JUNE 15, 2015
*(Edited for the Web dissemination on November 9, 2015)*

## PROPOSAL NARRATIVE

# EXPLORING THE BENEFITS FOR USERS OF LINKED OPEN DATA FOR DIGITIZED SPECIAL COLLECTIONS

*Submitted to the*
**SCHOLARLY COMMUNICATIONS PROGRAM**
**OF THE ANDREW W. MELLON FOUNDATION**

**Principal Investigator:** **Timothy W. Cole**
Professor
CIRSS Coordinator for Library Applications
Graduate School of Library and Information Science
Mathematics Librarian
University Library
University of Illinois at Urbana-Champaign

**Co-Principal Investigators:** **Myung-Ja Han**
Associate Professor
Metadata Librarian
University Library
University of Illinois at Urbana-Champaign

**Caroline Szylowicz**
Associate Professor
Kolb-Proust Librarian
Curator of Rare Books and Manuscripts
University Library
University of Illinois at Urbana-Champaign

# EXPLORING THE BENEFITS FOR USERS OF LINKED OPEN DATA FOR DIGITIZED SPECIAL COLLECTIONS

## 1. Executive Summary

Tangible special collections of primary sources have long been central to humanities research. At times there is still no substitute for physical access to a primary source,[1] but scholar interest in digital resources is growing. Today digitized special collections play a major role in humanities scholarship and pedagogy. Digital collections facilitate the initial exploration, discovery and disambiguation of sources. Well-connected digital collections can help satisfy the need for contextual mass,[2] enable complex connective research, and provide a powerful way to collate and contextualize physically dispersed primary sources. Given the core mission of libraries to facilitate the discovery and use of resources that support scholarship, high priority has been given in the last 20 years to the digitization of special collections. A question naturally follows: After digitization, what more needs to be done to maximize the usefulness of these digitized resources?

The relatively modest levels of use that many digitized special collections get and the low share of this use attributable to faculty and students suggest that more does need to be done post-digitization. There are multiple factors, of course, but in large part the full potential utility of digitized special collections has not yet been realized because digitized special collection resources, though accessible via the World Wide Web, are not woven into the fabric of the Web, and especially are not integrated much at all into the emerging and increasingly important data-centric subset of the Web known as the Semantic Web.[3] Digitized special collections are *on* the Web, but not *part of* the Web, at least not to the degree that they could be.

Transforming legacy special collections item-level metadata into Linked Open Data (LOD) and integrating LOD into services and end-user interfaces will help address this problem. This is not a new or unique insight, but within the library community the paradigm shift to LOD is proving difficult, both technically and socially. Library experience with LOD, especially LOD for special collections, is limited. Best practices for transforming legacy metadata into LOD are still developing, and the hypothesized benefits of LOD for our users remain to be demonstrated. As a result libraries are hesitant to take on this task without outside assistance. Incentivizing the transition to LOD for digitized special collections is especially challenging given the diversity of descriptive practices and sophisticated user requirements in this domain. Further experimentation and proofs-of-concept are needed to establish the costs of transforming legacy special collections metadata into LOD and to demonstrate the near-term benefits of doing so.

We propose a 20-month project conducted collaboratively by the University Library and the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign to further our understanding of four translational research questions:

[1] For example, the *Motley Collection of Theatre and Costume Design* includes fabric swatches; while these can be imaged, to fully appreciate their texture and weight there is no substitute for physical access. Similarly, while digital facsimiles of Marcel Proust's letters give intellectual access to content, physical access helps a scholar judge paper weight and quality, which in turn is indicative of the esteem in which Proust held the recipient. But even in these scenarios, digital surrogates can make travel to remote collections more efficient and help avoid unnecessary trips.

[2] See section 2.1 and Palmer et al. (2010) for further discussions of *contextual mass*.

[3] The distinction between the World Wide Web and the emerging Semantic Web is described in subsection 2.2.

1. As compared to general collection catalog records, item-level metadata for digitized special collections are frequently more granular, richer in non-bibliographic entities, and expressed using custom vocabularies and schemas. What differences and additional challenges are encountered when transforming legacy special collections metadata records into LOD?
2. Typically interfaces used to discover and view digitized special collections are disconnected from the online public access catalogs and ancillary services used to provide user access to general library collections. Can LOD reconnect library special and general collections?
3. Digitized special collections are also disconnected from external, non-library information resources on the Web. How can LOD be leveraged to help identify and establish useful connections to these resources, and do non-library sources have the potential to enrich item descriptions and provide context for discovering and interpreting digitized special collections?
4. Often descriptions of special collection items include extensive references to people and relationships. Can emerging visualization and annotation technologies add a social network view of a special collection that usefully complements traditional bibliocentric perspectives?

We propose to investigate these four questions and demonstrate findings concretely by transforming legacy string-based item-level metadata and then experimenting with user services for three modestly sized digitized special collections hosted by the University of Illinois – the *Motley Collection of Costume and Theatre Design*, the *Portraits of Actors, 1720 – 1920 Collection*, and the *Kolb-Proust Archive for Research*.[4] The first two collections are typical of theatre-themed image special collections hosted in CONTENTdm or similar content management systems. While loosely based on Dublin Core (DC), the metadata schemas used for these digitized collections have been customized and extended to express attributes and types germane to such image collections. The Proust Archive metadata, on the other hand, are expressed using a profile of the Text Encoding Initiative (TEI) schema and provide context for Proust's letters, literary works and relationships. The metadata for all three collections are rich in person, place and event entities, but these contrasts in descriptive model and collection content will allow us to highlight findings that have applicability beyond a single metadata schema or collection type. Additionally, working with three collections will help us differentiate between collection-specific and generic remediation and transformation requirements. Finally, because the Proust Archive metadata are especially rich in information about Proust's social relationships, they will provide good fodder for question 4 above.

Our goal is to provide evidence helpful to understanding these research questions and gain experience with these issues, demonstrate potential benefits of LOD, and learn more about the resources required to transform and utilize LOD, both as a way to inform transformation best practices and as a means to add to a collective assessment of the likely benefits of LOD for library users. In undertaking our work we will take advantage of related past and ongoing research into the use of LOD across all kinds of library collections. This includes our own experience with *Emblematica Online* and in transforming a MARC-based snapshot of our library catalog into LOD, work that has been done by OCLC Research, the efforts of the World Wide Web Consortium (W3C) schema.org Community Group, and the research being conducted by the Linked Data for Libraries (LD4L) and the proposed Linked Data for Production cataloging (LD4P) projects. [5]

[4] These collections and their primary intended audiences are described in Section 3.

[5] See section 6 for further discussion of where our proposed project fits relative to other projects and initiatives.

THIS PAGE INTENTIONALLY LEFT BLANK

# Table of Contents

# 2. Making Web-Accessible Digitized Special Collections More Useful

## 2.1. The challenge – better connectedness

The primary sources found in research library and archives special collections support a broad range of advanced humanities scholarship and pedagogy and can be thought of as the basic data of the humanities (Brockman et al., 2001). Palmer, Teffeau and Pirmann (2009, p. 4) found that digitization does not lessen the intrinsic value of these special collections. "What has changed in the digital environment is not the value of these kinds of sources but rather how they are searched, accessed and used in the scholarly process." Scholars rely on digitized special collections for identifying and evaluating collections that might be worth visiting (Tibbo, 2003). Increasingly they also rely on these digital collections for finding and accessing pieces of evidence directly, either for analysis and interpretation or for use in publications (Palmer, 2005; Ciula & Lopez, 2009). Digitized special collections now have demonstrated value as pools of evidence for various aspects of scholarly work. Green and Courtney (2015) found a growing interest on the part of humanists in using digitized special collections in ongoing research and/or in teaching. Rogers (2008) goes further, suggesting that while digitized special collections, like physical special collections, begin by offering a space (albeit a virtual space) to explore evidence necessary for historical scholarship, they also have the potential to facilitate new more complex, connective work by historians. This growing receptivity to digitized special collections has not gone unnoticed by libraries.

In 2012 the *Association of Research Libraries (ARL)* and *Ithaka S+R*, surveyed ARL library directors and staff regarding digitized special collections. Key findings from this survey confirm the value placed on digitized special collections (Maron and Pickle 2013, p. 2):

- "Most library leaders feel that digitized special collections are critical to the libraries' future, but few feel their institutions' investments in updates and upgrades are sufficient."
- "Although the ability to offer greater access emerged as a key motivator for digitizing collections, investments in understanding the needs of the audience are quite low."
- "Libraries are spending far more to create new resources than they are on maintaining and enhancing the ones they have already created…. This suggests a scenario where digitized collections, once created, are intended to essentially run without much active management, a situation that could ultimately hamper the ability of these institutions to sustain their projects and achieve the impact they desire."

These findings suggest that along with a consensus that digitization is a priority there is a shared anxiety about what needs to be done post-digitization. The last finding in particular recognizes the limitations (to abuse a cliché) of the 'digitize and they will come' assumption. Digitization of library special collections has proven a good first step in making unique collections and archival materials more visible and available to scholars, but alone this step is not enough. Too often there is a disconnect between anticipated or intended use of digitized special collections and actual use, both in terms of who is using these resources and for what they are using them. Michele Reilly and Santi Thompson surveyed users of the University of Houston Digital Library (UHDL). The UHDL, built on top of CONTENTdm, holds primarily digitized special collections images and "was created to be used by academics and researchers at the University of Houston (UH) and around the world" (Reilly and Thompson 2014, p. 197). Yet, the survey found that 65% of the users of the UHDL identified themselves as non-UH visitors and only 13% identified themselves as UH faculty or student. In terms of use, less than 30% was described as fitting into a scholarly or pedagogical category. Rather many respondents simply wanted an image with which to decorate home or office or wanted an image as remembrances of important life moments. Not to disparage these use cases, but this was not what the University of Houston Library had in mind when the decision was made to digitize these special collections.

We believe that two of the main reasons for this disconnect between intended and actual use are poor discoverability and insufficient connectedness to support *contextual mass* – both of which stem from a tendency to isolate digitized special collections in the Web equivalent of silos. Because digitized special collections are presented on the World Wide Web in isolation from other relevant information resources, both scholarly (e.g., the library's general collections) and popular culture (e.g., Wikipedia), discovery is impeded. Connections to external resources that could facilitate serendipity are absent. Metadata schema and authorities are string-based and often idiosyncratic to each collection, content management system and/or institution. Such descriptions are less useful to Web search engines than are descriptions that favor links over strings and rely on community-based semantics.

Additionally, studies of the information practices of humanities scholars also suggest that collections are most effective for facilitating research and pedagogy when they exhibit contextual mass (Palmer et al., 2010). Contextual mass is achieved by collating items that work together as a system of sources, with enough meaningful interrelationships between materials of different kinds and on different topics, to support research inquiry. Rather than being driven only by the obtainment of critical mass, digital collecting in support of scholarship is driven by criteria that produce dense, rich, and cohesive groupings of sources for research. Special collections hold valuable evidence for research in the humanities, but on their own may fail to obtain contextual mass, as they are circumscribed by institutional and technical boundaries. Connecting items in these collections with related sources beyond the immediate local collection supports the infusion of contextual mass into digital special collections by affording the expression of different kinds of relationships between sources of evidence and other entities. Systematic connections make possible the collation of related sources, regardless of origin, so that scholars and collection developers can create virtual groupings of evidence that have contextual mass. Ideally, contextual mass is achieved by each scholar as she assembles an aggregation of primary sources sufficient to satisfy the needs of her specific scholarly inquiry. We can anticipate that researcher-defined thematic collections "will play an important role in how research materials are reconfigured in the digital environment, as libraries become more involved in providing access to digital resources collected and organized by scholars, who contribute important expertise in selection, collocation, interpretation, and integration of the sources they study" (Palmer et al., 2010, p. 3). Essential to achieving this vision is the connectedness of items across a range of digitized general and special collections, spanning institutional boundaries and integrated with non-library sources. Libraries need to seize this opportunity. Better linking of digitized special collections and items one to another and to external, non-library information resources on the data-centric Semantic Web will significantly improve both the discoverability and utility of these digital resources.

## 2.2. The emergence of a data-centric Web

Part of the current silo effect has come about because of the inherent document-centric nature of the legacy World Wide Web. HTML (HyperText Markup Language) and HTTP (HyperText Transfer Protocol), both created in 1989 by Tim Berners-Lee, enabled the creation of a document-centric World Wide Web. This model has proven powerful and durable, but it is limited, especially when it comes to knowledge management. Documents and document-like objects (e.g., scanned images) on the traditional World Wide Web are (for computational purposes) largely opaque and indivisible, and the links between objects (e.g., between Web pages, scanned images, etc.), when present in the object itself or in the metadata describing the object, are undifferentiated. On the original document-centric World Wide Web, a librarian could link together two resources, but there was no standard way to express the nature of the link, why the link was created or which facet or element of object A related to which facet or element of object B. It was not possible, for example, to express in a standard, machine recognizable way that two objects were connected because both are related to the same individual, event, place or bibliographic item. This in turn meant that there was no efficient standards-based way to task computers to seek out and identify new linkages between objects on the legacy document-centric World Wide Web.

Over the last decade, the emergence of best practice guidelines for Linked Open Data (LOD),[6] in combination with the increasing maturity of RDF (the Resource Description Framework, introduced in 1998) and other closely-related protocols, have enabled the creation of the *Semantic Web*, a next generation enhancement of the World Wide Web that is data-centric rather than simply document-centric (*data* is used here in its broadest sense). Documents and document-like objects are still to be found on the Semantic Web, of course, but through the use of RDF, the way they are connected is different. RDF supports the use of properties and classes, which in turn support standard, precise differentiation of objects and the links between them. Coupled with agreements on how to reference and talk about non-Web entities (people, places, intellectual works, events), a broader range of linkage types and granularity can be expressed, ultimately making it possible for scholars to conduct their research in the digital environment more efficiently and to better extend traditional research methods and paradigms. Librarians, aggregators and ultimately users and software can now discover and connect resources on the basis of their LOD-described relationships. Using an RDF-compatible ontology, links can now be differentiated as to their role and entities within or connected to Web resources can be labeled and differentiated – e.g., as a person, a place, an event, a bibliographic entity. Links to related resources can be followed and resources can be collated based on their linkages, better supporting the creation of scholar-driven thematic collections that achieve contextual mass.

