Title

Enhancing an OAI-PMH Service Using Linked Data: A Report from the Sheet Music Consortium

Journal
Journal of Library Metadata, 13(2-3)

ISSN
1938-6389 1937-5034

Authors
Davison, Stephen
Sugiyama, Yukari
McAulay, Elizabeth
et al.

Publication Date
2013-07-01

DOI
10.1080/19386389.2013.826067

Peer reviewed

Enhancing an OAI-PMH Service Using Linked Data: A Report from the Sheet Music Consortium

STEPHEN DAVISON, YUKARI SUGIYAMA, ELIZABETH McAULAY and CLAUDIA HORNING
University of California, Los Angeles, California, USA

Running title:

Enhancing an OAI-PMH Service Using Linked Data

This article discusses the metadata records aggregated by the Sheet Music Consortium using the Open Archives Initiative Protocol for Metadata Harvesting. The Consortium's web site serves as a union catalog for over a quarter of a million records, harvested from 27 providers. During a recent period of revision, the Consortium decided to experiment with normalizing data and publishing authority data related to sheet music. A pilot project focused on sheet music publishers is discussed and the results are presented.

Keywords: cataloging, data normalization, linked data, metadata harvesting, semantic web

INTRODUCTION

The Sheet Music Consortium (SMC) is a collaboration of several universities and other sheet music repositories to publish an online metadata catalog of sheet music.[1] The Consortium began in 2002, inspired by the opportunities for metadata sharing offered by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). In 2013, the Consortium released a revised web site and set of services to make it easier for contributors to submit metadata using OAI-PMH. The development of these new services has attracted over a dozen additional institutions, and this expansion means that the SMC's metadata now comes from a wider variety of systems and is created according to different schema and descriptive standards. While this wealth of metadata ensures that the SMC is a significant resource for the search and discovery of sheet music, the diversity of the metadata contributions results in a non-normalized data set.

During the recent, major revision of the SMC, the two lead institutions--the University of California, Los Angeles (UCLA) and Indiana University (IU)—grappled with the well understood problem of metadata disparity. In this article, we review the current state of the SMC metadata, discuss why publishing some metadata as Linked Open Data (LOD) might be meaningful, and present our pilot project to publish information related to sheet music publishers from the SMC records.

[1] The Sheet Music Consortium's web site available at: http://digital2.library.ucla.edu/sheetmusic. "Sheet music" as understood by the community of music librarians: "... is best described as single sheets printed on one or both sides, folios (one sheet folded in half to form four pages), folios with a loose half-sheet inserted to yield six pages, double-folios (an inner folio inserted within the fold of an outer folio to make eight pages) and double-folios with a loose half-sheet inserted within the fold of an inner folio to produce ten pages" (http://library.duke.edu/digitalcollections/hasm/about/#define). Generally, a piece of sheet music contains a single popular song, aria, or piano piece, and is intended for domestic rather than for professional use.

At present, the Consortium web site provides access to 228,463 items from 27 institutions. These institutions include 24 universities, one public library and academic library collaboration, and two national libraries--the Library of Congress and the National Library of Australia. The SMC web site has been well received, appears on numerous lists of resources, and is the most heavily used collection hosted by the UCLA Digital Library. The service has appeared in Martha Brogan's reviews of OAI-PMH services, *Digital Library Aggregation Services* (2003), and its companion/update, *Contexts and Contributions: Building the Distributed Library* (2006). The vision of the Consortium service was to be a union catalog for sheet music—to be *the* place to go to discover online sheet music by pulling together the metadata records of several significant collections and offering them for search and browse from one interface.

While the SMC has enjoyed both early and continuing recognition as a useful service, it became apparent that the wide range of metadata practices, even among relatively similar institutions, significantly limited the services that the SMC was able to offer. Those findings inspired a second phase of development (2007-2013), in which the Consortium greatly expanded its reach. The SMC added capacity to harvest metadata records in MODS and qualified Dublin Core (in addition to original simple Dublin Core harvesting) and mapped all the incoming metadata to a standard MODS format. The SMC also implemented an OAI Static Repository Gateway and offered conversion tools to help organizations publish their metadata in a harvestable format without building or maintaining a data provider (see Table 1). At the same time, SMC metadata experts developed more guidance for potential participants and published new metadata guidelines for digitized sheet music. During this same period, the Consortium also improved the end user interface (i.e., the "service provider" in OAI terminology). The new web site offers improved browsing and searching options, including a chronological browse facilitated by normalized date metadata. Finally, the SMC added a function for users to add structured metadata and comments. As participation increased and the web site functionality improved, the divergence of metadata practices utilized for sheet music description became more apparent. To improve the quality of both searching and browsing, the Consortium wanted to provide more metadata normalization, but did not know how to disseminate the normalized values without asking contributors to update their records. This request would increase the effort required for effective aggregation—the exact opposite of the impetus for the improvements on which the Consortium had focused over the past several years.

Table 1 Schemas and Workflows used to harvest records for the Sheet Music Consortium.

SCHEMA	# institutions	# records
Dublin Core	14	98,199
Qualified Dublin Core	9	26,466
MODS	4	103,798
WORKFLOW		
Direct harvesting using the OAI protocol	25	206,026
Harvesting the metadata via the Static Repository Gateway	1	2,222
Manual extract of MARC records from an integrated library system and mapping to MODS and ingest	1	20,215

At this point, our small group at UCLA decided to try a new activity: publishing LOD. We realized we could run normalization routines and then publish the results as LOD so that others could use the data. LOD is being discussed excitedly in many communities, and by running a pilot project we hoped to learn how we could contribute data based on the Consortium’s records. After identifying several fields that could be fruitfully normalized and published, we decided to conduct our pilot using metadata about publishers. We recognized that there were many similarities between publishers and corporate names maintained in name authority files, while at the same time names of publishers were completely outside the authority workflow. Therefore, we recognized that the data we might publish would be new to the broadening field of authority control.

LITERATURE REVIEW

The SMC began as an OAI-PMH focused service provider, and while significant aggregations have been created using OAI-PMH—most notably OAIster—the PMH method was recognized early on as not fulfilling its stated potential. This shortcoming was in part because of the difficulty of aggregating metadata effectively and in part because the ease of using the Protocol had been overestimated at its beginning (Shreeves, Habing, Hagedorn, & Young, 2005; Lagoze et al., 2006). Reports on strategies to ameliorate aggregation problems soon appeared. In his report for the California Digital Library about improving OAI-PMH results, Tennant (2004) advised that service providers enact normalization routines and additional metadata parsing on harvested records. Similarly, Hagedorn's (2003) initial report on the OAIster project outlined desired improvements for the service, including data normalization and additional processing of incoming records. For the most part, these problems were not intrinsic or unique to OAI-PMH implementers, but surfaced during OAI work because the new protocol allowed for aggregations of records that had previously been siloed. In fact, general metadata problems are so commonly observed no matter what the platform that one recent article focused merely on grouping the problems into categories (Yasser, 2011). As a result of these early findings, though, some service providers soon invested significant staff time into developing methods for normalization and maintenance of harvested metadata in order that their aggregations were useful portals for end-users.

