

How libraries use publisher metadata

With the proliferation of electronic publishing, libraries are increasingly relying on publisher-supplied metadata to meet user needs for discovery in library systems. However, many publisher/content provider staff creating metadata are unaware of the end-user environment and how libraries use their metadata. This article provides an overview of the three primary discovery systems that are used by academic libraries, with examples illustrating how publisher-supplied metadata directly feeds into these systems and is used to support end-user discovery and access. Commonly seen metadata problems are discussed, with recommendations suggested. Based on a series of presentations given in Autumn 2012 to the staff of a large publisher, this article uses the University of Washington Libraries systems and services as illustrative examples. Judging by the feedback received from these presentations, publishers (specifically staff not familiar with the big picture of metadata standards work) would benefit from a better understanding of the systems and services libraries provide using the data that is created and managed by publishers.

Introduction

The purpose of this article is to provide an overview of how libraries provide access to publisher content using publisher-provided metadata. This data is most commonly seen in three systems which are managed or profiled by libraries:

- MARC-based library catalogs
- OpenURL link resolvers
- Library discovery systems.

These three systems are not mutually exclusive since link resolution and MARC catalog records often support library discovery systems. Library service has never exclusively been about the library catalog but it is even less so in the current service environment.

University of Washington Libraries

The University of Washington (UW) includes a large, heterogeneous research operation that is the largest public university recipient of US federal research funding (totalling US\$1 billion for fiscal year [FY] 2012). The UW Libraries consists of 16 libraries on three campuses with 5.2 million visits last year. Our digital collection includes 500,000 licensed electronic books, 100,000 online journals and 600,000 locally digitized items in 300 collections. Our physical collection includes 8 million print volumes, 6 million microforms and 60,000 print serials. In FY2011, we had 6 million licensed journal articles downloaded, 1.8 million check-outs, answered 50,000 reference questions in person and another 15,000 reference questions online. *It is not possible to provide this amount of service (both in person and remotely) without systems in place that use metadata and services supplied by publishers and content providers.*

In 2009, the Libraries' Information Technology Services (ITS) office at the UW Libraries developed a series of library user 'personas' to help guide system development.¹ Five personas were developed:

- Brooke the Beginner (~31,000 undergraduates)



STEVE SHADLE
Serials Access
Librarian
University of
Washington Libraries

- Richard the Researcher (~11,000 graduate students)
- Sharon the Scholar (~4,500 faculty and academic staff)
- Paul the Professional (~1,800 students in professional programs)
- April the Alumna (~500 alumni users).

Because two of these personas (Brooke and Richard) account for about 85% of our users, this paper will focus on the user experience from their perspectives.

Brooke is a 19-year-old undergraduate who has not yet declared a major. Currently, she is taking classes in English, history and biology. She is new to the research process and academia and will be working on several research assignments in humanities and social science disciplines but will not be a content expert. She generally uses the Libraries' website for course support (reserves, library open hours/study spaces) and tends to start research in Google or with the Libraries' default search. Quote: "I'd rather use an online article that works than go to the hassle of find a book in the library."

Richard is a 29-year-old doctoral student in the College of Built Environments. His dissertation research is on modeling public transportation utilization and incentives. Richard is an experienced researcher who generally uses the library website to obtain research materials and to use licensed databases (e.g. Web of Science). Quote: "Accessing full-text articles online is my primary use of the library and is central to my research ... but I still go to the library for some reference materials that aren't online."

Scenario 1: The MARC-based Library catalog

Historically, the library catalog was the record of the library's physical holdings. Beginning in the mid 1990s, libraries started including online licensed resources in the library catalog with links to access the online content (so the catalog began to take on some aspect of being a very selective web portal). Most library catalogs still use the MARC (MACHINE Readable Cataloging) record format.

A typical use of the UW Library catalog by Richard is to find a conference proceeding. Richard's advisor has mentioned a recent conference on transport management systems that could be of interest to Richard. However, the full text of the conference proceedings is not available from the conference website. As an experienced researcher, Richard knows that the Library frequently has conference proceedings so he searches the Library catalog to see if the Library can provide him with any recent proceedings.