The implications and affordances of RDF and LOD for digital humanities scholarship are still being explored and need to be assessed discipline by discipline, but these new technologies have the potential to help overcome previous limitations of scale and the distributed nature of the Web to allow scholars to answer new questions using virtual thematic collections and work sets of items gathered together from previously disjoint, dispersed collections and leveraging the LOD-described relationships that link digitized

[6] Berners-Lee. (2006/2009.) *Linked Data*. Accessed 5/10/15: http://www.w3.org/DesignIssues/LinkedData.html

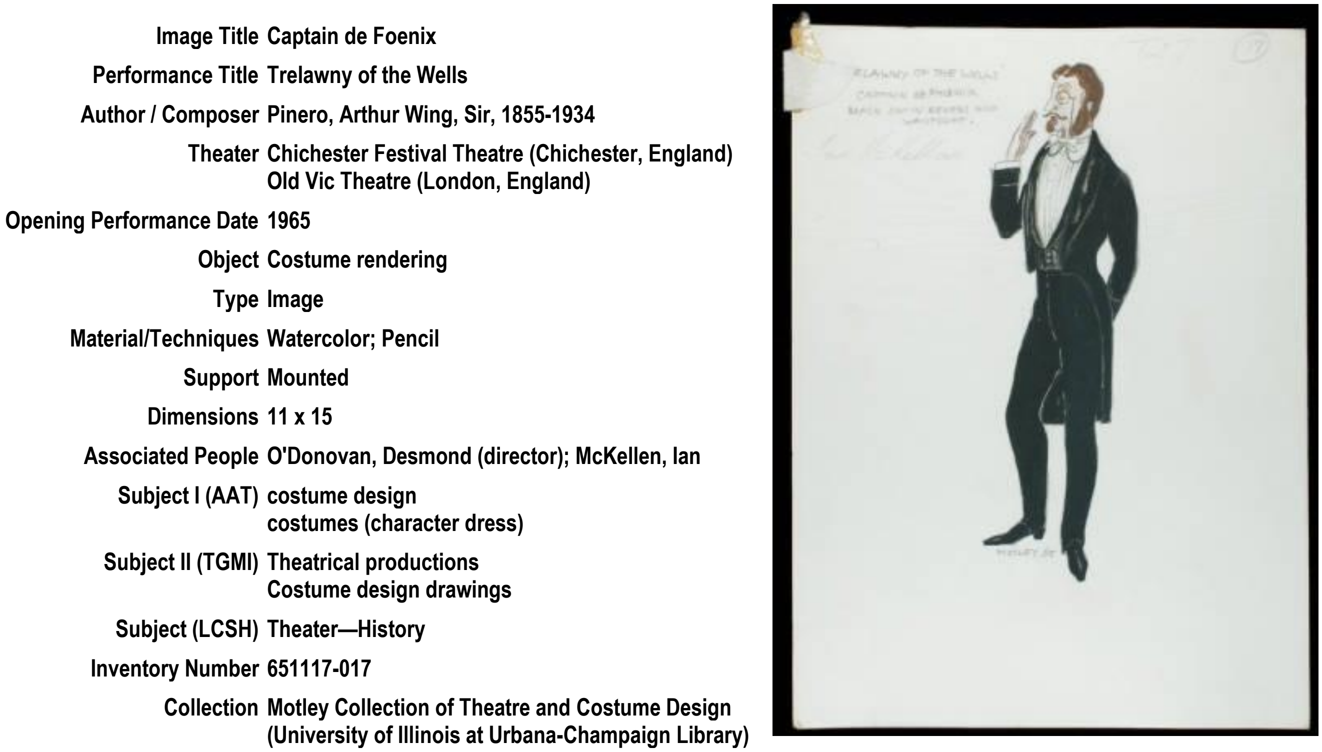
special collection items on a variety of shared attributes. Ultimately, as Stefan Gradmann has argued, by taking advantage of LOD, RDF and other technologies of the Semantic Web, libraries have opportunities to help scholars interact with content and its context in innovative ways. “Libraries are particularly apt for this role because of their traditional co-operational discipline in metadata generation and organisation and also because of their excellent contextualisation data (authority files of all kinds) which they are used to apply[ing] for enriching object descriptions with contextual links” (Gradmann 2015, p. 255). Gradmann identifies three burgeoning areas of scholarly activity that libraries will be able to better support with increasing adoption of LOD: (1) semantic abstracting and named entity recognition to enable “strategic reading” (Renear and Palmer, 2009) across collections; (2) using links to context about sources as a basis for generating new knowledge; and (3) inferring from and reasoning over linkages to stimulate deeper understanding.

## 2.3. Task 1: Transforming special collections metadata to LOD

Libraries routinely refine and improve the practices and methods used to describe and manage curated resources. As libraries look today to update the methods used to describe digitized special collections items, the difference this time is the broader technical context represented by the data-centric Semantic Web (i.e., the availability of LOD best practices, RDF and schema.org to serve as our starting point) and the opportunity to link to relevant, non-library information resources available on the Semantic Web. The first step and the pre-requisite for making our digitized special collections part of the Semantic Web (and therefore better connected to both library and non-library information resources) is to transform existing item-level descriptive metadata into LOD.

While serial titles and individual books held in library collections are typically described by Machine-Readable Catalog (MARC) records, non-book, special collection resources held by libraries, e.g., archival materials, items in collections of images, manuscripts, letters, etc. are described using non-MARC descriptive schemas and at varying levels of granularity (e.g., in the case of many archives, at the folder, box or collection level rather than at the level of individual items). Indicative of this is the relatively low proportion of special collection items represented in library online catalogs (most of which ingest and index only MARC records). For example, Jackie Dooley and Katherine Luce (2015, p. 7) report that only 25% of visual materials curated by libraries are visible in online catalogs. There are a variety of reason why MARC is not used for describing special collections resources. The creation of MARC records is labor intensive and requires specialized knowledge, especially when describing special format items such as maps and images; using MARC to describe all the images in a large collection quickly becomes impractical. MARC is not easily extensible and its expressiveness is limited. MARC does not express whole-part relationships well, which can be important when describing archival relationships.

In lieu of MARC record descriptions, digitized special collections and special collections items are described according to other, often custom metadata models relying on domain-specific or locally defined metadata semantics. A variety of metadata schemas are used by libraries to describe digitized special collections. At Illinois we use locally customized, collection-specific extensions of Dublin Core (DC) to describe our digitized image collections. Figure 1 illustrates how images in the *Motley Collection* are described. As demonstrated by this example, the departure from the DC standard is substantial to the point that the metadata is not recognizable as DC. (A mapping to Qualified DC is maintained within CONTENTdm, but it does not encompass all metadata elements used locally and is lossy, i.e., not fully reversible.) The descriptions in our Kolb-Proust Archive are expressed using the Text Encoding Initiative (TEI) schema. Archival special collections (beyond the scope of this project) are often described using the Encoded Archival Description (EAD) schema in order to preserve whole-part and hierarchical relationships.



| | |
|---|---|
| Image Title | Captain de Foenix |
| Performance Title | Trelawny of the Wells |
| Author / Composer | Pinero, Arthur Wing, Sir, 1855-1934 |
| Theater | Chichester Festival Theatre (Chichester, England)<br>Old Vic Theatre (London, England) |
| Opening Performance Date | 1965 |
| Object | Costume rendering |
| Type | Image |
| Material/Techniques | Watercolor; Pencil |
| Support | Mounted |
| Dimensions | 11 x 15 |
| Associated People | O'Donovan, Desmond (director); McKellen, Ian |
| Subject I (AAT) | costume design<br>costumes (character dress) |
| Subject II (TGMI) | Theatrical productions<br>Costume design drawings |
| Subject (LCSH) | Theater—History |
| Inventory Number | 651117-017 |
| Collection | Motley Collection of Theatre and Costume Design<br>(University of Illinois at Urbana-Champaign Library) |

**Figure 1**: Metadata describing a costume design included in the *Motley Collection*

Though all of these approaches differ substantially from MARC in many respects, they all hew relatively closely, like MARC, to pre-digital traditions of library descriptive cataloging and bibliographic control. Titles, author names, subject headings, etc. contained in existing metadata records are expressed as strings (with varying degrees of consistency and exactitude). But the language of the World Wide Web in general is based on HTTP uniform resource identifiers (URIs), not strings, and the descriptive syntax of the Semantic Web is RDF in conjunction with well-defined semantic labels like those defined at schema.org. Early work at Illinois[7] and elsewhere creating RDF from MARC records and applying the principles of LOD in transforming these general collection catalog records, suggests that the use of URIs as identifiers for the bibliographic, name, place and event entities commonly found in library descriptive records results in more reliable (i.e., less ambiguous), more consistent, and more interoperable entity identification. In order to explore the efficacies and affordances of LOD for non-MARC digitized special collection metadata, our first challenge

[7] See: http://catalogdata.library.illinois.edu/

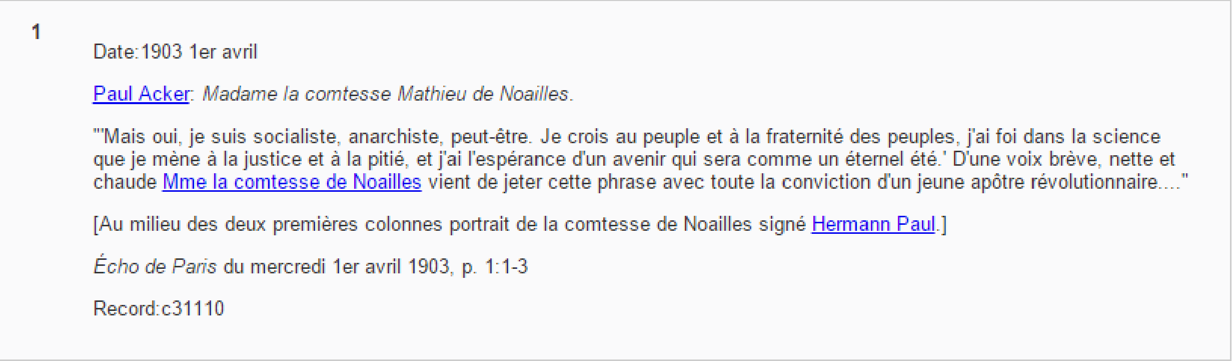
will be to map and transform our custom legacy metadata into RDF and schema.org, replacing strings with URIs where possible, thereby linking our resources and resource descriptions to other resources, authorities, and services on the Semantic Web.

```
[…]
 a schema:CreativeWork ;
 schema:name "Trelawny of the Wells" ;
 schema:author <http://viaf.org/viaf/9975067> ;
 schema:sameAs <http://worldcat.org/entity/work/id/2113881> .
[…]
 a schema:VisualArtwork ;
 schema:name "Captain de Foenix" ;
 schema:sameAs <http://www.worldcat.org/oclc/902642810> ;
 schema:url <http://imagesearchnew.library.illinois.edu/cdm/ref/collection/motley/id/8423> ;
 schema:description "Costume rendering" ;
 schema:material "Watercolor", "Pencil" ;
 schema:surface "Mounted" ;
 schema:width "11" ;
 schema:height "15" ;
 schema:about <http://id.loc.gov/vocabulary/graphicMaterials/tgm010747>,
               <http://id.loc.gov/vocabulary/graphicMaterials/tgm002607>,
               <http://id.loc.gov/authorities/subjects/sh85134531>,
               <http://vocab.getty.edu/aat/300163423>,
   <http://vocab.getty.edu/aat/300266810>;
 schema:mentions [
  a schema:TheaterEvent ;
  schema:location <http://dbpedia.org/page/Chichester_Festival_Theatre> ;
  schema:startDate "1965" ;
  schema:organizer "O'Donovan, Desmond" ;
  schema:performer <http://dbpedia.org/page/Ian_McKellen> ;
  schema:workPerformed <http://worldcat.org/entity/work/id/2113881>
  ], [
  a schema:TheaterEvent ;
  schema:location <http://viaf.org/viaf/140548301> ;
  schema:startDate "1965" ;
  schema:organizer "O'Donovan, Desmond" ;
  schema:performer <http://dbpedia.org/page/Ian_McKellen> ;
  schema:workPerformed <http://worldcat.org/entity/work/id/2113881>
  ] ;
 schema:copyrightHolder <http://id.loc.gov/authorities/names/no2006022679> ;
 schema:provider <http://id.loc.gov/authorities/names/no2006022679> ;
 schema:creator <http://id.loc.gov/authorities/names/n50006654>;
 schema:isPartOf <http://imagesearchnew.library.illinois.edu/cdm/landingpage/collection/motley> .
```

Figure 2: the metadata in Figure 1 transformed into LOD.

As a concrete illustration of this task, Figure 1 shows metadata for an image from the *Motley Collection of Theatre and Costume Design* (this collection is described further in subsection 3.1 below). Figure 2 shows a preliminary, hand-crafted transformation of the metadata in Figure 1 into LOD using schema.org semantics. The content access system for the Motley Collection is CONTENTdm. The metadata is entirely string based. None of the metadata fields contain links to external resources elsewhere on the Web, although links to collection-specific information and to facilitate searching for similar items within the Motley Collection are available. Having discovered this resource through our CONTENTdm service, a user can further explore the Motley Collection, but she is not connected to external library or non-library resources. In this sense the items in the Motley Collection appear as Web dead-ends.

The string values in Figure 1 for the author, theater, collection, at least one of the associated people and Art & Architecture Thesaurus (AAT), Library of Congress Thesaurus for Graphic Materials (TGMI) and Library of Congress Subject Headings (LCSH) subjects can all be replaced with URIs. An explicit URI can also be added for the work performed, which is represented in the current metadata record by the performance title. Figure 2 gives a sense of how the resulting LOD graph, RDF serialized in Turtle, might appear. (Note how the single metadata record depicted in Figure 1 conflated, from an LOD perspective, a VisualArtwork, the play itself as a CreativeWork, and two TheaterEvents.) Similarly the *Kolb-Proust Archive* record in Figure 3, might look as shown in Figure 4 once transformed.



1

Date: 1903 1er avril

Paul Acker: *Madame la comtesse Mathieu de Noailles*.

"'Mais oui, je suis socialiste, anarchiste, peut-être. Je crois au peuple et à la fraternité des peuples, j'ai foi dans la science que je mène à la justice et à la pitié, et j'ai l'espérance d'un avenir qui sera comme un éternel été.' D'une voix brève, nette et chaude Mme la comtesse de Noailles vient de jeter cette phrase avec toute la conviction d'un jeune apôtre révolutionnaire...."