During the same period the OAI was started, Tim Berners-Lee introduced the concept of the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001). Several years later, the concept of Linked Data was proposed as a mechanism to enact the Semantic Web (Bizer, Heath, Idehen, & Berners-Lee, 2008). Recently, at some national levels, libraries, archives and museums have begun to reconceive of their metadata and cataloging activities in terms of Linked Data (Dunsire, Harper, Hillmann, & Phipps, 2012). Significant initiatives are underway at the Library of Congress; the national libraries of Sweden, Germany, France and Great Britain; and OCLC (Dunsire et al., 2012). The first instances of published library Linked Data are authority values, but now there is also a growing body of linked bibliographic data (Harper, 2012).

Publishing authority data is a normal practice for libraries, especially national libraries, and it makes sense to begin forays into linked data with this special and rich set of data that libraries consistently maintain. Indeed, the concept of name authority control has been an important one in the library community for many decades. Authority control has allowed libraries and users to disambiguate similar headings, to collocate works by the same author or on the same subject, to execute precise and accurate searches, and to navigate between related headings (Harper & Tillett, 2007). However, efforts to create an international authority value for each entity foundered because cultural differences led to disagreements on canonical forms of names (Dunsire et al., 2012). The solution to this impasse came when the Virtual International Authority File (VIAF) was formed and it eschewed a universal “preferred form” by collecting multiple authority values under a unique identifier (Dunsire et al., 2012). VIAF has been widely embraced as an effective tool and solution, and therefore Dunsire et al. argue that VIAF offers a viable model to emulate for publishing authoritative bibliographic data.

Based on this understanding of linked data in the library realm, we decided to begin with publishing authority data, but we wanted to publish data that was rare or unique and that would contribute to the linked data already published by other libraries and related organizations. We also wanted to work with data from the SMC without getting overwhelmed with the volume of information. From this perspective we decided to experiment using the names of publishers included in the Consortium records. Because publisher names are not usually controlled, we would be creating new authority data about the publishers, while at the same time it was simple to extract a few well known publishers' names and variants from the SMC data set.

Catalogers have worked cooperatively to standardize the creation of name authority records (NARs) for agents, such as corporate bodies, that are used as access points in bibliographic records. NARs for corporate bodies have long allowed the tracking of sequential forms of names that result from actions such as mergers and splits; under the new cataloging code, RDA (Resource Description & Access), new fields allow catalogers to add even more information to authority records. For example, corporate NARs may now contain elements indicating information such as the start/end dates of a corporate name, associated places and addresses, the field of business of the corporate body (e.g. music publisher), and the language(s) used in their publications.

Authority control for publishers would certainly be an asset within the SMC. However, there are a number of challenges when trying to use authority control in this context. Sheet music occupies a position in the spectrum of published materials somewhere between ephemera (brochures, advertising, instructional manuals, etc.) and formally published resources such as books or journals, with many variations in the ways that bibliographic information is presented.

In addition, publishers may describe or name themselves differently from piece to piece. Moreover, many of these sheet music resources are not being described by trained catalogers, or in accordance with existing cataloging rules. Finally, even if the traditional cataloging method was used to describe these materials, publisher names are not normally subject to authority control because they typically appear in parts of the MARC record (MARC21 fields 260 and 264) that are not considered access points. Therefore, they are normally transcribed as they are on the resource itself, which may or may not conform to the authorized form of the publisher's name—if an authorized form does exist. Thus, name variants abound. Frequently, in metadata records describing sheet music, the publisher field also includes the address of the publisher, which makes disambiguation and collocation difficult.

Despite these challenges, we believe authority records derived from the SMC metadata would be a rich source of consistent, machine-actionable information, particularly considering the new fields allowed under RDA. To take one example from the SMC, we can look at the music published by Oliver Ditson. In the 19th century, Ditson founded one of the most prominent music publishers in the United States. With his partner, Samuel H. Parker, Ditson founded the firm Parker & Ditson (1836-1842). After Parker's death, Ditson changed the name of the company to Oliver Ditson (1842-1856). When Ditson's employee John C. Haynes joined the firm as a partner, they changed the name to Oliver Ditson & Co. And finally, after Ditson's death, the company was renamed the Oliver Ditson Company. Somewhat unusually for sheet music publishers, all of these names have been established in the Library of Congress Name Authority File (LCNAF); the record for Oliver Ditson Company even includes a short note describing the history of this company. All of these authority records could be enhanced to include additional information, such as associated dates, as well as information describing the relationships between these entities (Marrocco, Jacobs & Krummel, n.d. [i]).

Despite the Oliver Ditson example, it may prove too much of a challenge to assert tight authority control over the forms of name for all of these entities. Dunsire, Hillmann, & Phipps (2012) suggest a different approach which could complement (or replace) the traditional library approach. Since Resource Description Framework (RDF)--a delivery mechanism for Linked Data--allows for the publication of "metadata statements" instead of complete bibliographic records, it might be preferable to "[link] identifiers for local data to an aggregator identifier without transforming or discarding the local data" (Dunsire et al., 2012, p. 165). In other words, instead of revising the contributed metadata record, we could add a link to an identifier that further describes the publisher. This approach provides a method for utilizing the bulk of metadata statements that exist in the Consortium metadata records without engaging in editing the actual contributed records. The Consortium data might be linked to aggregator identifiers by a manual process, or the process might be partially or completely automated based on dates of publication or other information.

THE CHALLENGES OF AGGREGATED METADATA

Karen Coyle (2009) contrasts the metadata "dumb down" and "smart up" strategies as follows:

> It's an unfortunate fact that many systems combine data from different sources using only the "dumb down" method, reducing the metadata to the few matching elements and resulting in the least rich metadata record possible. This results in a tremendous loss of data and an inferior user experience. The "smart up" method uses all or most of the data from the different sources, resulting in enhanced information. (Coyle, 2009, p. 10)

In this article, Coyle argues that combining information from different types of records rather than confining metadata to one schema means that the aggregation is additive rather than reductive. She discusses the way Linked Data allows for this by escaping the limitations of being a "record" of something rather than a constellation of information nodes that can lead in a variety of directions, and through it, users can more effectively move across systems and domains.

