linked data sources for entities included in the collections studied for this project. For Motley and Portraits the student worked 10-12 hours per week for six month, totaling about 240 hours. The student significantly improved the metadata quality and consistency, and successfully de-conflated legacy metadata values. Manual process affected the reconciliation work significantly, reinforcing the suggestion both manual and automated processes are valuable in special collections legacy metadata remediation and reconciliation work.

Appendix 1: Local file names used for the Motley Collection of Theatre and Costume Design

Field name	DC map	Note	Controlled vocabulary
Image Title	title	Title of image	No
Dimensions	extent	Size of the physical item	No
Associated People	description	Add the qualifiers after the name, usually director, producer and actor's names are available	Yes (LC NAF)
Inventory Number	none	Inventory number used locally	No
Description	description	Additional information about the item or play	No
Inscriptions	description	Information inscribed on item	No
Repository	source	Holding library where the physical item is housed	Yes (LC NAF)
Collection	isPartOf	Collection title	No (one default value)
Author/Composer	contributor	Creator of the play or opera	Yes (LC NAF)
Production Notes	relation	Additional information of the performance (URL)	No
Performance Title	references	Title of the performance	Yes (LC NAF)
Theater	description	Name of the theater where the performance was held	Yes (LC NAF)
Opening Performance Date	date	Performance date	No (ISO 8601)
Materials/Techniques	medium	Materials/technique used for the item	Yes (TGM II)
Object	format	Describe the genre of the item	Yes (AAT)
Type	type	Type f the item	Yes (DCMI Type)
Subject I (AAT)	subject	Descriptions of costume and furnishings as well as concepts and style	Yes (AAT)
Subject II (TGM I)	subject	Descriptions of physical characteristics of the item	Yes (TGM I)

Appendix 2: Local file names used for the Portraits of Actors

Field name	DC map	Note	Controlled vocabulary
ID Number	identifier	Alpha-numeric code based on name of actor	not necessary- there is no duplication
Title	title	"Portrait of [name of actor]" or "Name of Actor as [role] in ["Play"]" or "Name of Actor 1 as [role] and Name of Actor 2 as [role] in a scene from ["Play"]	no
Date	created	4-digit year print was made, if known.	not necessary
Role	description	Controlled list of role names	local
Play	description	Controlled list of play titles (short titles, not the long titles many of these 18th century plays have)	local
Subject	subject	Name of actor from LC NAF- For those not in NAF, a name authority is created and recorded on the spreadsheet **Actors_portraits_data.xls**; some of these may not be detailed enough. Other LCSH subject headings, including headings that have general date ranges corresponding to when the actor was working [ie. eighteenth century] LCSH was chosen rather than Thesaurus for Graphic Materials because it had terms like "theatrical manager" and "breeches parts" that seemed necessary to describing this collection. The term "costume" was used whenever the actor is depicted in a role. It was difficult to decide whether to include a subject heading like "blackface entertainer" when a particular portrait was not in blackface, but in general we did so.	LCSH NAF
Type	type	Type of print, photograph, etc. *photomechanical prints* was used as a fairly catch-all category	AAT
Dimensions	extent	Indicate whether the measurement is: image sheet mounted sheet plate marks …or whatever it says on card, ie.: image 3 x 2 ½ inches sheet 5 x 5 ¼ inches	

Technique	medium	Artistic/technical technique(s) used to create the print	AAT
Creator	creator	Name of artist whose painting/drawing the print was based on; name of printmaker; name of photographer	No – hard to control. Used the CV function of CONTENTdm just to catch typos
Publisher	publisher	Name of publisher; other corporate name responsible for making print available (i.e. name of lithography firm) – when difficult to distinguish between creator/publisher, an attempt was made to use publisher for corporate names, but there is some lack of consistency with this. Another issue: usually the publisher is an individual (i.e. John Bell) – and we used the name J. Bell rather than the title of the publication (*Bell's British Theatre*)	no
Description	description	Free-text description, including some details from costume, scenery. Including whether portrait is whole-length, half-length, bust, etc., for costume researchers who may only want certain portrait types. (I began to include the word "portrait" in the description because terms like "whole-length" seemed a little vague. This isn't consistent throughout the collection, ideally, it would be.)	no
Rights	rights	Not sure	blanket statement
Digital Collection	isPartOf	Portraits of Actors, 1720-1920 University of Illinois Digital Collections	blanket statement
Repository Collection	source	University of Illinois Theatrical Print Collection, #35. (It seems important to tie the actors portraits to the larger physical collection from which they were derived. Would be good to link to the rbml/archon finding aid when it is given a stable url.)	blanket statement
Digital File Creation	none	[administrative metadata about how scanning and conversion was done]	blanket statement
File Name	identifier	Name of the image file	