[Au milieu des deux premières colonnes portrait de la comtesse de Noailles signé Hermann Paul.]

*Écho de Paris* du mercredi 1er avril 1903, p. 1:1-3

Record: c31110

Figure 3: A record from the *Kolb-Proust Archive for Research*

```
@prefix schema: <http://schema.org/> .

[...]
 a schema:Article ;
 schema:name "Madame la comtesse Mathieu de Noailles" ;
 schema:author <http://viaf.org/viaf/71420451> ;
 schema:mentions <http://viaf.org/viaf/95303465>, <http://viaf.org/viaf/100903063> ;
 schema:pageStart "1" ;
 schema:pageEnd "1" ;
 schema:description "'Mais oui, je suis socialiste, anarchiste, peut-être. Je crois au peuple et
        à la fraternité des peuples, j'ai foi dans la science que je mène à la justice et à la
        pitié, et j'ai l'espérance d'un avenir qui sera comme un éternel été.' D'une voix brève,
        nette et chaude Mme la comtesse de Noailles vient de jeter cette phrase avec toute la
        conviction d'un jeune apôtre révolutionnaire....", "[Au milieu des deux premières colonnes portrait de la comtesse de Noailles
        signé Hermann Paul.]" ;
 schema:isPartOf [
  a schema:CreativeWork ;
  schema:name "Écho de Paris" ;
  schema:url <http://gallica.bnf.fr/ark:/12148/bpt6k813488q.langEN> ;
   schema:datePublished "1903-04-01"
 ] .
```

Figure 4: the metadata in Figure 3 transformed into LOD

Doing these transformations algorithmically will be challenging. We will begin by adapting and then applying across all three collections the mappings and algorithmic methods we developed for finding links for and transforming 5 million library catalog records into LOD.

Although the special collection metadata schema semantics are non-MARC, this work is still a good starting point given the bibliographic character of legacy special collections metadata, and given that target semantics for both transformations is the same, i.e., schema.org. The process of adapting our MARC to schema.org RDF transformations also will be informed by RDF-compatible schema analyses, comparisons and modeling work emerging from the Workset Creation for Scholarly Analysis project (Nurmikko-Fuller et al., 2015).

Subsequent analysis of our initial transformation pass will help identify collection-specific mappings and entity reconciliation requirements. For example, we can anticipate that many of the individuals mentioned in the Kolb-Proust metadata records are not in the Virtual International Authority File (VIAF), i.e., are not authors, actors or artists. A preliminary analysis indicates only about one-third are in VIAF. So we will need to consult additional resources to help with reconciliation – e.g., the Léonore database (Légion d'honneur)[8] maintained by the French National Archives and the Mémoire des hommes[9] database of World War I documentation maintained by the French Department of Defense. This will give us more information (e.g., dates) about and possible identifiers for individuals. With our generic link identification and transformation workflow augmented by collection-specific remediation, identifier sources, and transformation rules, we will then make a second transformation pass. For the experimentation outlined in subsections 2.4 – 2.6, we will rely on the output of this second transformation pass.

Because our special collection metadata are more diverse and variable in structure, formatting of strings and semantics than the MARC records that we have transformed in prior LOD experimentation, there of course will be a number of new challenges and issues that will need to be addressed in creating our algorithmic transformation workflows. Not all of these issues are foreseeable, but we can already anticipate needing to:

- Create custom mappings from idiosyncratic, collection-specific metadata schemas.
- Identify and deal with collection-specific semantics and string value constructions.
- Separate/parse out concatenated values, e.g., Associated People in Motley, birth/death dates in Kolb-Proust. (We did some similar work in the MARC to LOD transform.)
- Reconcile local with community authorities i.e. with VIAF, TGMI, AAT, etc.
- Deal with string-to-URI reconciliation uncertainties, ambiguities and errors.
- Provide persistent URIs (and a service to de-reference these URIs) for entities in our local authority database that we are unable to reconcile with external authorities.
- Recognize conflated classes (e.g., artwork, associated performance, theater entities).
- Identify requirements to use additional namespaces and/or the need for extensions to schema.org semantics.[10]
- Rely on formatting and simple natural language processing to extract additional entities and improve reconciliation and link (i.e., URI) acquisition.[11]

[8] http://www.culture.gouv.fr/documentation/leonore/recherche.htm

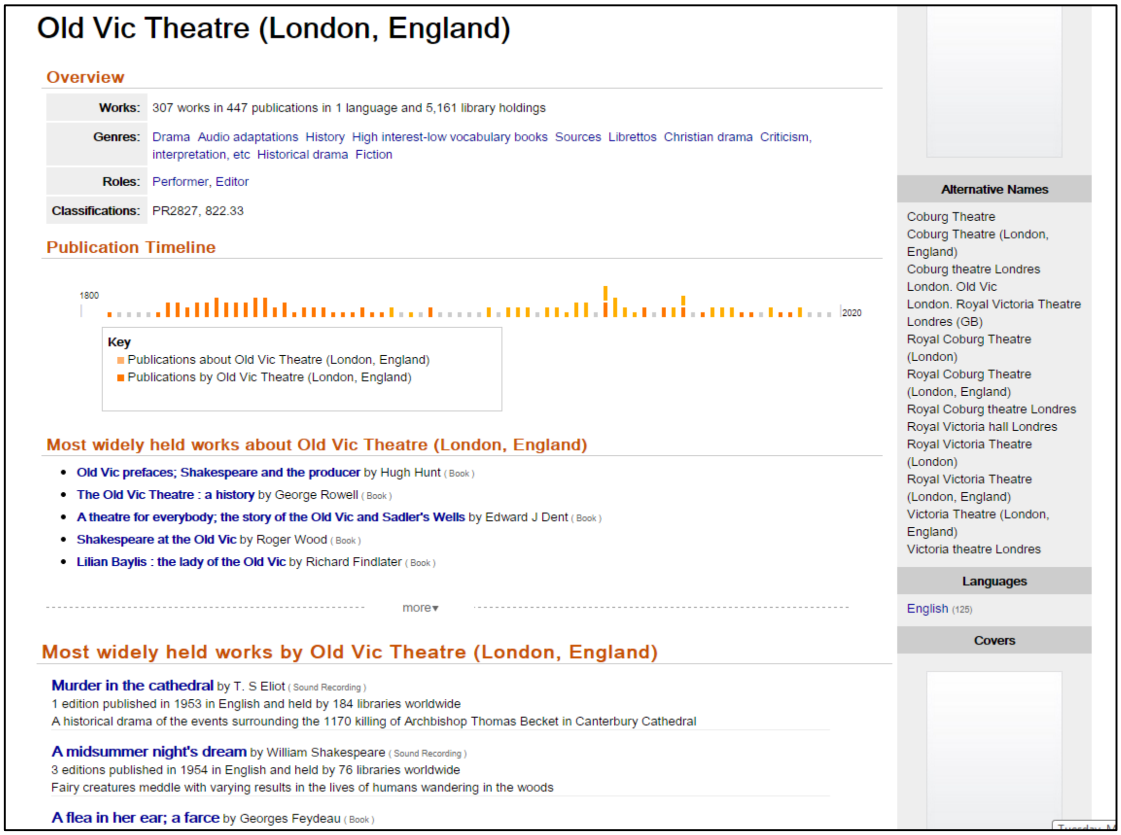
[9] http://www.memoiredeshommes.sga.defense.gouv.fr/en/article.php?larub=108&titre=individual-careers

[10] The need for extensions has been anticipated by schema.org, see: https://schema.org/docs/extension.html

[11] Based on prior experience and allowing for greater breadth, we anticipate obtaining URIs for above 50% of the entities in our metadata using basic heuristics. This will do for this project. Greater levels of reconciliation will be desirable (see: https://mellon.org/news-publications/articles/linked-open-data/), but beyond our scope for now.

## 2.4. Task 2: Special collections LOD as an entrée to general collections

In most physical libraries, special collections are segregated from the general circulating collections of the library. One-of-a-kind image, manuscript and rare book collections are kept in more secure locations with better environmental controls. This separation has largely been carried forward into virtual space and in some ways made even more stark, encouraged not only by how things have been done for print special collections, but also by the tendency today to use specialized digital library content management systems for digitized special collections resources (e.g., CONTENTdm). Yet there is no compelling technical or end-user service reason to maintain such strong separation in the digital realm.



Old Vic Theatre (London, England)

Overview

| | |
|---|---|
| Works: | 307 works in 447 publications in 1 language and 5,161 library holdings |
| Genres: | Drama Audio adaptations History High interest-low vocabulary books Sources Librettos Christian drama Criticism, interpretation, etc Historical drama Fiction |
| Roles: | Performer, Editor |
| Classifications: | PR2827, 822.33 |

Publication Timeline

Key
■ Publications about Old Vic Theatre (London, England)
■ Publications by Old Vic Theatre (London, England)

Most widely held works about Old Vic Theatre (London, England)

- Old Vic prefaces; Shakespeare and the producer by Hugh Hunt ( Book )
- The Old Vic Theatre : a history by George Rowell ( Book )
- A theatre for everybody; the story of the Old Vic and Sadler's Wells by Edward J Dent ( Book )
- Shakespeare at the Old Vic by Roger Wood ( Book )
- Lilian Baylis : the lady of the Old Vic by Richard Findlater ( Book )

more▾

Most widely held works by Old Vic Theatre (London, England)

Murder in the cathedral by T. S Eliot ( Sound Recording )
1 edition published in 1953 in English and held by 184 libraries worldwide
A historical drama of the events surrounding the 1170 killing of Archbishop Thomas Becket in Canterbury Cathedral

A midsummer night's dream by William Shakespeare ( Sound Recording )
3 editions published in 1954 in English and held by 76 libraries worldwide
Fairy creatures meddle with varying results in the lives of humans wandering in the woods

A flea in her ear; a farce by Georges Feydeau ( Book )

Alternative Names

Coburg Theatre
Coburg Theatre (London, England)
Coburg theatre Londres
London. Old Vic
London. Royal Victoria Theatre
Londres (GB)
Royal Coburg Theatre (London)
Royal Coburg Theatre (London, England)
Royal Coburg theatre Londres
Royal Victoria hall Londres
Royal Victoria Theatre (London)
Royal Victoria Theatre (London, England)
Victoria Theatre (London, England)
Victoria theatre Londres

Languages

English (125)

Covers

**Figure 5**: OCLC WorldCat Identities screenshot showing entry for the Old Vic Theater.

An immediate advantage of having LOD-based descriptions of digitized special collections items is the opportunity to better integrate (in virtual space) these collections with general library collections (here and elsewhere) and with library authority services. For example the Old Vic theater name mentioned in the description of the Captain de Foenix costume design sketch (Figure 1) appears in the Virtual International Authority File (VIAF).[12] (There is also an entry for the Old Vic in the Library of Congress Name Authority file.) Once linked to LOD vocabulary services, an LOD-aware interface can be designed to access more than a dozen alternative forms of the Old Vic theater name, including some preferred name variants that are used in national-level catalogs in Europe. These alternate name forms can be made

[12] http://viaf.org

visible to our users to facilitate additional discovery. Usefully VIAF provides a further link to the WorldCat Identities[13] page for the Old Vic Theater (Figure 5).

Through WorldCat Identities users can also be linked to both digital and print books about the Old Vic, as well as to library-curated texts, scripts and audio or video recordings of selected plays performed at the Old Vic. Similarly, books about and other works by the playwright, Sir Arthur Wing Pinero, can be linked through his name authority and WorldCat Identities records. And bibliographic references in Kolb-Proust records (Figure 3), currently presented as plain, unlinked text, can be linked to scanned copies of newspaper articles, fiction, magazine pieces and Proust correspondence held in libraries and archives around the world. It is important to note that once the LOD descriptions have been created many of these kinds of connections can be detected and made automatically by software. Because URIs are more precise than strings (to a computer at least), adding URIs to metadata and transforming the record into an RDF graph conforming to schema.org semantics facilitates computer-based recognition and use of connections.

As part of this proposed project, we will create prototype extensions to our current digitized special collections content management system interfaces (currently CONTENTdm[14] for our image collections and the eXtensible Text Framework [XTF][15] for the Kolb-Proust Archive). These prototype extensions will link users from digitized special collection resources such as depicted in Figures 1 and 3 to related print and digital library resources elsewhere in our library and in other major research libraries worldwide. To gain a preliminary sense from users of the potential benefits we will sample changes in initial user reaction (on a small scale of no more than 5 to 7 student users) before and after these extensions are implemented. User interactions will be one-time short-duration (e.g., ~ 30 minutes) and feature a self-chosen (rather than tester-assigned) Web search testing methodology (Russell and Grimes, 2007). Testers will observe any changes in user actions and attitude when links to library resources outside the immediate collection are added to displays of digitized special collection resources. While too informal and small a sample to be definitive, this preliminary testing can inform strategies for more extensive and systematic testing of LOD benefits for scholarly users of digitized special collections.

LOD enables bi-directional connections. This will allow us to also prototype an interface linking from general library collection resources to related digitized special collection resources based on shared URIs in LOD graphs (e.g., for persons, play name, publications, etc.). Development of this demonstration, proof-of-concept prototype will leverage prior work done at OCLC with an integration of the open source VuFind library online public access catalog and WorldCat Discovery API (Application Programming Interface) software libraries.[16] Because the WorldCat Discovery API can disseminate schema.org RDF descriptions for items in WorldCat, we will be able to augment the standard VuFind display of general collections item metadata with links to related (based on common URIs) digitized special collections items.

[13] https://www.worldcat.org/identities/
[14] https://www.oclc.org/contentdm.en.html
[15] http://xtf.cdlib.org/
[16] Developed by Karen Coombs; used for Developer House: https://www.youtube.com/watch?v=sq4UKD3D48Q

## 2.5. Task 3: Descriptive enrichment and enhanced discovery

Schema.org semantics was initially developed as a collaboration between Google, Yahoo, Microsoft Bing and Yandex as a way to improve Web search engine indexing. Resources on the Web described by schema.org LOD should be more discoverable in the major Web search engines. To enhance discoverability of our special collections resources we will add LOD descriptions created (Task 1) to our item-level splash screens (following the RDFa 1.1 Lite standard[17]). Using transaction logs, we will measure changes in level of use and referrer pages (e.g., to determine if more people are being referred by Web search engines).