The screenshot shows a search interface with a search bar containing 'Transport Systems Telematics' and 'Entire Collection'. Below the search bar, it indicates '9 results found. Sorted by relevance | date | title'. The search results list includes:

- Author:** International Conference "Transport Systems Telematics TST" (11th : 2011 : Katowice-Ustroń, Poland)
- Title:** Modern transport telematics [electronic resource] : 11th International Conference on Transport Systems Telematics, TST 2011, Katowice-Ustroń, Poland, October 19-22, 2011, selected papers / Jerzy Mikulski (ed.).
- Pub Info:** Berlin ; Heidelberg : Springer, c2011.

Below the title, there is a link: [SpringerLink Connect to this title online; UW restricted](#). A red arrow points to this link. Below the link, it says 'Part of the package: Springer Computer Science'. The 'Details' section includes:

- Description:** 1 online resource (xiv, 418 p.) : ill.
- Series:** Communications in computer and information science, 1865-0929 ; 239
- Series:** Communications in computer and information science ; 239.
- Bibliography:** Includes bibliographical references and author index.
- LC SUBJECTS:** Telematics -- Congresses, Transportation -- Data processing -- Congresses.
- Subj Keywords:** Computer science, Computer Communication Networks, Computer software, Database management, Information storage and retrieval systems, Information Systems, Information Systems Applications (incl. Internet), Information Storage and Retrieval, Algorithm Analysis and Problem Complexity, Management of Computing and Information Systems.
- Genre/Form:** Electronic books.
- Other Author:** Mikulski, Jerzy, 1946.

Figure 1. Example showing results of a keyword search in the Library catalog

292 The results of a keyword search of the conference name in the Library catalog (see Figure 1) include the online proceedings for the 2010 and 2011 conferences, and links in the record take Richard to the full text of the proceedings (which the Library has licensed as part of a publisher package). Perfect!

This record was not created by a Library cataloger, but was instead created by the publisher and loaded into the Library catalog by one of our catalogers. 'Cataloging' through the use of record-set loads has been around for a long time for microform sets, but with an increasing amount of electronic publishing, library cataloging is moving away from title-by-title handling towards the management of sets of publisher-supplied records. When librarians ask publishers and vendors for 'record sets', they are typically asking for a set of MARC records following a content standard such as AACR2 or RDA.

"... library cataloging is moving away from title-by-title handling towards the management of sets of publisher-supplied records."

Figure 2 is the underlying MARC record for what is displayed in Figure 1. Because the MARC format and library content standards are so specialized, publishers who provide MARC records generally have specialized staff or contractors who create and manage these records. MARC record creation and management is typically separate from the day-to-day processes of editorial, marketing and production staff. However, with MARC editing tools (such as MARCedit), catalogers are now able to convert non-MARC metadata into the MARC format and can load this metadata into the library catalog.