A typical bibliographic citation contains these elements: creators, title, edition, publisher, place and date. To these basic elements a typical bibliographic system, such as a library catalog, will add notes and subjects of various sorts. Sheet music metadata can include all these elements, with the exception of edition. The informality and variety of sheet music editions is such that publishers typically do not identify variations as "editions", except to indicate that an edition is a different format or for specific instruments.

SMC metadata reflects the diversity of the institutions from which it harvests. As an illustration of the variation in the description of a single piece of sheet music we have chosen one that appears in a number of the SMC source collections, described using all three of the representative metadata schema: MODS, DC, and qualified DC. The record from Indiana University (IU) is taken here as a reference, because it is the most complete and most closely tracks library cataloging practice as represented in the MODS schema (Example 1). The music is a song entitled "California and you" by the composer Harry Puck and the lyricist Edgar Leslie. The differences between the metadata in the records for "California and you" discussed below are typical of those found across the SMC collections. The title is represented in the Consortium by eight records from seven institutions, as shown in Table 2. In the discussion that follows these records are referred to by the abbreviation in the second column.

Example 1 MODS <titleInfo> segment for “California.” Indiana University. Entire MODS record available: http://n2t.net/ark:/21198/r2td9v76.

```
<mods:titleInfo>
    <mods:title>California</mods:title>
</mods:titleInfo>
<mods:titleInfo type="alternative" displayLabel="First line">
    <mods:title>Oh! You old pacific coast, oh! you land I love the most,</mods:title>
</mods:titleInfo>
<mods:titleInfo type="alternative" displayLabel="First line of chorus">
    <mods:title>Don't you remember California in September?</mods:title>
</mods:titleInfo>
```

Table 2 Sheet Music Consortium MODS records for: Harry Puck and Edgar Leslie, *California and you* (New York: Kalmar & Puck, 1914). Formatted displays of these records can be found at: http://n2t.net/ark:/21198/r2z60kz2

CONTRIBUTING INSTITUTION	SMC MODS RECORD URL (Abbreviation)
Mississippi State University	http://n2t.net/ark:/21198/r2jw8bs4 (MSU)
Johns Hopkins University	http://n2t.net/ark:/21198/r2pn93hv (JHU)
Indiana University	http://n2t.net/ark:/21198/r2td9v76 (IU)
University of Illinois at Chicago	http://n2t.net/ark:/21198/r2f769gf (UIC)
Duke University	http://n2t.net/ark:/21198/r29g5jrq (DU1) http://n2t.net/ark:/21198/r2x63jtp (DU2)
Southern Illinois University, Edwardsville	http://n2t.net/ark:/21198/r25q4t1d (SIUE)
York University [Canada]	http://n2t.net/ark:/21198/r21z4292 (YU)

Titles and creators

The combination of title and creators is generally sufficient to define a musical “work,” and it would be highly desirable to normalize and aggregate records by work. However, this normalization process is too large a project to undertake at this moment because the variation amongst the Consortium data is too great. Sheet music titles are harder to define than those for published books or articles. Titles may be taken from the first line of text, the first line of the chorus, be independent of the text, or if the song is from a larger work that work may be the designated title. Frequently, the same song may be published under different titles (Example 2), while, on the other hand, a variety of distinct songs may have the same title.

Example 2 Titles from other SMC data providers.

> California and you -- MSU, JHU, DU1, DU2, SIUE, YU
> California (and You) -- UIC
> Oh! you old Pacific coast [first line] -- JHU
> Don't you remember California in September? [first line of chorus] -- JHU
> Don't you remember California in September? -- SIUE (in a note field)

Most of the variations we find are due to the lack of discrimination between the various title elements found in simple Dublin Core records; a lack that that carries through when mapped to MODS.

Most sheet music repositories treat their collections as ephemera and therefore do not do any significant authority work. In addition, metadata sources for some repositories are archival finding aids or inventories, in which personal names may not even appear in inverted order. A sheet music publication is highly collaborative, and the hierarchy of creators as represented in the metadata may depend on the focus of the collecting institution. Although a music library will prioritize the role of the composer in formal cataloging through the personal name access point, historical societies and other types of special collections often organize collections by song topics, cover art, or other ways that reflect the collection’s social or historical significance, and prioritize creators accordingly without placing a high priority on making names easily browsable by last name.

Together, the titles and creators define a specific “work”--which is the logical focus for exposure as Linked Data--and the “distinct intellectual or artistic creation” in the FRBR model (International Federation of Library Associations and Institutions, 1998). It would be highly desirable for the SMC to normalize the names of creators and their works and to expose them as Linked Data. Although we have not done a quantitative study to determine the extent, there are clearly large numbers of creators represented in the SMC data for whom authority records do not exist. One strategy for normalizing names would be to follow the normal cataloging route of matching names to an existing authority file (e.g. Library of Congress) and establishing new headings where needed. However, the resources for this work are often not available at the point of initial description, and even less available at the point of aggregation. The very existence of the SMC, with our published guidelines for sheet music description are, we hope, encouraging more complete records at the point of description, and this process will make the need for normalization services at the point of aggregation less urgent.

Example 3 MODS <name> segment from “California” (Indiana University)

```
<mods:name>
        <mods:namePart>Puck, Harry</mods:namePart>
        <mods:role>
        <mods:roleTerm type="text"
                authority="marcrelator">Composer</mods:roleTerm>
        <mods:roleTerm type="code" authority="marcrelator">cmp</mods:roleTerm>
        </mods:role>
</mods:name>
<mods:name>
        <mods:namePart>Leslie, Edgar</mods:namePart>
        <mods:role>
        <mods:roleTerm type="text" authority="marcrelator">Lyricist</mods:roleTerm>
        <mods:roleTerm type="code" authority="marcrelator">lyr</mods:roleTerm>
        </mods:role>
</mods:name>
<mods:name>
        <mods:namePart>Dooley & Joyce</mods:namePart>
        <mods:role>
        <mods:roleTerm type="text"
            authority="marcrelator">Performer</mods:roleTerm>
        <mods:roleTerm type="code" authority="marcrelator">prf</mods:roleTerm>
        </mods:role>
    </mods:name>
```

The MODS name element, as cataloged by IU, is rendered as in Example 3. Of these three names, Harry Puck has a record in VIAF (http://viaf.org/viaf/7074281) linked to authority records at the Library of Congress and the National Library of Australia; Leslie Edgar has no authority record, but is listed in Wikipedia/DBpedia (http://en.wikipedia.org/wiki/Edgar_Leslie; http://live.dbpedia.org/page/Edgar_Leslie) with a link to a SMC search on his name; and the performers "Dooley and Joyce" do not have any discernible presence in the bibliographic universe outside this singular metadata record. Normalizing these names within the SMC and publishing them as Linked Data, including references to VIAF and Wikipedia/DBpedia, would improve linkages between these information resources, and publish some information about the lesser known creators associated with this piece of music, the performers Dooley and Joyce. It is impossible to tell what the particular importance of these performers might be in music and cultural history, but a Linked Data record would provide a hook to link any additional information with when it surfaces.