Beyond this initial step, we will also algorithmically leverage links in our LOD descriptions to enhance user services by identifying connections between the digitized special collections resources and related non-library resources elsewhere on the Web. This experimentation will encompass non-library – i.e., potentially less scholarly and more popular – Web resources such as Wikipedia.[18] We will implement linking in a way that allows users to move seamlessly from library to external resources. We will manually explore options for linking from Wikipedia to special collection resources. We also will take this opportunity to learn more about the resources we curate, enriching our displays and LOD graphs to enhance discovery and the completeness of the presentation of our resources to end users.

This latter facet of this proposed experimentation is important. A study of scholarly users of digital collections (done as part of the Workset Creation for Scholarly Analysis project) found that the enrichment of item-level metadata poses a critical challenge to digital library development going forward (Fenlon et al., 2014). Item records are the primary – indeed sometimes the only – access point for scholars seeking evidence for their research in digital collections. The study found that scholars consider traditional catalog records to be limited in their capacity to support advanced discovery, interpretation, and analysis. LOD offers a ripe opportunity to improve special collections metadata, by connecting information in records to external sources that provide rich context and authoritative detail.

Fenlon et al. (2014) also found that scholarly users of digital collections have a keen interest in making use of emerging technologies to embed links to other, authoritative sources into item records, and to integrate alternative tools and methods of exploration with existing descriptions and metadata. This can be understood as an explicit request for the implementation of LOD in digital collections, and in particular for the implementation of advanced annotation tools and services (e.g., as enabled by the standards of the W3C Open Annotation Community Group[19] and now also the W3C Web Annotation Working Group[20]). In addition, scholars expressed needs for secondary kinds of information that LOD can demonstrably support. To support both discovery and use, scholars want to see relevant pieces of context integrated with the information found in common metadata fields. For example, beyond an author name, scholars want to discover items based on the gender of the creator of an item. This kind of contextual information – not about the item itself, but about

[17] http://www.w3.org/TR/rdfa-lite/

[18] http://www.wikipedia.org/

[19] Co-Founded by the PI: https://www.w3.org/community/openannotation/

[20] http://www.w3.org/annotation/

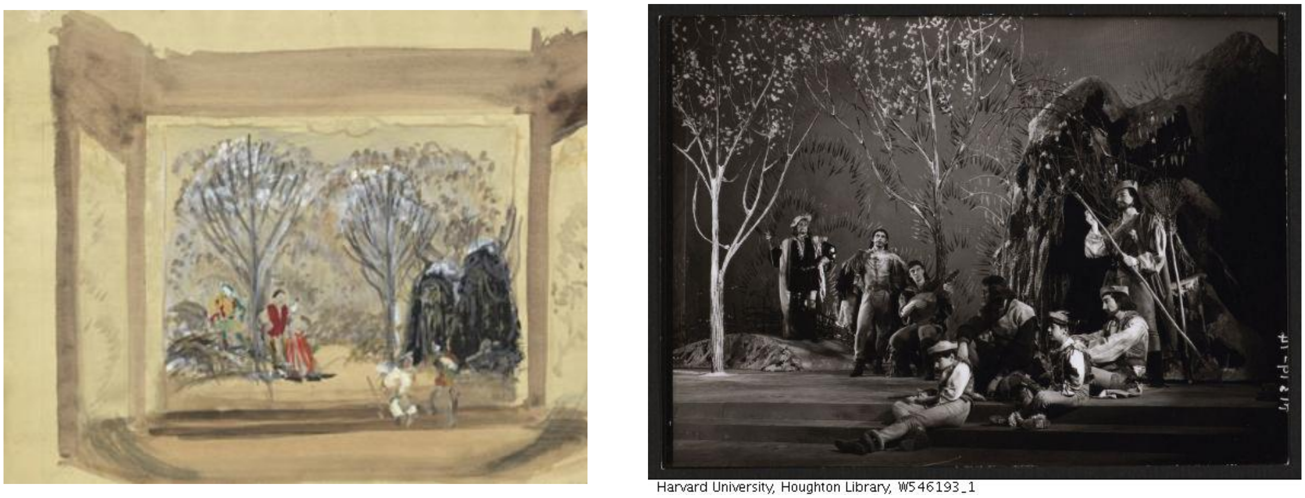
entities (such as creators, dates, and subjects) associated with an item through descriptive metadata – does not normally appear in static, library item records, but it often does exist in external resources (e.g., Wikipedia articles and associated DBpedia LOD graphs[21]) that can be linked from library metadata. LOD opens up digitized special collections to more advanced routes of discovery and interpretation.

Returning to the example of the "Captain de Foenix" costume rendering shown in Figure 1, a scholar might find this item relevant or interesting for any number of reasons: for its relevance to a particular historical performance; for its part in the history of the Old Vic theater; for its expression of the character, Captain de Foenix; etc. Including links to the wider Web for any of these facets would allow scholars to pivot on the components of information that are most critical to their work and in doing so leverage the contextual mass that exists outside of and (currently) unconnected to digitized special collection metadata records. As a concrete illustration, consider the value of a link from the play performed information in the metadata for the Captain de Foenix costume sketch image to the Wikipedia article on the play.[22] The Wikipedia article provides additional information about original staging, a brief synopsis, information on two film adaptations, and a brief bibliography. Similarly a link to Wikipedia from the "Associated People" field, which currently gives the name, "McKellen, Ian" without explicit indication of McKellen's part in the play, would allow a scholar interested in McKellen's career to see that this was one of McKellen's earliest performances; that in the same year he also played in "*Much Ado About Nothing*"; and that he performed this role as part of Laurence Olivier's National Theatre Company at the Old Vic. The connection of contextual information potentially casts the item in a more interesting light. Any of this added context, currently invisible to a user of this special collection, might prove relevant to a scholar's inquiry and make the costume sketch more useful. Additionally, such context can be potentially valuable in facilitating discovery.

This last observation suggests a need to investigate how, once integrated in the larger knowledge graph of the Semantic Web, we can take advantage to enrich our own resource descriptions and identify new connections. We anticipate being able to do this both by mining external LOD (e.g., DBpedia) and by making it possible for scholars to add their own links through annotation. By design, linkages between LOD described sources can be detected algorithmically, but not all related resources are currently described using LOD, and because the significance of some relationships is contextual, LOD descriptions will not include all possible relationships. Scholars often broach the task of enriching metadata for analytic purposes on their own, as part of their "preprocessing" labor prior to analysis. While implementing LOD may help offload some of this burden, expert metadata enrichment also poses an opportunity for LOD to support feedback of enriched data into discovery systems. Fenlon et al. (2014) found that scholars expressed a desire to contribute their own, enriched data back into digital collections or discovery systems. As one scholar noted, "[Y]ou've done all this work, and you then have the authoritative metadata. You have the best metadata in the world, and no one will take that from you" (Fenlon et al., 2014).

[21] http://wiki.dbpedia.org/about

[22] http://en.wikipedia.org/wiki/Trelawny_of_the_%27Wells%27



Harvard University, Houghton Library, W546193_1

**Figure 6**: A set design sketch from the Motley Collection (left) and a photograph from the Harvard Theatre Collection, both from a 1957 Royal Shakespeare Co. performance of *As You Like It*

An illustration of the potential benefit of being able to ingest scholar enrichment is shown in Figure 6. On the left is, from the UIUC Motley Collection, a set design sketch for a 1957 performance of *As You Like It* at the Shakespeare Memorial Theatre in Stratford-upon-Avon. On the right is, from the Harvard Theatre Collection, a photograph of the realized set design that was taken during this performance. If both images had LOD description their connection could be found algorithmically. Lacking this, an LOD description for the UIUC owned image allows the connection to be made by a scholar annotation.

This phase of our research will shed light on the difficulties of using LOD to establish connections from our special collection items to external resources such as found in Wikipedia, the Internet Broadway Database, etc.; how feasible and automatable it is to leverage LOD to identify and establish links from Wikipedia back to relevant items in our special collections; and how, once linked through our own LOD to other segments of the Semantic Web (e.g., DBpedia), we can best enrich our descriptions, both algorithmically and with the help of user contributed annotations, to facilitate discovery. Ultimately we anticipate that this increased connectedness to external resources and the associated enhanced navigation will stimulate and in some cases enable new research investigations and insights. The logical initial step from where we are today is simply to begin connecting more dots, more efficiently and more effectively. This in turn will open the way to further research into the efficacies and trade-offs of an LOD approach.

## 2.6. Task 4: Visualizing the social network of Marcel Proust

Most library application interfaces have intrinsically bibliocentric perspectives. Library Online Public Access Catalogs (OPACs) present search results as a list of bibliographic items, ordered according to the relevance of each item to the query as determined from inspection of the item's bibliographic attributes (e.g., author name, title, subject). Even library search interfaces for digitized special collections, e.g., cultural heritage image collections, tend to adapt this paradigm, replacing author name with photographer or artist name, title with image label, and so on. Much has been made of 'serendipity' when discovering resources in print libraries, but in reality serendipity when browsing library shelves relies on collation by call number (i.e., subject) and author; other associations and resource properties or relationships as a basis for browsing are neglected.

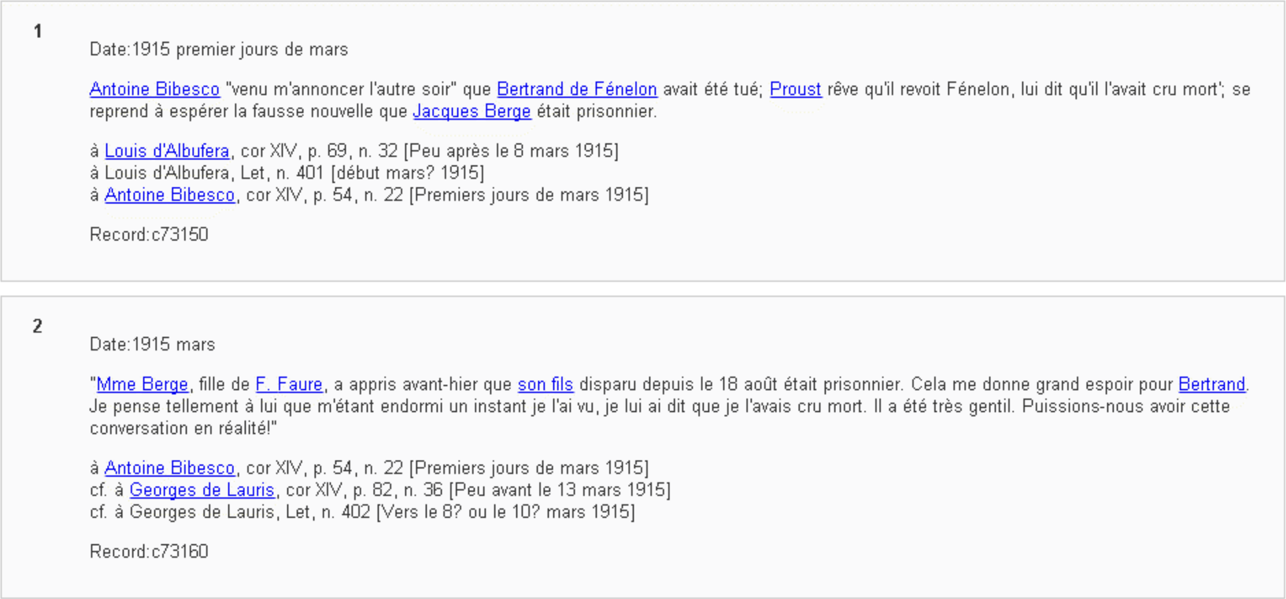
Digital technologies provide an opportunity to explore additional views of digitized special collections, supporting different ways of browsing and different modes of discovery (and serendipity). Newer networked archival and special collection projects are beginning to experiment with innovative discovery and exploration paradigms, including a few based on the social networks relating to specific historical periods, individuals or special collections archives. Notable examples are the *Six Degrees of Francis Bacon Project: Reassembling the Early Modern Social Network*[23] and the *Social Networks and Archival Context (SNAC)* project.[24] The SNAC prototype research tool includes a radial graph demo component that allows users to browse a graphical representation of the network of organizations and people mentioned in described archival resources that have been indexed by the service. Lynch suggests that this functionality is illustrative of a data-centric approach that "enables structures and relationships to emerge in the system that would otherwise go unnoticed without a lucky combination of serendipity and painstaking manual research by humans" (Lynch 2014, p. 17).

The precision and fine granularity of LOD descriptions support a broad range of perspectives on the resources described. As part of this project we will explore the potential benefits of one alternative perspective on the resources described in the Kolb-Proust Archive, providing as part of an experimental interface a social network view of the people mentioned and described in resource descriptions. (This interface will be built atop XTF using available Open Source SVG utilities or an Open Source, off-the-shelf javascript visualization library.) For example, the record from the Kolb-Proust Archive shown in Figure 3 above, references three individuals, Paul Acker (a journalist and novelist); Anna, Comtesse Mathieu de Noailles (a novelist and poet of Romanian descent); and René Georges Hermann-Paul (French artist). These three individuals can be thought of as nodes in the social network graph surrounding Marcel Proust. In this instance all three are mentioned in other records (the Comtesse alone, some 217 times) in connection with other individuals in Proust's circle. Expanded to include all at once the 7,000 names mentioned in the Archive would create a network of names too large too usefully visualize, but views centered on an individual can be taken in, especially if the interface makes it possible to further constrain by date range or other filter (e.g., publication where mentioned, etc.). Having discovered an individual of interest on the network, the user can then explore the mentions in the Archive, and as enabled by the LOD descriptions, the user can connect out to other related library or Web resources – e.g., to the contemporary newspaper articles in which the person was mentioned, to Wikipedia articles, to entries in the resources such as the French National Archive's Légion d'honneur database, and to other bibliographic resources by or about the individual.

[23] http://sixdegreesoffrancisbacon.com/ and also Finegold et al. (2013).

[24] http://socialarchive.iath.virginia.edu/

There are of course limitations to this approach and to what can be algorithmically divined from the existing string-based metadata. To help overcome these limitations we would again enable limited user annotation of the edges of the Proust social network graph. Consider the two entries from the Kolb-Proust Archive concerning Jacques Berge as shown in Figure 7.



1

Date:1915 premier jours de mars

Antoine Bibesco "venu m'annoncer l'autre soir" que Bertrand de Fénelon avait été tué; Proust rêve qu'il revoit Fénelon, lui dit qu'il l'avait cru mort'; se reprend à espérer la fausse nouvelle que Jacques Berge était prisonnier.

à Louis d'Albufera, cor XIV, p. 69, n. 32 [Peu après le 8 mars 1915]
à Louis d'Albufera, Let, n. 401 [début mars? 1915]
à Antoine Bibesco, cor XIV, p. 54, n. 22 [Premiers jours de mars 1915]

Record:c73150

2

Date:1915 mars

"Mme Berge, fille de F. Faure, a appris avant-hier que son fils disparu depuis le 18 août était prisonnier. Cela me donne grand espoir pour Bertrand. Je pense tellement à lui que m'étant endormi un instant je l'ai vu, je lui ai dit que je l'avais cru mort. Il a été très gentil. Puissions-nous avoir cette conversation en réalité!"

à Antoine Bibesco, cor XIV, p. 54, n. 22 [Premiers jours de mars 1915]
cf. à Georges de Lauris, cor XIV, p. 82, n. 36 [Peu avant le 13 mars 1915]
cf. à Georges de Lauris, Let, n. 402 [Vers le 8? ou le 10? mars 1915]

Record:c73160

Figure 7: Two references to Jacques Berge in the *Kolb-Proust Archive for Research*

These are excerpts of letters from 1915 where Proust talks about friends of his who died in combat, or are missing in action. Jacques Berge, son of his friend Antoinette Faure (daughter of French president Félix Faure), is missing in action, his mother got the false news that he was made prisoner, while Proust had a dream that Berge had been killed. Berge's fate was not known until much later. The WWI *Mémoire des hommes* archive includes a record[25] showing that Jacques Berge actually died very early in the war, on August 22, 1914 in Belgium, but the bottom part of the document shows that his death was not confirmed until June 27, 1919 by a Parisian tribunal, with a judgment recorded on September 9, 1919. In this case, machine processing of our existing metadata would not make clear all of these details, only that Jacques Berge was mentioned together with Antoinette Faure (aka Ms. René Berge), Félix Faure, and others in correspondence between Proust and Antoine Bibesco. If provided, user annotation of these relationships and the implication of the information provided by the record from the external *Mémoire des hommes* archive could enrich the LOD graph significantly. This suggests that a social networked perspective coupled with linking to a mix of traditional library bibliographic sources and biographical Web resources may represent a new and different way for libraries to facilitate new more complex, connective work by historians and other scholars, e.g., along the lines suggested by Rogers (2008).

[25] http://www.memoiredeshommes.sga.defense.gouv.fr/fr/ark:/40699/m005239d8af1655f/5242bc2cb859c

# 3. Collections

The primary sources and descriptive metadata found in the three collections chosen for this project support a broad range of humanities pedagogy and advanced scholarly inquiry spanning multiple disciplines, e.g., art, design and theater scholars, historians of the 18th, 19th and 20th century, literary scholars, and scholars interested in the broader relationships between literature, theater, culture and society. Each of the collections makes available unique digitized primary sources and/or provides context for humanities scholarship in a specific domain.

## 3.1. *Motley Collection of Costume & Theatre Design / Portraits of Actors Collection*

The Motley Group (Mullin, 1996), which consisted of Margaret Harris, her sister Sophia Harris, and Elizabeth Montgomery, designed sets and costumes from 1932 to 1976 for performances of plays by Shakespeare as well as for performances of modern classics, opera, and ballet. Their designs were used in productions in the West End of London, the Royal Shakespeare Theatre, the English National Opera, and in the United States on Broadway and at the Metropolitan Opera in New York City. The Motley Group was highly innovative in designing sets and costumes that suggested the mood, architecture, and styles of the original setting of the play, but was not the rote duplication that had been done so many times before. They wanted to create an atmosphere that was artistic, in addition to having an air of authenticity. Motley set the standard for how Shakespearean productions should be staged. The Group's work diversified in 1940 when Margaret Harris and Elizabeth Montgomery went to New York to design a production for Laurence Olivier and had to remain there for the duration of World War II, while Sophia Harris continued to work in London.

The *Motley Collection of Costume & Theatre Design*, a unique and valuable source for documentation on the history of theatre that was purchased outright by the University of Illinois Library in 1981, includes more than 5,000 items, mostly original sketches for costume and set designs (all have now been digitized) used in over 150 productions at 61 different venues in the UK and the United States. About 49 of these 61 venues have a Wikipedia page, and some have specific production history pages. As illustrated in Figure 1, metadata for this collection is quite rich, referencing the actors who wore the costumes (e.g., Ian McKellen), the author of the play (e.g., Sir Arthur Wing Pinero), the director (e.g., Desmond O'Donovan), as well as the theatres in which the play was performed using the set or costume design. Two of the named entities from Figure 1 do appear in VIAF (O'Donovan does not), but all three men are described in other Web sources, e.g., Wikipedia, IMDb, IBDB, etc. Though not all of these Web resources currently support LOD, they do have persistent URIs for entities that can be used as links. Similarly, the play, the character, the venues and the various performances of the play are referenced in multiple Web sources, and of course the script for the play itself is held by a number of libraries.

The *Portraits of Actors Collection, 1720 – 1920*, includes more than 3,500 studio portraits of actors posing in costume for a particular role or performing a scene from a play. To help provide a more complete view of historic play performances, the collection also includes portraits of dramatists, theatrical managers, singers and musicians, but the majority of the portraits are of British and American actors who worked between about 1770 and 1893. The images in this collection were digitized from etchings, engravings, lithographs, mezzotints, aquatints, wood engravings, photographs, and photo-mechanically-reproduced prints, all from the University of Illinois Theatrical Print Collection.

The *Motley* and *Portraits* collections are complementary in time spans covered, but also in another, more nuanced way. The Motley collection comprises the working, unpublished records of a design group, whereas the items in the Portraits of Actors Collection were all published or destined for mass consumption. The images in the Portraits of Actors Collection are among the earliest examples of the mass production of celebrity images, in many ways the forerunners to *Us* Magazine and *Entertainment Tonight*. Entities such as individuals (actors, directors) and specific performances (venues and dates) should make it possible to link many of these digitized portraits to announcements and reviews of performances increasingly to be found online in digitized national and regional press repositories.

These collections today are of interest to many scholarly and lay users, including theater historians and biographers and performing arts and theater professionals such as costume and set designers doing research for current and future productions. Motley images have been licensed for inclusion in Shakespeare textbooks and we have anecdotal evidence[26] that these images are used in classroom settings and for course projects.

In transforming the metadata for these collections, it will be a challenge to adapt the lessons learned in transforming MARC metadata and Early Modern Emblem-level metadata[27] (Cole et al., 2013) to the distinctive metadata schemas of these collections. To meet this challenge, we will identify and integrate into our workflows new authorities from both library and non-library sources. We anticipate that what we learn and strategies used will be of interest to and potentially adaptable for transforming metadata of other digitized special collections.

## 3.2. The *Kolb-Proust Archive for Research*

The digitized documents that comprise the Kolb-Proust Archive for Research (Szylowicz and Kibbee, 2004) are different from the two collections discussed above. These are not primary sources themselves but rather are an archive of raw scholarly output, representing an extensive body of scholarly research notes compiled over fifty years by Philip Kolb, a University of Illinois faculty member and the editor of Marcel Proust's correspondence. These notes provide context for and descriptions of Proust's correspondence and publications. They identify individuals, places and events mentioned in Proust's letters. A selection of these notes were then "distilled" into the critical apparatus of the print edition of Proust's correspondence (Paris, Plon, 1970-1993, 21 volumes).

Kolb's research files, about 40,000 methodically cross-referenced index cards, can be understood as a first layer of metadata around the primary sources that are Proust letters and manuscripts, a map of his literary and cultural universe. The Archive contains considerably more data than Kolb was able to fit in the print edition, and draw from known letters, biographical sources, newspapers and other periodicals of the time, social directories, etc.

[26] E.g., the recommendation of the *Protraits of Actors Collection* as a source of images for students taking Theatre 101 at Williams College: http://library.williams.edu/subjectguides/theatre/thea101/images.php

[27] See our OpenEmblem Portal (http://emblematica.library.illinois.edu/oebp/ui/)

Kolb, one of the first scholars to do research on Proust also obtained unpublished information directly from Proust's contemporaries (Proust died in 1922 at age 51) and consigned those details to his files.

The digitization and encoding of Kolb's research notes adds a second layer of useful metadata and authority control: all cited individuals have been assigned a unique identifier, all literary and creative works cited have been assigned a genre category (fiction, poetry, music, sculpture, etc.), and all bibliographic references were standardized, which will facilitate linking these metadata to resources such as digitized newspapers (most French press of the time have been scanned and made available by the BnF[28]) and other digital surrogates (digitized books, and image or sound repositories, and Proust's own manuscripts, also digitized and hosted by the BnF[29]).

The local name authority file created for the Kolb-Proust Archive augments name strings with dates (birth, death, marriage, etc.), and includes notes on profession and/or kinship. In reconciling names with external authorities, this ancillary information associated with each name will facilitate identification and disambiguation. In the current interface, our local authority file serves also as an index and has become a biographical resource of its own, from which users can link back to name mentions in the Kolb-Proust papers. We will leverage these metadata to help users link out to other resources, including non-bibliographic repositories: genealogical sources, and institutional sites such as the official records of the Legion of Honor or the French Ministry of Defense database listing all WWI casualties. There is also a growing body of other digitized vital records resources. Though most do not currently include LOD descriptions of entities indexed, we anticipate the potential for user-contributed annotations linking names found in the Kolb-Proust Archive to entries in these additional sources.

The *Kolb-Proust Archive for Research* is a bi-lingual site and first became available online in the mid-1990's. It has an established user base including professional and lay Proust scholars in France and Europe, Japan, Brazil, and the US. It also has many other users interested in other aspects (i.e., not just Proust) of the literature, history, art history, and music history of turn-of-the-century French society and culture. Other online scholarly resources about prominent figures from Proust's time present additional potential for making connections, both through machine mediation and through user-contributed annotations. For example, a developing site about the French poet Anna de Noailles[30] (see her mention in the *Archive* record shown in Figure 3). Or the Web resources concerning Reynaldo Hahn,[31] a French composer, music critic, and lifelong friend of Proust. It is worth noting that the Hahn resource has already extensively mined (manually, we assume) all mentions of Hahn from the *Kolb-Proust Archive* and presents these within the larger Hahn Website.[32] This contributes to our hypothesis that a social network perspective on Proust, his letters and publications and his circle of acquaintances will be useful and of interest to scholars.

[28] http://gallica.bnf.fr/html/presse-et-revues/les-principaux-quotidiens

[29] http://gallica.bnf.fr/Search?ArianeWireIndex=index&p=1&lang=FR&q=Fonds+marcel+proust&x=0&y=0

[30] http://www.annadenoailles.org

[31] http://reynaldo-hahn.net/index.htm

[32] http://reynaldo-hahn.net/Html/ecritsdiversProust.htm

# 4. Staff and Organization Qualifications

## 4.1. Organizational strengths

The University of Illinois at Urbana-Champaign is a nexus for digital humanities, information science and knowledge management research and development. Building on a now long history of close and successful collaboration, the proposed project will be a joint endeavor of the University Library and the Center for Informatics in Science and Scholarship (CIRSS) in the Graduate School of Library and information Science (GSLIS). We will draw on the strengths of these two entities as well as experience gained in past and ongoing research partnerships with the Illinois Program for Research in the Humanities, the HathiTrust Research Center, and the National Center for Supercomputing Applications. Close working partnerships with these and other specialized research centers and consortia beyond the University enable the Library and CIRSS to create, develop and provide forward-looking services to scholars across the disciplines.

The *Center for Informatics in Science and Scholarship*[33]
CIRSS conducts research on information problems that impact scientific and scholarly inquiry with a specific focus on how digital information can advance the work of scientists and scholars, the curation of research data, and the integration of information within and across disciplines and research communities. CIRSS researchers, faculty and staff bring a range of expertise to the center's projects in areas including empirical studies of scientific information use, information modeling and representation, ontologies, data curation, and digital research collections and technologies. The center's staff includes project coordinators, research assistants and other academic staff with experience in project management, quantitative and qualitative methods, research with human subjects, and the design and conduct of multi-method research and evaluation studies in information science and cognate social sciences. CIRSS builds on synergies in four key intellectual areas: 1) digital humanities; 2) collections, curation, and metadata; 3) e-Science; and 4) socio-technical data analytics. CIRSS is a core research center within GSLIS. Founded in 1893, GSLIS, the iSchool at Illinois, is a world leader in library and information science education, research and practice. Consistently ranked as one of the very best in the field, GSLIS has earned its reputation by creating pioneering and innovative educational opportunities, by leading groundbreaking research to advance preservation of and access to information in both traditional and digital libraries, and through its services and strong commitment to outreach and community development.

The *University of Illinois Library at Urbana-Champaign*[34]
The Library at Illinois is one of the preeminent research libraries in the world. As the intellectual heart of the campus, the Library is committed to maintaining the strongest possible collections and services and engaging in research and development activities in pursuit of the University's mission of teaching, scholarship, and public service. The Library provides a rich range of services geared to support the curricular and research needs of students and faculty and serve the dynamic needs of scholars in the digital age both local and remote. The Library was established in 1867 with only 644 books purchased with $1,000

[33] http://cirss.lis.illinois.edu/
[34] http://www.library.illinois.edu

appropriated by the State of Illinois. Today it houses more than 22 million items, and it is known for the depth and breadth of its collections. Materials from the library are actively used, with more than 1.4 million items circulated annually and subscriptions and licenses for over 50,000 e-journals resulting in over 7 million user click-throughs per year via an e-resource registry and over 11 million full-text downloads. The Library currently employs approximately 90 faculty and 300 academic professionals, staff, and graduate assistants who work in multiple departmental libraries located across campus, as well as in an array of central public, technical, and administrative service units. The Library also encompasses a variety of virtual service points and "embedded librarian" programs that provide library services to scholars across the spectrum of research environments. Librarians are full faculty members of the University and contribute significantly to scholarly literature in their respective fields of study. The Library plays a leadership role in regional, national, and international organizations; provides services to users throughout the State of Illinois; and serves as an integral part of the worldwide scientific and scholarly community.