As with the titles, most of the variant forms of names stem from Dublin Core records because that standard requires names and roles to be concatenated (Example 4).

Example 4 Names from other SMC data providers

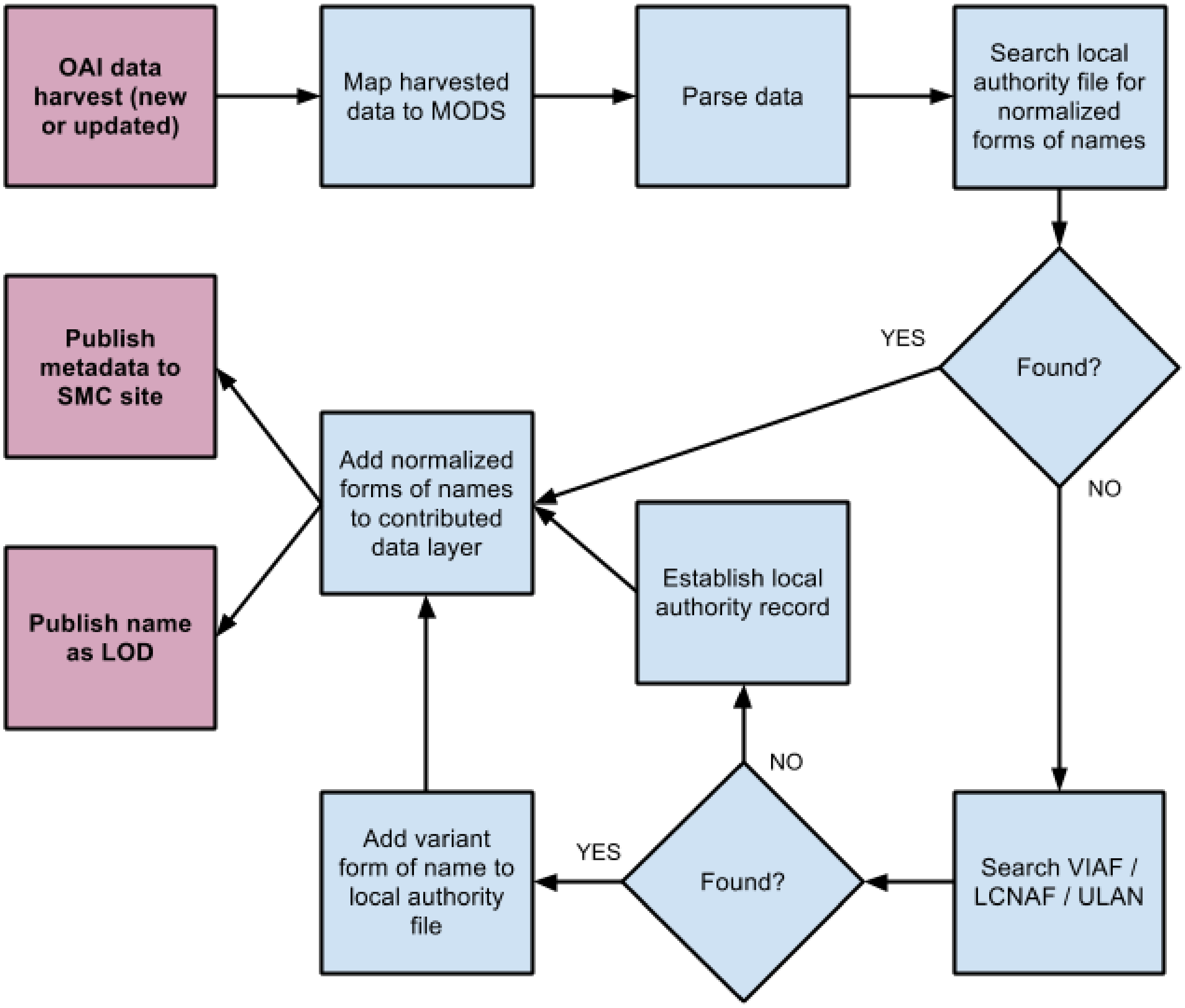
Harry Puck (composer) -- JHU
Puck, Harry, 1890-1964 -- SIUE
Puck, Harry [composer] -- YU
Leslie, Edgar [lyricist] -- YU
Edgar Leslie (lyricist) -- JHU
Richard Burton -- DU1, DU2
John Frew -- DU1, DU2

Additional names reflect interest in roles other than composer and lyricist. For instance, John Frew created the cover art, and Richard Burton is a performer whose portrait appears on the covers of some "editions."

These performers, like Dooley and Joyce above, have not been noted in any authority files, primarily because their recordings--if they did indeed record performances--are not present in library catalogs. Similarly, the names of the cover artists were never included in bibliographic descriptions until covers were digitized and described partly, and sometimes primarily, to expose the cover art. The exposure of artists' names as Linked Data would be as potentially beneficial as it would be for performers because it would begin documenting the existence and relationships that these supporting figures had. One potential synergy for SMC information published as Linked Data is with the archival community. As more names are published using the Encoded Archival Context (EAC) standard, the EAC records could be transformed into Linked Data and serve as a link between archival collections of musicians and artists and their published works in repositories.

Figure 1 illustrates a possible workflow for the republication of harvested metadata as LOD for names of creators represented in the SMC data. This is a traditional workflow in that it starts with names extracted from the published items themselves, includes reference to standard name authority files such as LCNAF, and the creation of shared or local authority records that are published in a variety of ways. The "Publish name as LOD" terminator in our process represents an addition to the existing network of name authority records, providing access to names appearing in sheet music metadata records that are otherwise difficult or impossible to find.

Figure 1 Proposed name authority workflow


OAI data harvest (new or updated)
Map harvested data to MODS
Parse data
Search local authority file for normalized forms of names
Found?
YES
NO
Publish metadata to SMC site
Add normalized forms of names to contributed data layer
Establish local authority record
Publish name as LOD
NO
YES
Found?
Add variant form of name to local authority file
Search VIAF / LCNAF / ULAN

.

Given the lack of resources to fully implement a program to create normalized names across the entire data set, the SMC has implemented a limited version of this workflow focusing on publishers, rather than creators' names, as will be discussed below. However, the establishing of name authority records is more familiar so it serves a useful purpose to discuss the workflow with respect to those.

Publication information

The relationships between the various people who created, published, performed and sold popular music in the nineteenth and early twentieth centuries are complex and quite different from the more formal relationships that exist between the authors and publishers of books. Composers, lyricists, performers, “song pluggers” (songwriter-pianists employed to plug songs in department stores), publishers, and music stores worked in a highly competitive and interrelated way. In many cases songwriters, song pluggers, and/or publishers exchanged or overlapped roles, most commonly when composers were also publishers and/or song pluggers. For example, Harry Puck is both a creator and publisher of the song “California and you” and Bert Kalmar is a co-publisher. Kalmar was also a longtime lyricist-collaborator with the composer Harry Ruby, with whom he formed the publishing company Kalmar & Ruby Music Corp. in Hollywood, California. This fluidity of roles demonstrates that the sheet music publishing process involved a network of individuals who collaborated with each other in a variety of ways. Documenting and presenting these relationships provides information both on the business of and the creation of popular songs in the United States during this period.

However, this multidimensional network is initially represented in single metadata records and is often described as a string of text that contains multiple pieces of information. For example. harvested records include publication information comprised of three elements: place, publisher, and date, either separate or strung together, depending on the source metadata schema. For example, see the IU MODS segment, which parses the publication statement into multiple fields (Example 5) and the DC example which keeps the data in one string (Example 6).

Example 5 MODS <originInfo> segment for “California” (IU)

```
        <mods:originInfo>
```

```
    <mods:place>
        <mods:placeTerm type="text">New York</mods:placeTerm>
    </mods:place>
    <mods:publisher>Kalmar & Puck Music Co. Inc.</mods:publisher>
    <mods:copyrightDate encoding="w3cdtf" keyDate ="yes">1914
    </mods:copyrightDate>
</mods:originInfo>
```

Example 6 Publication information from other SMC data providers.

New York : Kalmar & Puck Music Co. Inc., -- MSU
Bert Kalmar & Harry Puck -- DU1, DU2
Kalmar and Puck Music Co. Inc. -- UIC
Kalmar & Puck Music Co., 152 West 45th St. -- JHU
Kalmar & Puck Music1914 -- SIUE
New York : Kalmar & Puck Music Co., 1914 -- YU

Many SMC records include a transcription of the publisher's address, as in the fourth item in Example 6. Typically, addresses are not included in bibliographic records, but they are very valuable for research. With sufficient resources the most desirable approach would be to normalize publisher information in a way that connects both personal and corporate names of publishers in a flexible, linked, and open manner. LOD manifestations of these names, along with dates and addresses would facilitate new and flexible ways of interacting with SMC data, including the ability to create timelines and map-based browsing.

Subjects

Subject headings, when present, can relate to the subject of the lyrics, subjects in the cover art, aspects of form and genre, or some combination of these. Generally, music cataloging practice has been to use a heading or two to describe the genre only, and not to include any headings relating to the lyrics or the artwork. However, sheet music collections are often digitized precisely for their artwork, and so digitized collections include headings relating to the cover art.

Although application of Library of Congress Subject Headings genre terms is straightforward and consistently applied, there is considerable variation in the applications of headings for artwork, and a variety of vocabularies in use, including those locally constructed.

Subjects appear to be a less fruitful field on which to attempt normalization, given the resources we have available at this time. Headings are applied in such a diverse number of ways, using a variety of vocabularies, with varying degrees of uniformity, that attempting any normalization of practice, granularity or vocabulary is not realistically feasible. In addition, the majority of the subject information deployed in the SMC is from standard sources, so re-publishing just the authority records is not inherently valuable as the source of the vocabulary is the best entity to publish that vocabulary as LOD. We hope that over time common practices will emerge, promoted by access through the Consortium and its descriptive guidelines. Certainly, normalized subjects could prove quite useful, particularly if it was possible to specify which subjects relate to which aspects of the work (e.g. lyrical content, art work, etc.).

Example 7 MODS <subject> segment for “California” (Indiana University)

```
<mods:subject authority="local">
        <mods:topic>States--Western</mods:topic>
</mods:subject>
```

Whereas the IU MODS record (Example 7) contains one single broad subject heading, the range of subject headings from other repositories (Example 8) is so great as to make browsing by subject in an aggregated environment ineffective. At best these subject headings provide terms for keyword searching.

Example 8 Subjects from other SMC data providers

Love -- JHU, YU

Separation -- YU
Songs with piano -- DU1
Piano; Voice -- MSU
Songs with piano ; California -- Songs and music -- DU2
Society and Culture--State songs--California -- DU1