## 4.2. Existing staff qualifications and roles

*Project PI: Professor Timothy W. Cole* (5%)
Timothy W. Cole is Mathematics Librarian (University Library) and CIRSS Coordinator for Library Applications (GSLIS). He is a co-PI for the *Workset Creation for Scholarly Analysis*[35] project and for the *Emblematica Online*[36] projects. He was previously the PI for the *Open Annotation Collaboration* projects (all phases, 2009-2013) and the *Digital Collections and Content* projects (phases 1 & 2, 2002 – 2007), as well as PI or co-PI for multiple other projects involving metadata and digital library system design, interoperability and implementation. A member of the Illinois faculty since 1989, he has held prior appointments as Interim Head of Library Digital Services and Development, Systems Librarian for Digital Projects and Assistant Engineering Librarian for Information Services. He is a member of the International Mathematical Union Committee on Electronic Information and Communication, a member of Library Hi Tech Editorial Board, a past member of the National Academies Committee for Planning a Global Library of the Mathematical Sciences,[37] past chair of the National Science Digital Library Technology Standing Committee and a former member of the Open Archives Initiative Protocol for Metadata Harvesting Technical Committee. He has published and presented on metadata and LOD best practices, OAI-PMH, digital library interoperability, Open Annotation, and the use of XML for encoding metadata and digitized resources in science, mathematics and literature. As PI for the proposed project, Cole will be responsible for the direction and overall execution of the project and in particular for overseeing the technical and interface design work carried out by project developer, post-doc and hourly staff.

*Project Co-PI: Associate Professor Myung-Ja Han* (5%)
Myung-Ja Han is Metadata Librarian (University Library). She is a co-PI for the *Emblematica Online* projects and is the 2015 recipient of the Esther J. Piercy Award, given annually by the Association for Library Collections & Technical Services (of the American

[35] 2013 – 2015, funded by the Andrew W. Mellon Foundation, http://worksets.htrc.illinois.edu/worksets/
[36] 2009 – 2015, funded by the NEH and the DFG, http://emblematica.library.illinois.edu/
[37] 2012 – 2014; this group published their final report in 2014: http://www.nap.edu/catalog/18619/developing-a-21st-century-global-library-for-mathematics-research

Library Association) to a technical services librarian with less than 10 years professional experience who has made outstanding contributions to date and has shown outstanding promise for continuing contribution and leadership. A member of the faculty at Illinois since 2006, Han has worked extensively with the Library's general and special collection catalog records and metadata and in support of the Library's participation in both the Open Content Alliance the Google Book mass digitization initiatives. She has published and presented extensively on library cataloging, metadata, XML and Linked Open Data. Han will assist in the direction and overall execution of the project and in particular will direct, help design and oversee implementation of metadata reconciliation and transformation workflows carried out by the project developer and hourly staff. She will also assist in directing Task 2 (using LOD for digitized special collections as an entrée to other library collections) and in the development of strategies for linking to non-library Web resources.

*Project co-PI: Associate Professor Caroline Szylowicz* (5%)
Caroline Szylowicz is French Subject Specialist, Kolb-Proust Librarian, and Curator of Rare Books and Manuscripts (University Library). A member of the faculty at Illinois since 1994, Szylowicz created the Kolb-Proust Archive for Research from the ground up, working closely with Kolb family members and with Proust scholars in the US and France. Her leadership and vision created an invaluable and powerful resource for readers and scholars of Proust. She writes and presents regularly on the life and letters of Marcel Proust in a wide range of venues in both the US and Europe, and in both English and French. Since joining the Rare Books and Manuscripts Library (RBML) at Illinois as a Curator of Rare Books and Manuscripts, Szylowicz has broadened her knowledge of additional digitized special collections. In addition to collection development, maintenance and reference responsibilities pertaining to our special collections materials, Szylowicz does extensive outreach and instructs both undergraduate and graduate students in the role and collections of RBML. For the proposed project Szylowicz will bring her in-depth domain knowledge, especially regarding Proust, and with the project post-doc and project advisors will lead in identifying subjects for user testing, assembling our project midpoint local scholar panel and otherwise soliciting scholar feedback and reactions to prototypes.

*Project Developer: M. Janina Sarol* (50%)
Janina Sarol is a Visiting Research Programmer (University Library). Her primary job assignment is to support digital library research. Sarol, who has a BS in Computer Science awarded by the University of the Philippines – Diliman in 2011 and is a member of the W3C Web Annotation Working Group and the W3C Schema.org Community Group, joined the University Library in early 2014 to take over as the lead developer for the second phase of the *Emblematica Online* project. In this role she has implemented a number of LOD features in the *OpenEmblem Portal*.[38] In 2014 she also served as lead developer for the Library's project to create a schema.org LOD snapshot of the UIUC general collection catalog (5+ million bibliographic MARC records and 10+ million holding records) and participated in OCLC Developer House, working on extensions to the instance of VuFind that Karen Coombs of OCLC has modified to work with RDF and the WorldCat API.[39] Sarol continues as the lead developer for *Emblematica Online* (through November 2015) and now serves

[38] http://emblematica.library.illinois.edu/oebp/ui/
[39] Available on GitHub under the GNU General Public License (2.0): https://github.com/librarywebchic/vufind

also as lead developer for the *Workset Creation for Scholarly Analysis* project (through September 2015). For the proposed project, Sarol will be the lead developer across all tasks, working under the direction of the PI and co-PIs and closely with other project staff involved in reconciliation and transformation workflows, interface modification tasks and the creation of the Proust social network visualization interface.

### 4.3. Project staff to be named / hired

GSLIS will be adding a new post-doc this fall. (Recruitment of this individual is already underway and a well-qualified candidate has been identified and expressed interest.) CIRSS will name a project coordinator for this project from among existing CIRSS staff prior to project start. Additional academic & graduate hourly staff with requisite skills and experience will be hired at project start and over the course of the project as described below.

*Project Post-Doc* (20%)
This individual will have a visiting appointment in GSLIS and an earned doctorate in Library and Information Science or a closely related discipline. He or she will bring to the project expertise and experience working with digitized cultural heritage resources and metadata and with Web interface functional design and implementation. For the proposed project, this person, working under the direction of Cole (PI), will participate across all project tasks as required, focusing especially on interface design pertaining to the integration of links to / from related resources (both library and non-library) and to the design of social network visualization functionality. He or she will also provide feedback on the efficacy of reconciliation and transformation workflows and will work closely with Szylowicz (co-PI) and Cole in helping to elicit feedback from scholars, students and other users of digitized special collections.

*Project Coordinator* (20%)
CIRSS project coordinators have a Master's degree in Library and Information Science or Museum Studies, relevant professional work experience and expertise in project management and coordination. For the proposed project, the CIRSS staff member selected will monitor, organize and help ensure the smooth running of the project and the timely execution of project deliverables, including white papers and interim and final reports. The Coordinator will work with Cole (PI) to help handle the basic administrative aspects of the project including research meeting planning, communications, time and effort reporting, budget monitoring, and gathering input for the interim and final reports. The Coordinator will work with Cole, Han (co-PI) and Szylowicz (co-PI) to recruit and oversee project hourly staff. The Coordinator will work with Szylowicz, the project post-doc and hourly staff to coordinate user testing (task 2) and the convening of the local scholars panel, and with Cole (PI) to convene the early 2017 Advisory Committee meeting.

*PhD Research Assistant & Academic Hourly Staff* (20 hours per week total, i.e., 50% FTE)
A 25% full-time-equivalent PhD Research Assistant (tentatively Jacob Jett) will be assigned to this project. (The PhD Research Assistant will be assigned to this project on an hourly basis, i.e., on average 10 hours / week. GSLIS will provide all tuition remission; consistent with Foundation policy, no tuition remission fees will be charged to the grant.) Additional CIRSS affiliate(s) with similar academic qualifications, i.e., having at least a Master's degree

in Library and Information Science (or equivalent), will be hired / assigned to this project *also on an hourly basis* at 10 hours per week to perform specific subtasks. Jett served previously as Project Coordinator for Phases 2 & 3 of the Open Annotation Collaboration and has both a Master's degree in Library and Information Science and a Certificate of Advanced Studies in Library and Information Science with a concentration in Digital Libraries. For this project Jett, who currently leads the collection modeling in RDF component of the *Workset Creation for Scholarly Analysis* project, will focus on metadata mapping, reconciliation and transformation issues; the generation of RDFa for embedding in resource splash screens; and the implementation of annotation and bi-directional linking functionality. He will work under the direction of Cole (PI) and closely with the project co-PIs, developer and post-doc. Additional academic hourly CIRSS affiliate(s) with similar degree qualifications but greater expertise in user testing, focus group facilitation and scholarly domain requirements will augment Jett's participation and work on other facets of the project.

*Grad Hourly Staff (Masters candidate level)* (20 hours per week total, i.e., 50% FTE)
Additional CIRSS staff (2 concurrently for most of the project) who are candidates for a Masters degree in Library and Information Science will be hired / assigned to this project *on an hourly basis*. Half of the hours under this budget line will be allocated for an individual to work with Han (co-PI) on metadata reconciliation and transformation issues and the implementation of linking to library and non-library resources. This individual will have expertise with library metadata and XML and at least some knowledge of TEI, RDF and schema.org semantics. The rest of the allocation will be for an individual with domain expertise who will work with and under the direction of Szylowicz (co-PI) to identify sources for reconciliation and potential linking as well as to help in soliciting user feedback.

### 4.4. Advisors and soliciting user feedback

Because this project focuses on growing connections across collections and on how this better connectedness can then be leveraged to better meet scholarly user needs, it is important that we solicit expertise from colleagues and feedback from users on an ongoing basis. Locally in the University Library and in CIRSS we are fortunate to be able to draw on a broad mix of local expertise, e.g.: Harriett Green, English and Digital Humanities Librarian and member of the Library's Scholarly Commons team, who helps connect humanities scholars to digital resources; Ayla Stein, metadata librarian, and Patricia Lampron, academic hourly metadata specialist, both of whom were involved along with Cole, Han and Sarol in transforming a snapshot of our general collection MARC catalog into schema.org LOD. Prior to project start we will assemble a formal Project Advisory Board of between eight and ten experts in LOD, related technologies, and the use, curation and/or delivery of digitized cultural heritage materials of the sort with which we will be working.

Over the course of the project we will keep Board members appraised of progress and consult with members one-on-one as appropriate to each member's expertise. In early spring 2017, we will convene a face-to-face meeting of the Board in Chicago to present preliminary findings and solicit feedback from Board members on the substance of our findings, prototypes and demonstrations; on the potential impact of findings; on best avenues for further dissemination and propagation of outcomes; on advice regarding next logical steps.

We also will seek some initial reaction and feedback from current and likely users of our test special collections. We will do this via two mechanisms. In January of 2016 we will conduct five to eight tests with individual graduate students interested in theater arts or the work of Marcel Proust. For these tests participants will use the current interfaces and services, i.e., without LOD descriptions and disconnected from resources in other collections are elsewhere on the Web. We will use a self-chosen Web search task testing model (Russell, 2007), observe what they do on finding resources in the collection, and at the end of the test ask questions about satisfaction with process and usefulness of results obtained. In January 2017 we will repeat the test using new, prototype interfaces with LOD descriptions and active links to additional resources, and compare results. As a second way to solicit scholar feedback, we will convene a local scholars' panel at the halfway point of the project. This panel of eight faculty and advanced graduate students actively doing relevant research will comment on the likely value and utility of links identified through LOD transformation and react to mockups of interfaces that will leverage these linkages. Several potential panel members have already been identified and expressed tentative interest in participating, e.g., Curtis Perry (Professor, English), Sara Thiel (PhD candidate, Theatre), Lori Newcomb (Associate Professor, English), François Proulx (Assistant Professor, French & Italian).

## 5. Work Plan and Expected Outcomes

### 5.1. Summary of expected outcomes and benefits

An immediate benefit of this project will be the increased visibility of the three collections featured in the research. We anticipate that adding LOD descriptions, serialized as RDFa, to item-level splash screens will make the items in these collections more discoverable and more useful in collaborative discovery contexts (e.g., Europeana, see below). By algorithmically adding links to and from items, items will be made both more discoverable and more useful, though how much so remains to be seen. More generically and of broader benefit, the concrete experience gained with reconciliation and transformation of legacy special collections metadata into LOD, with initial integration of LOD into user interfaces, and with early user feedback, will help inform decisions by libraries and library leadership regarding next steps to make digitized special collections more visible and useful. To maximize the project's impact, we will generate the following specific work products:

- ***White Paper 1***: describing reconciliation & transformation, workflows implemented, resources required and lessons learned, and providing advice on transforming special collections metadata to LOD;
- ***White Paper 2***: describing strategies and early user feedback for use of LOD descriptions of digitized special collections to connect collections to the larger semantic Web, and identifying at least qualitatively potential benefits.
- ***Scripts & code used in reconciling and transforming legacy metadata***: As used to process legacy metadata, extract entities, replace strings with URIs, and generate LOD descriptions. Scripts will be made available via a project GitHub repository under the NCSA/UIUC Open Source license. Should be considered of early maturity (e.g., a late alpha release) and will need to be customized to use for records of different collections, but scripts will represent a significant time-saving starting point for new projects.

- ***Extensions to XTF (XML stylesheets) to accommodate LOD***: As *Kolb-Proust* metadata is transformed to LOD, we will modify the search and presentation layers in XTF (managed through XML stylesheets) to expose and make use of the LOD data model and URIs. (Released via GitHub, with licensing as above.)
- ***Extensions to special collection image services to accommodate LOD***: As we transform legacy theater image descriptions into LOD, we will modify search and presentation layers of our image content management system (presently CONTENTdm, but we are investigating an alternative built atop Fedora 4[40] and the International Image Interoperability Framework,[41] and released via GitHub, with licensing as above).
- ***Prototype interface for viewing the social network of Proust: C***ode may not be easily adaptable, but will be useful to demonstrate proof-of-concept. (Via GitHub, as above.)

In addition to the work products listed above, we will disseminate outcomes through articles and conference papers / posters at venues such as Digital Humanities, the Joint Conference on Digital Libraries, Coalition for Networked Information, ASIS&T Annual, etc.