Dates

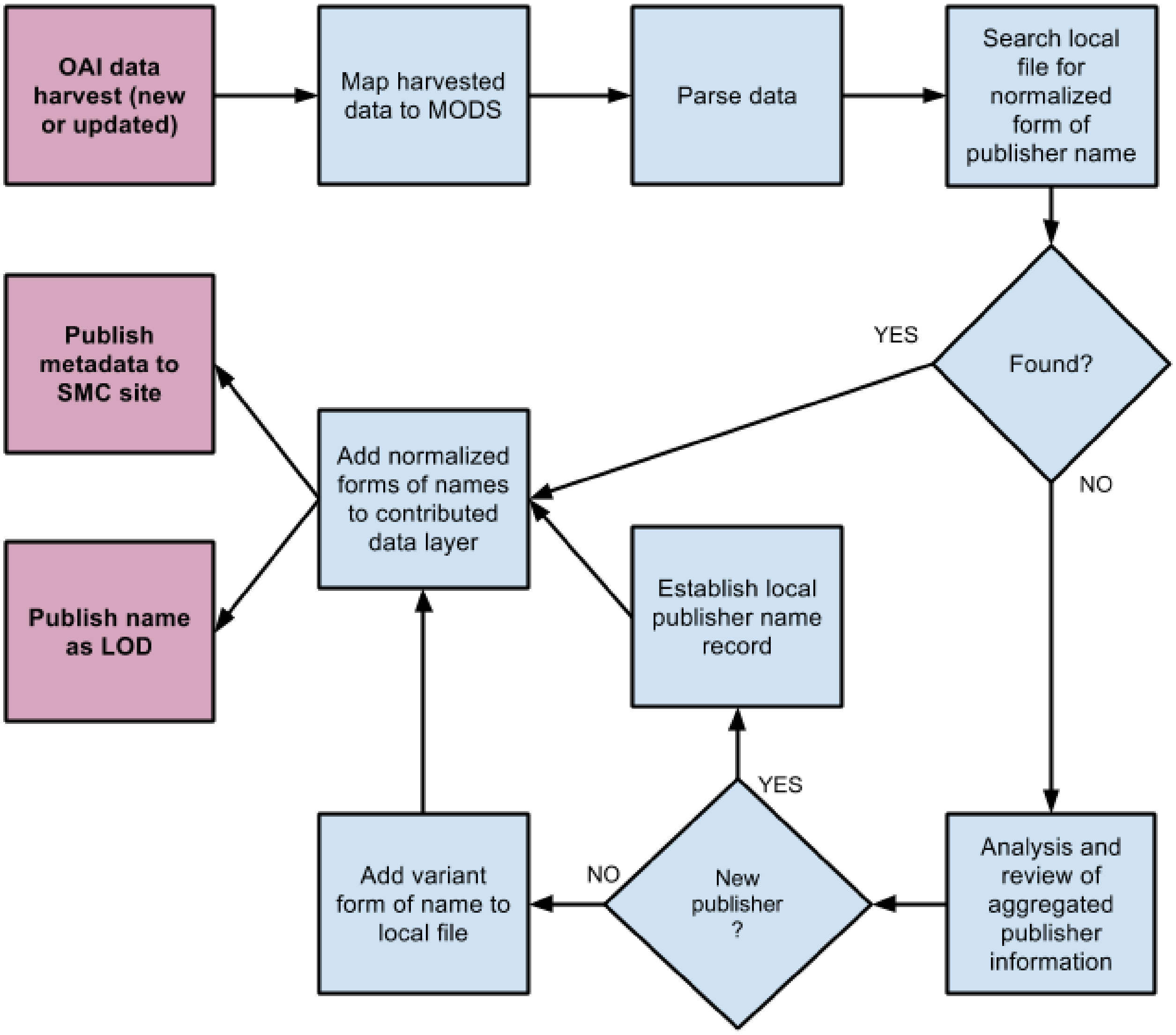
Among the problems encountered in harvested metadata is the confusion between the date of publication of the digitized object and its date of digitization, both of which can appear in the record. Dates also appear in forms that are not machine actionable--data providers generally format dates as strings according to descriptive cataloging practice--but the SMC has mitigated that problem through use of the California Digital Library's Date Normalization Utility (California Digital Library, 2005). This has allowed us to generate actionable dates and to provide end-user date browsing.

PUBLISHING AGGREGATED DATA AS LINKED DATA: A PILOT PROJECT

To test our working method (diagrammed in Figure 1) and to experiment with publishing LOD, we chose a pilot project focusing on sheet music publishers. As noted above, the roles of performers, composers, lyricists and publishers are more fluid than is typical in more formal publications, so adding a Linked Data layer that exposes some of these relationships has the potential to create new data for users and researchers. The inclusion of address information in many bibliographic records for digitized sheet music will also allow us to expose standardized information that is not normally available to end users in an actionable way.

The workflow we used to establish publisher names is very similar to the name authority workflow described above, with analysis of the aggregated publication data replacing reference to external name authority files. A few publishers are represented in authority files, but these are the exception rather than the rule.

Figure 2 Publisher name normalization workflow.



OAI data harvest (new or updated)
Map harvested data to MODS
Parse data
Search local file for normalized form of publisher name
Found?
YES
NO
Publish metadata to SMC site
Add normalized forms of names to contributed data layer
Publish name as LOD
Establish local publisher name record
YES
Add variant form of name to local file
NO
New publisher ?
Analysis and review of aggregated publisher information

Our first step (the "Parse data" box in Figure 2) was to use the text analysis tools Google Refine and Voyeur to identify and normalize names of publishers across all the aggregated SMC collections. Google Refine, currently in transition to the community-based OpenRefine (http://openrefine.org/), is a powerful data “cleaning,” or normalization tool. Voyeur Tools (http://hermeneuti.ca/voyeur) are a suite of text analysis tools developed by scholartechnologists to support work in the Digital Humanities. Although developed for different purposes, both applications provide frequencies for the occurrence of various forms of data elements such as publishers, names, titles and subjects.

The normalization features of Google Refine allow variant forms to be grouped and the source data updated if desired. An analysis of publisher information containing the word “Kalmar” for instance, reveals simple variants such as "Kalmar & Puck," "Bert Kalmar & Harry Puck," "Kalmar Puck & Abrahams," and "Kalmar & Ruby Music Corp." A little extra research in other sources identifies the publisher “Maurice Abrahams” as the same Abrahams working with Kalmar and Puck, and Ruby as a later collaborator. Using both Google Refine and corroborating research we can normalize data so, for instance, variants such as “Kalmar Puck & Abrahams," "Kalmar, Puck and Abrahams," "New York : Kalmar Puck & Abrahams, 1915," and "Kalmar Puck & Abrahams, New York" are all linked together. Normal cataloging practice would be to establish name headings using the form by which the publisher is commonly identified, but as we are not looking so much to normalize as simply to improve access through Linked Data, we have chosen to identify the most common form of name. Our effort is to aggregate variations of publisher names and present relationships, and therefore choosing a preferred name based on prevalence in the metadata records is sufficient. Besides, the promise of Linked Data is that different aggregators and systems can settle on different principal forms without a detrimental effect on user services.

Returning to the publishers of our representative title, “California and you,” Kalmar and Puck, we find that there are 72 distinct publisher names that include “kalmar” and 69 that include “puck.” These include various forms of “Kalmar & Puck,” as well as three additional publishers: Kalmar & Ruby Music Corp, based in Hollywood; Kalmar Puck & Abrahams, New York; and Maurice Abrahams Music Co., New York, with whom Kalmar and Puck collaborated.