## 5.2. Detailed task breakdown and schedule of completion

### Task 1. Transforming special collections metadata into LOD, Nov. 2015 – Oct. 2016

- ***Extract & clean metadata***: Extract metadata records as currently stored in XTF and CONTENTdm servers; then clean, normalize and filter for entities.
- ***Register resources***: As necessary (i.e., where no pre-existing URI) register persistent URIs for items in collections using the UIUC Library's handle server.
- ***Initial reconciliation***: Adapt our existing name and subject entity reconciliation workflows to find URIs for strings that identify entities in extracted metadata.
- ***Pass 1 transform***: Identify additional mappings and adapt existing schema.org transformation workflows to create schema.org graphs for extracted metadata.
- ***Analyze pass 1***: Analyze transformed metadata to identify areas requiring collection-specific reconciliation and/or transformation and amend workflows.
- ***Register unreconciled names***: Register persistent URIs for unreconciled name entities in *Kolb-Proust Archive* using the UIUC Library's handle server and implement service to return appropriate RDF when these URIs are de-referenced.
- ***Pass 2 transform***: Execute a final, definitive reconciliation and transformation of the extracted metadata records into schema.org LOD graphs (descriptions).
- ***Implement triple-store***: Implement a fresh instance of Virtuoso triple-store to facilitate storage and efficient serialization of the transformed descriptions. This triple store will allow fast querying by URI, e.g, for use when connecting from VuFind displays (Task 2) to special collection items.

[40] https://wiki.duraspace.org/display/FF/Fedora+Repository+Home

[41] http://iiif.io/

- ***Author White Paper 1***: describing reconciliation & transformation and lessons learned & providing advice re transforming special collections metadata to LOD.

**Task 2. Special collections LOD as an entrée to general collections, Jan. – Dec. 2016**

- ***Baseline user test***: Conduct a small sample of user tests of the existing interface to establish a baseline assessment of its functionality and utility.
- ***Add links to library resources mentioned***: Modify interfaces to integrate links to digitized bibliographic resources mentioned (especially Kolb-Proust Archive), e.g., to books, scanned newspaper articles, scanned Proust letters, etc.
- ***Add links from WorldCat Identities***: Modify interfaces to integrate links & info from WorldCat Identities pages (or successors) into resource splash screens.
- ***Post-LOD user test***: Conduct a small sample of user tests of the modified interfaces to preliminarily identify impacts and changes in user response.
- ***Implement RDF-VuFind***: Implement OCLC's VuFind integrated with WorldCat Discovery API online public catalog interface and modify so it can use URIs present in WorldCat schema.org graphs to link through our Virtuoso server to our digitized special collections resources.

**Task 3. Descriptive enrichment and enhanced discovery, July 2016 – June 2017**

- ***Add RDFa***: Serialize LOD descriptions created in Task 1 in RDFa 1.1. Lite and integrate into resource splash pages (creating splash pages as needed); monitor before and after referrer data and traffic to collections for indication of impact.
- ***Identify non-library Web sources***: Algorithmically identify Web resources (e.g., Wikipedia articles) to which special collection items are relevant.
- ***Linking to non-library Web sources***: Modify special collection item displays to link to relevant Web resources (e.g., Wikipedia, Internet Broadway Database, ...)
- ***Enrich with triples discovered***: As possible (e.g., for Web sources having LOD descriptions), extract useful triples from non-library sources to enrich our LOD descriptions of digitized special collections; add triples to Virtuoso triple store.
- ***Proof-of-concept linking to special collections***: Demonstrate a prototype workflow for identifying algorithmically and manually adding to Wikipedia (at least) links to our digitized special collections described by LOD.
- ***Adding links by annotation***: Implement an annotation service allowing users to propose additional relevant links to add to digitized special collections LOD.

***Task 4. Visualizing social network of Marcel Proust*, Jan. – June 2017**

- ***Extract social network graph data***: From Kolb-Proust LOD extract name relationship and co-occurrence metadata needed to create social network graph.

- ***Select network graph tool***: Assess candidates and select visualization tool for creating and displaying interactive social network graph.
- ***Create proof-of-concept social network graph visualization interface***.
- ***Add relationship annotation feature***.

**Task 5. Project Management, Nov. 2015 – June 2017**

- ***Ongoing project oversight & management***.
- ***Prepare and submit Year 1 interim report.***
- ***Convene local scholar panel***: Panel will provide reaction to and feedback regarding first half progress and mockups illustrating plans for second half.
- ***Convene advisory committee meeting***: Will provide feedback on outcomes and advise on dissemination strategy, ways to maximize impact, next steps.
- ***Author White Paper 2***: describing strategies for using LOD descriptions of digitized special collections to connect collections to the larger semantic Web.
- ***Prepare and submit final report***.

## 5.3. Schedule of Completion

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

**SCHEDULE OF COMPLETION (page 1 of 2)**

| | | | | **Year 1** | | | | | | | | | | | | **Year 2** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | **1 Nov 2015 - 31 Oct 2016** | | | | | | | | | | | | **1 Nov 2016 - 30 June 2017** | | | | | | | |
| | | | | no | de | ja | fe | mr | ap | my | je | jl | ag | se | oc | no | de | ja | fe | mr | ap | my | je |
| | ***Task*** | ***Sub-Task Activity*** | ***Personnel*** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1. | **Transforming Special Collections Metadata to LOD** | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | |
| | | *Extract & clean metadata* | MH, CS, JS, GH | ■ | ■ | ■ | | | | | | | | | | | | | | | | | |
| | | *Register resources* | CS, JS, GH | | | ■ | | | | | | | | | | | | | | | | | |
| | | *Initial reconciliation* | JS, AH | | | ■ | ■ | | | | | | | | | | | | | | | | |
| | | *Pass 1 transform* | JS, GH | | | | ■ | ■ | | | | | | | | | | | | | | | |
| | | *Analyze pass 1* | TC, MH, CS, PD, AH | | | | | | ■ | ■ | | | | | | | | | | | | | |
| | | *Register unreconciled names* | JS, AH | | | | | | | | ■ | | | | | | | | | | | | |
| | | *Pass 2 transform* | JS, GH | | | | | | | | ■ | ■ | | | | | | | | | | | |
| | | *Implement triple-store* | JS, AH | | | | | | | | | | ■ | | | | | | | | | | |
| | | *Author White Paper 1* | All | | | | | | | | | | | ■ | ■ | | | | | | | | |
| 2. | **Special Collections LOD as an Entrée to General Collections** | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | |
| | | *Baseline user test* | CS, PD, AH, GH | | | ■ | | | | | | | | | | | | | | | | | |
| | | *Add links to library resources mentioned* | TC, MH, PD, GH | | | | | | ■ | ■ | | | | | | | | | | | | | |
| | | *Add links from WorldCat Identities* | TC. MH, PD, GH | | | | | | | | ■ | ■ | | | | | | | | | | | |
| | | *Implement RDF-VuFind* | TC, JS, PD, AH | | | | | | | | | | ■ | ■ | ■ | ■ | | | | | | | |
| | | *Post-LOD user test* | CS, PD, AH, GH | | | | | | | | | | | | | | ■ | | | | | | |
| 3. | **Descriptive Enrichment & Enhanced Discovery** | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| | | *Add RDFa* | TC, MH, JS, AH | | | | | | | | | ■ | ■ | | | | | | | | | | |
| | | *Identify non-library Web sources* | PD, CS, AH, GH | | | | | | | | | | | | ■ | ■ | ■ | | | | | | |
| | | *Linking to non-library Web sources* | PD, CS, AH, GH | | | | | | | | | | | | | | ■ | ■ | ■ | | | | |
| | | *Enrich with triples discovered* | TC, MH, JS, GH | | | | | | | | | | | | | | | | ■ | ■ | ■ | | |
| | | *Proof-of-concept linking to special collections* | PD, CS, AH | | | | | | | | | | | | | | | | | | ■ | ■ | |
| | | *Adding links by annotation* | TC, JS, AH | | | | | | | | | | | | | | | | | | | ■ | ■ |

| **Code** | **Full Name / Role** |
|---|---|
| TC | Timothy Cole / PI |
| MH | Myung-Ja Han / co-PI |
| CS | Caroline Szylowicz / co-PI |
| JS | Janina Sarol / Developer |
| PD | Post-Doc Fellow |
| PC | Project Coordinator |
| AH | PhD RA/Acad. Hourly |
| GH | Grad. Hourly, MS Level |

| | Exploring the Benefits for Users of Linked Open Data for Digitized Special Collections | | | Year 1 | | | | | | | | | | | | Year 2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SCHEDULE OF COMPLETION (page 2 of 2) | | | 1 Nov 2015 - 31 Oct 2016 | | | | | | | | | | | | 1 Nov 2016 - 30 June 2017 | | | | | | | | | | |
| | | | | no | de | ja | fe | mr | ap | my | je | jl | ag | se | oc | no | de | ja | fe | mr | ap | my | je | | | |
| | *Task* | *Sub-Task Activity* | *Personnel* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | | | |
| 4. | Visualizing the Social Network of Marcel Proust | | | | | | | | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | **Code** | **Full Name / Role** |
| | | *Extract social network graph data* | TC, CS, JS, PD, AH | | | | | | | | | | | | | | | ▒ | ▒ | ▒ | | | | | TC | Timothy Cole / PI |
| | | *Select network graph tool* | PD, JS | | | | | | | | | | | | | | | | | ▒ | | | | | MH | Myung-Ja Han / co-PI |
| | | *Create social network graph visualization interface* | CS, PD, JS, AH | | | | | | | | | | | | | | | | | ▒ | ▒ | | | | CS | Caroline Szylowicz / co-PI |
| | | *Add relationship annotation feature* | TC, JS, AH | | | | | | | | | | | | | | | | | | | ▒ | ▒ | | JS | Janina Sarol / Developer |
| | | | | | | | | | | | | | | | | | | | | | | | | | PD | Post-Doc Fellow |
| 5. | Project Management | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | PC | Project Coordinator |
| | | *Ongoing project oversight & management* | TC, MH, CS, PC | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | ▒ | | AH | PhD RA/Acad. Hourly |
| | | *Prepare and submit Year 1 interim report* | TC, PC | | | | | | | | | | | | | ▒ | | | | | | | | | GH | Grad. Hourly, MS Level |
| | | *Convene local scholar panel* | CS, PD, PC | | | | | | | | | | | | | ▒ | | | | | | | | | | |
| | | *Convene advisory committee meeting* | TC, PD, PC | | | | | | | | | | | | | | | | | ▒ | | | | | | |
| | | *Author White Paper 2* | All | | | | | | | | | | | | | | | | | | | ▒ | ▒ | | | |
| | | *Prepare and submit final report* | TC, PC | | | | | | | | | | | | | | | | | | | | | ▒ | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | |

## 6. Related Work

The desire for greater precision and connectedness in the way libraries describe and curate their growing digital holdings drives library interest in Linked Open Data. LOD is inherently distributed and collaborative which resonates well with library approaches to cataloging, metadata creation, resource sharing, and union catalogs (among other things), suggesting that LOD is, in some ways at least, a natural fit for libraries. However, LOD also represents a break with traditions of library descriptive cataloging and bibliographic control. Despite having in common themes of shared adoption and distributed and collaborative approaches to resource description, LOD eschews many traditional library cataloging practices in favor of a wholesale movement away from the string descriptions found in MARC records toward, machine-actionable data that relies on URIs, is subject to ongoing change and updating, and depends on semantics that are defined by a broad, Web-based community including, in addition to librarians, Web search engine vendors, publishers, other commercial entities, Web developers, etc. LOD and the Semantic Web are here to stay, but the cost/benefit ratio of LOD in a variety of library contexts remains to be demonstrated.

Our proposed project will provide concrete evidence in regard to the challenges and benefits of LOD for digitized special collections. To better understand in other library contexts both the potential benefits of LOD and its likely cost in time and effort, multiple other research projects and prototype implementation initiatives are currently in the planning stages or already ongoing in parallel. We are tracking these other initiatives and plan to take full advantage of synergisms wherever possible.

*Europeana*[42] is an example of a large-scale cultural heritage aggregation that has in the past relied on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[43] and descriptive semantics developed by the Dublin Core Metadata Initiative.[44] By bringing together descriptive metadata records from cultural heritage institutions across Europe, Europeana achieves a kind of contextual mass and has helped raise the visibility and utility of many digitized special collections held by these institutions. Similar projects in the US, such as the IMLS-funded *Digital Content and Collections* project (conducted here at Illinois from 2002 - 2013[45]) and now the *Digital Public Library of America (DPLA)* project,[46] also facilitate discovery and access for digitized special collections resources. While these projects have begun to address the issue of contextual mass, they are built largely on library-specific technologies and traditional string-based approaches to resource description. Europeana realized early on the potential advantages of using URIs and RDF for resource description. Accordingly they developed an RDF-compatible Europeana Data Model (EDM)[47] which has since been adopted in large part by DPLA. Initially EDM was primarily a behind-the-scenes model for description that helped to connect and collate resources within the Europeana

[42] http://www.europeana.eu/
[43] http://www.openarchives.org/OAI/openarchivesprotocol.html
[44] http://dublincore.org/
[45] http://imlsdccweb.grainger.illinois.edu/
[46] http://dp.la/
[47] http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation

metadata aggregation. EDM remains a library-centric model for description, more so than the semantics of schema.org on which we will rely.

For example, EDM's reliance on the Open Archives Initiative-Object Reuse and Exchange data model[48] is well suited for the task of merging metadata records within a large scale aggregation, where preserving metadata records gathered together from multiple libraries and archives and building a sort of provenance view of who said what about a resource is critical, but it is not necessary for the interconnectivity we want to explore, where the goal is simply to connect items in one library's collection with items elsewhere. Additionally, the primary feature of our work will be its focus on enriching metadata at a stage where local domain expertise is available. We believe that this expertise, much of it vested in special collection librarians, can be leveraged to create some linked data more efficiently and more accurately than can be done once metadata is aggregated. As it evolves over the long term, we can expect that the EDM and the closely-related DPLA Metadata Application Profile (MAP) will facilitate broader connectedness. The challenge will be to develop the EDM and DPLA-MAP in the right way (i.e., presumably in the direction of schema.org) and to determine which entity reconciliations (conversions of strings to URIs) are most efficiently done post-aggregation and which need to be done by metadata providers prior to aggregation. Currently DPLA is dependent on URIs included in metadata furnished by partner libraries, but they have begun to explore which kinds of entities they can reconcile post-aggregation.[49] Our work is complementary not redundant to the work being done by Europeana and now DPLA. Our findings regarding the identification and reconciliation of entities in digitized special collections metadata will help inform the larger scale work being done by Europeana and DPLA. We also anticipate that for the kinds of special collections being considered in this project, our findings will reveal more about the trade-offs of doing entity reconciliation and transformation to LOD at the curating library versus doing this at the metadata aggregator, i.e., Europeana or DPLA. Because digitized special collections tend to be unique and involve a high degree of active curation, we anticipate that doing entity reconciliation and transformation to LOD collection by collection at the curating library holds greater potential for more broadly enhancing the connectedness of digitized special collection items.