Table 3 Summary of publisher information generated from SMC data

PUBLISHER NAME	PUBLISHER ADDRESS	DATES OF PUBLICATIONS
Kalmar & Puck		1905
Kalmar & Puck	152 West 45th Street, New York	1913-1915
Kalmar & Puck	New York	1913-1916
Bert Kalmar & Harry Puck	New York	1914-1915
Maurice Abrahams Music Co.	New York	1913-1915
Maurice Abrahams Music Co.	1570 Broadway, New York	1913-1916
Kalmar Puck & Abrahams	New York	1915-1918
Kalmar Puck & Abrahams	1570 Broadway	1917
Kalmar Puck & Abrahams	Strand Theatre Building at 47th St	1917-1918
Maurice Abrahams, Inc.	1591 Broadway, New York	1923
Maurice Abrahams, Inc.		1923-1926
Kalmar & Ruby Music Corp.	6301 Sunset Boulevard, Hollywood	1937-1939

Additional analysis provides publisher names, addresses and date ranges as represented by publications in SMC data (Table 3). From the aggregated bibliographic data alone it is possible to infer that the firms of Kalmar & Puck and Maurice Abrahams were operating independently; that they joined forces in 1915; but that Abrahams continued as an independent imprint as well. In the 1930s Kalmar teamed up with Ruby to publish music in Hollywood. This timeline is confirmed in the music historical literature (Sanjek & Sanjek, 1996).

Table 4 Archival Resource Keys (ARK) for publishers

PUBLISHER	IDENTIFIER
Kalmar & Puck	ark:/21198/r23x84k8
Maurice Abrahams Music Co.	ark:/21198/r27p8w9m
Kalmar Puck & Abrahams	ark:/21198/r2cc0xm5
Kalmar & Ruby Music Corp	ark:/21198/r2057cvv

We have established permanent identifiers for the four distinct publisher entities using the University of California Curation Center's EZID service (California Digital Library, 2013). These are given in the Table 4. Each of these IDs has an associated permanent URL that will resolve to a web page with a link to an RDF record for that publisher. The permanent URIs are of the form: http://n2t.net/IDENTIFIER, so, for example, the URI for Kalmar Puck & Abrahams becomes http://n2t.net/ark:/21198/r2cc0xm5. These four normalized publisher names are written back into the Consortium data as "user supplied" data, using the LOD system username to identify that data. In addition, RDF records are constructed for each publisher and this is published. The RDF record for Kalmar Puck and Abrahams is shown in Example 9. This data record provides links between related publishers, addresses in forms that are actionable, dates, and variant forms of publishers' names.

Example 9 Sample RDF record: "Kalmar Puck & Abrahams"

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:owl="http://www.w3.org/2002/07/owl#"
        xmlns:skos="http://www.w3.org/2004/02/skos/core#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:dc="http://purl.org/dc/terms/#"
        xmlns:time="http://www.w3.org/2006/time/#"
        xmlns:madsrdf="http://www.loc.gov/mads/rdf/v1#">

<rdf:Description rdf:about="http://n2t.net/ark:/21198/r2cc0xm5/">
```

```
        <rdf:type rdf:resource="http://www.w3.org/ns/org/Organization"/>
        <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Organization"/>
        <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
        <rdf:type rdf:resource="http://purl.org/dc/terms/Agent"/>
        <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>

<skos:prefLabel>Kalmar Puck &amp; Abrahams</skos:prefLabel>
<skos:altLabel>Kalmar, Puck &amp; Abrahams</skos:altLabel>
<skos:altLabel>Kalmar, Puck &amp; Abrahams Consolidated Inc.</skos:altLabel>
<skos:altLabel>Kalmar, Puck &amp; Abrahams Consol't'd, Inc.</skos:altLabel>
<rdfs:seeAlso rdf:resource="http://n2t.net/ark:/21198/r27p8w9m/"/>  <!--Maurice Abrahams Music Co.-->
<rdfs:seeAlso rdf:resource="http://n2t.net/ark:/21198/r23x84k8/"/>  <!--Kalmar & Puck-->
<rdfs:seeAlso rdf:resource="http://n2t.net/ark:/21198/r2057cvv/"/>  <!--Kalmar & Ruby Music Corp-->

<madsrdf:Address>
        <rdf:Description>
                <madsrdf:streetAddress>Strand Theatre Building at 47th Street</madsrdf:streetAddress>
                <madsrdf:city rdf:resource="http://sws.geonames.org/5128581/"/> <!--New York-->
                <time:year>1917</time:year>
                <time:year>1918</time:year>
         </rdf:Description>
</madsrdf:Address>

<madsrdf:Address>
        <rdf:Description>
                <madsrdf:streetAddress>1570 Broadway</madsrdf:streetAddress>
                <madsrdf:city rdf:resource="http://sws.geonames.org/5128581/"/> <!--New York-->
                <time:year>1917</time:year>
         </rdf:Description>
</madsrdf:Address>

<madsrdf:Address>
        <rdf:Description>
                <madsrdf:streetAddress>1370(?) Broadway</madsrdf:streetAddress>
                <madsrdf:city rdf:resource="http://sws.geonames.org/5128581/"/> <!--New York-->
                <time:year>1916</time:year>
         </rdf:Description>
</madsrdf:Address>