Similarly our proposed project will complement the ongoing work with BIBFRAME[50] by the Library of Congress et al., and more specifically the work on the current Linked Data for Libraries project (LD4L – Cornell, Stanford and Harvard),[51] the IMLS-funded BIBFLOW project[52] (a collaboration between the University of California at Davis and Zepheria), and the potential Linked Data for Production Cataloging (LD4P) project currently in development by Stanford University Libraries and five other major partners. These projects focus largely on the challenges of transitioning large technical service units and library vendors from MARC to the RDF, LOD-friendly BIBFRAME model of resource description. Here again our focus on

[48] http://www.openarchives.org/ore/1.0/toc

[49] "At the moment DPLA's plans for LOD include associating URIs that are already present in the records we get from our partners, as well as looking up and populating URIs for place names when we can." Retrieved from http://dp.la/info/2015/03/05/dpla-map-version-4-0/ on 31 May 2015.

[50] http://www.loc.gov/bibframe/

[51] https://wiki.duraspace.org/pages/viewpage.action?pageId=41354028

[52] http://www.lib.ucdavis.edu/bibflow/

custom legacy special collection metadata formats (rather than MARC) as input and schema.org (rather than BIBFRAME) as output differentiates our proposed project. As outlined by Jean Godby and Ray Denenberg, BIBFRAME and schema.org have clear differences in development and anticipated use. While libraries are experimenting with schema.org as a "vehicle for exposing library metadata to Web search engines in a format they seek and understand," BIBFRAME was developed as "linked data alternative to MARC" that uses the elements used in MARC as its predicates and FRBR as its data model (Godby and Denenberg 2015, p. 4). For many of the same reasons that library special collections tend not to be described using MARC records, it makes sense to consider non-BIBFRAME semantics when describing digitized special collections. (Conversely, it can be argued that it makes more sense to use BIBFRAME for describing digital analogs to general collection library resources that would normally be described using MARC.) Schema.org also makes sense given our interest in search engine uptake and connecting items in digitized special collections to non-library, non-bibliographic resources, none of which will be described by BIBFRAME. Ultimately, however, neither semantic set is wholly sufficient for describing items in library digitized special collections. Semantic extensions will be needed to use either with digitized special collections. It also may be (more study and analysis will be required as more examples of Library LOD become available) that the difference between schema.org and BIBFRAME can be largely mitigated algorithmically. The work previously cited by Nurmikko-Fuller et al. (2015) suggests this possibility. In addition to the BIBFRAME versus schema.org distinction, the LD4L, LD4P and BIBFLOW are all concerned with transforming or creating metadata at scale, whereas our focus, as is often the case with library special collections, allows for more specialized, smaller-scale treatments. This is not to say that there are no synergisms. In particular some of the use cases we will explore, e.g., connecting from special collections to general collection metadata and vice versa, overlap with LD4L use cases (e.g., 4.1: Identifying Related Works[53]). We are already in close touch with the ongoing LD4L project and plan to stay equally well in touch with the proposed LD4P project.

On the whole our proposed project is most synergistic with ongoing OCLC Research investigations of LOD.[54] We plan to make extensive use of OCLC LOD services and like OCLC we plan to work primarily with schema.org semantics. Scale and a less bibliocentric perspective differentiates our project from the main thread of OCLC's research into LOD. Our project will take a closer look at digitized special collection-specific metadata and user needs.

Collectively, our project and all of the projects listed above will add greatly to the community's understanding of the state of the art for reconciliation and transformation of legacy metadata and cataloging records to LOD. Data about user first reactions to LOD-based services will also be generated by several of these projects, including our proposed project. Much less well addressed is the next logical question – how easy or hard will it be to maintain distributed, dynamic LOD resource descriptions? This is a potential question for further research if outcomes from our project and those described above are sufficiently positive.

[53] https://wiki.duraspace.org/display/ld4l/Use+Case+4.1%3A+Identifying+related+works

[54] http://www.oclc.org/research/themes/data-science/linkeddata.html?urlm=168906

## 7. Intellectual Property

This project will be subject to the Foundation's intellectual property policy.[55] All software deliverables will be made available to the non-profit educational, scholarly and charitable communities on a royalty-free basis under an open source license allowing free redistribution, derived works, etc.; all pre-existing software that will be embedded in or used to derive deliverables is already made available under appropriate open source license. Reports and Web-posted deliverables will be made freely and openly available to the non-profit educational, scholarly and charitable communities on a royalty-free basis, under a Creative Commons Attribution license permitting non-commercial use and modification.

The scripts used for reconciliation and transformation will be released under the University of Illinois / NCSA Open Source license.[56] Local (UIUC) authority records about Kolb-Proust names that cannot be reconciled with a national or international authority, will be made available freely and openly with reuse allowed under the Open Data Commons Public Domain Dedication and License (PDDL).[57] We do not of course control licensing of other LOD reconciliation services, but wherever possible we will prefer reconciliation services that license their data in accord the Open Data Commons PDDL, the Open Data Commons Attribution License (ODC-By),[58] or an equivalent license. Some software deliverables we release (all under the Illinois / NCSA Open Source license as discussed above) will be designed to work with XTF or with CONTENTdm. XTF is covered by three Open Source license (Mozilla, BSD and Apache).[59] CONTENTdm is not available under Open Source license; however, any code we write will work with the publicly documented CONTENTdm API[60] and so will be useable by any institutions running CONTENTdm. The visualization interface that will be developed as part of Task 4 will depend only on Open Source libraries.

## 8. Sustainability

All White Papers, scripts, codes and data developed by this project will be maintained on a project Website and / or in a project-specific GitHub repository linked from that Website. The project Website will remain operational and publicly accessible through at least 2020.

LOD descriptions themselves (such as will be created as part of Task 1 and integrated into splash screens as part of Task 3) require little effort to maintain. We can anticipate that by the end of this project the three collections being used for this research (all of which are closed and static as regard items in each collection) will be better described and will be more easily indexed by Web search engines (given the use of schema.org semantics in our LOD descriptions). This represents a positive persistent and durable outcome of this project. However, effort is required to make use of the links contained in LOD descriptions, and over time LOD best practices will evolve and new LOD-compatible authorities will be created or augmented, suggesting a need to continue to add to LOD descriptions. In considering the sustainability of this project as an

[55] http://www.mellon.org/about_foundation/policies/AWMF-IP-October-2011.pdf/at_download/file
[56] http://opensource.org/licenses/NCSA
[57] http://opendatacommons.org/licenses/pddl/
[58] This is the licensed used by VIAF: http://opendatacommons.org/licenses/by/1.0/
[59] http://xtf.cdlib.org/download/
[60] http://www.contentdm.org/help6/custom/customize2a.asp

initiative to encourage broader use of LOD for digitized special collections, it also is important to recognize that the cost in time and effort and the ultimate utility of LOD descriptions in various library contexts and in regard to different library use cases is still being determined. We believe based on evidence to date that the cost/benefit ratio for digitized special collections is strongly positive, but one of the goals of this project is to provide (through the two white papers that will be part of the output from this project) some early, qualitative inputs to any subsequent analyses that might try to assess cost/benefit ratio of adopting LOD for library digitized special collections. .

So ultimately whether LOD descriptions are maintained, and whether the Proust social network visualization proof-of-concept interface and modifications to our special collections content access system interfaces serve as prototypes for production service enhancements, will depend in part on result from this project, as well as results from synergistic projects on LOD being undertaken elsewhere in the Library community. The opportunity is certainly there. UIUC Library alone has more than 25 special collections of digitized images. Though each collection is distinct, these collections (including *Motley* and *Portraits of Actors*) share many traits, including substantive overlap in metadata design and similarity in the interfaces and systems used for search and discovery. All scripts and code developed on this project will be written to maximize adaptability and to facilitate long-term sustainability. Assuming positive outcomes from the research undertaken, we are confident that the outcomes of this project will be sustained, will be reused at this library and others, and will contribute to the community's understanding of LOD and how it can be used to enhance the connectedness and usefulness of digitized special collections.

## 9. Reporting

Since the proposed project will span 20 months, from November 1, 2015 to June 30, 2017, we anticipate the submission of two formal project reports (i.e., one Interim Report to be submitted before the end of January 2017, i.e., within 90 days after the grant project start date 1 year anniversary, and one Project Final Report to be submitted within 90 days of the grant project's end date (30 June 2017). The reports will include narrative commentary on the activities, successes and challenges of the project. In particular the Interim report will describe results of reconciliation and transformation workflows; our goal and metric for success will be the creation of schema.org LOD graphs for all items in all three collections by the end of year 1. White Paper 1, describing the lessons learned in developing workflows, will be attached to our Interim Report. The Project Final Report will detail the results of Tasks 2, 3 and 4; our goal and metric for success will be working prototype search and discovery interfaces with added linking (Tasks 2 and 3) and a demonstrable dynamic and interactive visualization interface of Proust's social network (Task 4). The Project Final Report will include White Paper 2 as an attachment. Both reports also will discuss grant expenditures for the period covered in conjunction with the official budgetary accounting provided by the University of Illinois grants and contracts accounting office. The reports will be prepared by Timothy Cole (Project PI) in collaboration with the Project Coordinator, co-PIs and Project Post-doc. Cole will have the ultimate responsibility for timely completion and submission of these reports.

## 10. Budget Narrative

NOT INCLUDED IN THIS COPY

## 11. Bibliography

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. J. (2001). *Scholarly work in the humanities and the evolving information environment* (No. 104). Washington, D.C.: Digital Library Federation, Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub104/pub104.pdf

Cole, Timothy W., Myung-Ja Han, Mara R. Wade, and Thomas Stäcker. (2013). Linked OpenData & the OpenEmblem Portal. In *Digital Humanities 2013: Conference Abstracts*, Lincoln, NE: Center for Digital Research in the Humanities, University of Nebraska-Lincoln: 146-150. Retrieved from http://dh2013.unl.edu/abstracts/ab-359.html

Ciula, A., & Lopez, T. (2009). Reflecting on a dual publication: Henry III Fine Rolls print and web. *Literary and Linguistic Computing*, *24*(2), 129–141. http://doi.org/10.1093/llc/fqp007

Dooley, Jackie & Luce, K. (2015). Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives, Executive Summary. In *Making Archival and Special Collections More Accessible*, OCLC Research. Retrieved from http://www.oclc.org/content/dam/research/publications/2015/oclcresearch-making-special-collections-accessible-2015.pdf

Finegold, M.A., Warren, C., Shalizi, C., Shore, D., & Wang, L. Six Degrees of Francis Bacon [long paper abstract]. (2013.) In *Digital Humanities 2013: Conference Abstracts*. Lincoln, NE: Center for Digital Research in the Humanities, University of Nebraska-Lincoln. Retrieved from http://dh2013.unl.edu/abstracts/ab-417.html

Godby, Carol Jean and Ray Denenberg. (2015). *Common Ground: Exploring Compatibilities between the Linked Data Models of the Library of Congress and OCLC*. Retrieved from http://www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015-a4.pdf

Gradmann, S. (2014). From containers to content to context. Journal of Documentation, 70(2), 241–260. Retrieved from http://doi.org/10.1108/JD-05-2013-0058

Green, Harriett E. and Angela Courtney. (2015). Beyond the Scanned Image: A Needs Assessment of Scholarly Users of Digital Collections. *College & Research Libraries* (In Press)

Lynch, Tom J. (2014.) Social Networks and Archival Context Project: A Case Study ofEmerging Cyberinfrastructure. *Digital Humanities Quarterly*, 8:3 2014. Retrieved from http://www.digitalhumanities.org/dhq/vol/8/3/000184/000184.html

Maron, Nancy L. and Pickle, Sarah. (2013). Appraising our Digital Investment: Sustainability of Digitized Special Collections in ARL Libraries, Association of Research Libraries and Ithaka S+R, 2013, 49 pages. Retrieved from http://sr.ithaka.org/research-publications/appraising-our-digital-investment

Mullin, Michael. 1996. *Design by Motley*. Newark: University Of Delaware Press; 1996. Print.

Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J., Maden, C., Cole, T., Fallaw, C., Senseney, M. & Downie, J. S. (2015). Building complex research collections in digital libraries: A survey of ontology implications [short paper]. In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (Knoxville, TN, 21-25 July 2015): in press.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society for Information Science and Technology*, *56*(11), 1140–1153. http://doi.org/10.1002/asi.20204

Renear, A. H., & Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. Science, 325(5942), 828.

Palmer, C. L., Teffeau, L. C., & Pirmann, C. M. (2009). *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development*. Dublin, OH: OCLC Research and Programs. Retrieved from http://www.oclc.org/content/dam/research/publications/library/2009/2009-02.pdf

Palmer, Carole L., Zavalina, Oksana, Fenlon, Katrina. (2010). Beyond size and search: Building contextual mass in digital aggregations for scholarly use. In A. Grove (ed.) *Proceedings of the ASIS&T Annual Meeting*. (Pittsburgh, PA, Oct. 22-27). Retrieved from http://www.asis.org/asist2010/proceedings/proceedings/ASIST_AM10/submissions/213_Final_Submission.pdf

Rogers, B. M. (2008). The historical community and the digital future. Presented at the 3rd Annual James A. Rawley Graduate Conference in the Humanities. Imagining Communities: People, places, meanings., Lincoln, NE. Retrieved from http://digitalcommons.unl.edu/historyrawleyconference/26/

Russell, D.M., Grimes, C. (2007.) Assigned tasks are not the same as self-chosen Web search tasks. Proceedings of the Annual Hawaii International Conference on System Sciences, art. no. 4076538.

Szylowicz, C., & Kibbee, J. (2004). The collaboration that created the Kolb-Proust Archive: Humanities scholarship, computing, and the library. In J. A. Inman, C. Reed, & P. Sands (Eds.), *Electronic collaboration in the humanities: Issues and options* (pp. 255-266). Mahwah, NJ: Lawrence Erlbaum.

Tibbo, H. (2003). Primarily history in America: How U.S. historians search for primary materials at the dawn of the digital age. *The American Archivist*, *66*(1), 9–50.