</rdf:Description>
</rdf:RDF>
```

The example of the publishers Kalmar and Puck is a relatively simple one. Returning to the firm of Oliver Ditson, discussed above, we find a much longer and more complex set of historical relationships, in which the company acquired a number of competitors in different cities, and was ultimately absorbed by Theodore Presser. A brief timeline for the Oliver Ditson and Theodore Presser firms is shown in Table 5 (Marrocco et al., n.d. [i], [ii]).

Table 5 Timeline for Oliver Ditson, Music Publisher

DATE	PUBLISHER	EVENT
1835	Oliver Ditson, Boston	firm founded by Oliver Ditson
1867	Oliver Ditson, Boston	acquired Firth, Son & Co., New York
1867	Charles H. Ditson, New York	firm founded by Oliver's son
1873	Oliver Ditson, Boston	acquired Miller & Beacham, Baltimore
1875	Oliver Ditson, Boston	acquired Wm. Hall & Son, New York acquired Lee & Walker, Philadelphia
1875	James E. Ditson, Philadelphia	firm founded by Oliver's son
1877	Oliver Ditson, Boston	acquired G. D. Russell & Co., Boston acquired J.L. Peters, New York
1879	Oliver Ditson, Boston	acquired G. André, Philadelphia
1883	Theodore Presser, Philadelphia	firm founded by Theodore Presser
1890	Oliver Ditson, Boston	acquired F.A. North & Co., Philadelphia
1931	Theodore Presser, Philadelphia	acquired Oliver Ditson

Normalization of publisher names in the Consortium data set, extraction of date ranges and addresses, and publication of these relationships as LOD would both enhance the services provided directly through the SMC website, and provide a powerful set of links between Consortium data and various other information resources, both bibliographic and otherwise.

The description of relationships between resources and persons or corporate bodies is an issue worthy of further consideration. The MARC relator terms and codes have already been published as LOD.[2] In addition, the Program for Cooperative Cataloging (PCC) RDA Task Group on Relationship Designator Guidelines has issued a report with several recommendations relevant to the Linked Data environment and the user community for sheet music. The Task Group recommends using the RDA relationship designators rather than MARC relator codes or terms to describe the relationships between resources and other entities. Under RDA, the list of available relationship designators is a controlled vocabulary, but it is a vocabulary to which catalogers may add new terms relatively easily. The Task Group recommends that PCC catalogers use these RDA relationship designators, unless they are following specialist community guidelines prescribing a different vocabulary. (The Task Group also suggested that a new PCC task force could be charged with evaluating lists of such terms used by other communities, and recommending terms or entire lists for inclusion among the RDA relationship designators.) The Task Group also notes that RDA relationship designators are available via the Open Metadata Registry, which may support the possibility of mapping between multiple existing vocabularies (Andrew et al., 2013). It is heartening to see the PCC recognize the importance of recording the specific roles played by persons and organizations in relation to resources, since such information may assist in the discovery, understanding and analysis of resources such as sheet music publications.

CONCLUSION

Although Open Archives Initiative services have the potential to provide a variety of end-user services around aggregated metadata, the reality is that these services are sometimes seriously hampered by the inconsistent application of metadata standards, lack of adherence to uniform values for names and subjects, and varying levels of granularity between data providers. All of these problems are evidenced through the experience of the SMC. Mitigating these problems at

[2] Available at http://id.loc.gov/vocabulary/relators

the point of aggregation is certainly beneficial, but without the ability to push enhanced metadata back to the data provider the benefits are limited, especially given that data re-harvests can potentially re-introduce the same, or similar, problems.

LOD standards and strategies provide OAI service providers with a new set of possibilities. By selectively normalizing aggregated metadata and adding an LOD layer derived from the aggregated data the service provider has the potential to publish data back out to the community in a form that can be used in a variety of ways, by a variety of users and systems. By working on a group of prominent sheet music publishers, we hope to demonstrate the utility of LOD within an information aggregator.

ACKNOWLEDGMENTS

The recent development of the Sheet Music Consortium was made possible in part by two National Leadership Grants from the Institute of Museum and Library Services; first a planning grant (LG-52-07-0114-07), followed by an implementation grant (LG-05-09-0035-09). Our colleagues at Indiana University have been instrumental in this revision of the SMC services, including Jenn Riley (now at the University of North Carolina at Chapel Hill), Michelle Dalmau, Jon Dunn, Dot Porter (now at the University of Pennsylvania), and Mark Notess. Finally, we are grateful to all the SMC participants who have worked diligently to digitize, describe and publish their content.

REFERENCES

1. Andrew, P. et al. (2013, March 13). *PCC Relationship Designator Guidelines Task Group report*. Retrieved from http://www.loc.gov/aba/pcc/rda/RDA%20Task%20groups%20and%20charges/PCC-Relat-Desig-TG-report.rtf

2. Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May 17). The Semantic Web. *Scientific American*, *284*(5), 34–. Retrieved from http://www.scientificamerican.com/article.cfm?id=the-semantic-web

3. Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked Data on the web (LDOW2008) Workshop at the 17th International World Wide Web Conference. In *Linked Data on the web (LDOW 2008)*. Presented at the 17th International World Wide Web Conference, Beijing, China.

4. Brogan, M. (2003). *Digital library aggregation services.* Washington, DC: Digital Library Federation. Retrieved from http://www.diglib.org/pubs/dlf101/

5. Brogan, M. (2006). *Contexts and contributions: Building the distributed library.* Washington, DC: Digital Library Federation. Retrieved from http://www.diglib.org/pubs/dlf106/

6. California Digital Library (2005). Date normalization utility documentation. Retrieved from http://www.cdlib.org/services/dsc/projects/docs/datenorm_documentation.pdf

7. California Digital Library (2013). EZID. http://www.cdlib.org/services/uc3/ezid/

8. Coyle, K. (2009). Metadata mix and match. *Information Standards Quarterly*, 21(1), 9–11. Retrieved from http://kcoyle.net/isqv21no1.pdf

9. Dunsire, G., Hillmann, D., & Phipps, J. (2012). Reconsidering universal bibliographic control in light of the Semantic Web. *Journal of Library Metadata*, *12*(2-3), 164–176. doi:http://dx.doi.org/10.1080/19386389.2012.699831

10. Dunsire, G., Harper, C., Hillmann, D., & Phipps, J. (2012). Linked Data vocabulary management: Infrastructure support, data integration, and interoperability. *Information Standards Quarterly*, *24*(2/3), 4–13.

11. Hagedorn, K. (2003). OAIster: A “no dead ends” OAI service provider. *Library Hi Tech*, *21*(2), 170–181.

12. Harper, C. (2012). Letter from the guest content editor. *Information Standards Quarterly*, *24*(2/3), 2–3.

13. Harper, C., & Tillett, B. (2007). Library of Congress controlled vocabularies and their application to the Semantic Web. *Cataloging & Classification Quarterly*, *43(3-4)*, 47-68.

14. International Federation of Library Associations and Institutions (1998). Functional Requirements for Bibliographic Records, final report. Retrieved from http://archive.ifla.org/VII/s13/frbr/frbr1.htm#3.2

15. Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., & Saylor, J. (2006). Metadata aggregation and “automated digital libraries”: a retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 230–239). New York, NY, USA: ACM. doi:10.1145/1141753.1141804

16. Marrocco, W. T., Jacobs, M., & Krummel, D. W. (n.d. [i]). Ditson, Oliver. *Grove Music Online*. Retrieved from http://www.oxfordmusiconline.com/subscriber/article/grove/music/07860

17. Marrocco, W. T., Jacobs, M., & Krummel, D. W. (n.d. [ii]). Presser. *Grove Music Online*. Retrieved from http://www.oxfordmusiconline.com/subscriber/article/grove/music/22310

18. Sanjek, R., & Sanjek, D. (1996). *Pennies from Heaven: the American popular music business in the Twentieth Century.* New York: Da Capo Press.

19. Shreeves, S. L., Habing, T. O., Hagedorn, K., & Young, J. A. (2005). Current developments and future trends for the OAI protocol for metadata harvesting. *Library Trends*, *53*(4), 576–589.

20. Tennant, R. (2004, May 14). Bitter harvest: Problems & suggested solutions for OAI-PMH data & service providers. Retrieved from http://roytennant.com/bitter_harvest.html

21. Yasser, C. M. (2011). An analysis of problems in metadata records. *Journal of Library Metadata*, *11*(2), 51–62.