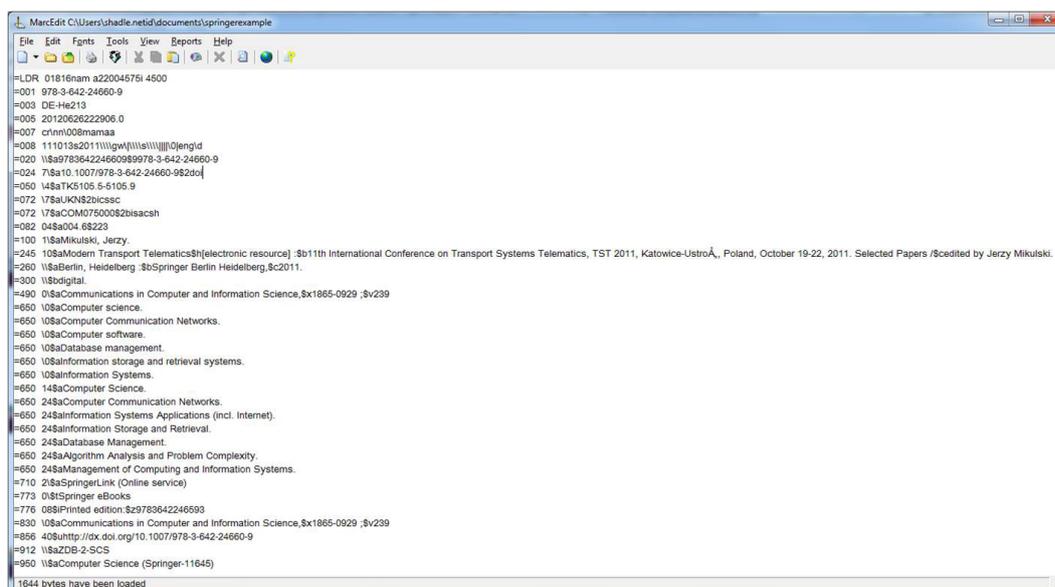


Figure 2. The underlying MARC record for the keyword search performed in Figure 1

Scenario 2: The OpenURL link resolver

An OpenURL link resolver is essentially a service that takes a citation formatted as an OpenURL and provides the user with library services related to that citation. These services can include accessing the online full text, placing an inter-library loan (ILL) request, searching a library catalog, or finding related resources. The citation is often referred to as the 'source' and the services are often referred to as 'targets'.

An OpenURL knowledge base (KB) is a database profiled by a library which contains information about electronic resources (e.g. e-journals, e-books) that are licensed by the library. The KB contains resource metadata including elements such as title, author, identifiers (ISBN/ISSN) journal coverage, resource provider and URL. Using the KB, an OpenURL link resolver can determine if an item (article, book, etc.) is available electronically and can identify the appropriate copy to serve to the user.

"... catalogers are now able to convert non-MARC metadata into the MARC format ..."

Librarians implement link resolvers for a number of reasons. Library catalogs are a time-consuming way for users to access article full text as the process entails searching in the catalog, identifying the correct journal, linking to the journal website from the catalog record and then drilling through several layers at the journal website (title, volume, issue, article) to get to the full text. Navigating the library catalog and journal website can take as many as eight clicks vs. one or two clicks in getting from source to target using a link resolver. Also, the link resolver gets the user to the correct copy in cases where the content is available from more than one source (e.g. EBSCOhost and Springer). And in cases where the library does not have the full-text content licensed, the link resolver can pass citation information to a document delivery request or a catalog search, helping the user get to additional library services.

“Navigating the library catalog and journal website can take as many as eight clicks vs. one or two clicks ...”

The link resolution process consists of essentially three steps:

- 1) The link resolver parses the citation elements from the source OpenURL,
- 2) The resolver tests those elements against a library's KB and identifies targets based on those test results, and finally
- 3) The resolver creates and offers links based on the linking logic of the target service using the citation elements from the OpenURL.

As mentioned previously, an OpenURL passes citation information to a library's link resolver. Here is an example an OpenURL sent from Web of Science to the UW link resolver:

http://resolver.lib.washington.edu/?&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&rft.atitle=Improving%20Group%20Attention%3A%20An%20Experiment%20with%20Synchronous%20Brainstorming&rft.aufirst=Antonio&rft.aulast=Ferreira&rft.date=2011&rft.spage=643&rft.epage=666&rft.genre=article&rft.issn=0926-2644&rft.issue=5&rft.jtitle=GROUP%20DECISION%20AND%20NEGOTIATION&rft.pages=643-666&rft.stitle=GROUP%20DECIS%20NEGOT&rft.volume=20&rfr_id=info:sid/www.isinet.com:WoK:UA&rft.au=Antunes%2C%20Pedro&rft.au=Herskovic%2C%20Valeria&rft_id=info:doi/10.1007%2Fs10726-011-9233-y

Note that the bolded elements include standard citation elements such as article title, article author, journal title and ISSN, issue date and numbering and article start and end pages. In addition, the OpenURL indicates that the source for this OpenURL is Web of Science (info:sid/www.isinet.com:WoK:UA) and that it is directed at the UW link resolver (resolver.lib.washington.edu).

As a final note, sources from providers must support OpenURL services (meaning a library can profile a source database to include a link or button that generates an OpenURL call to their link resolver) and targets must have a consistent linking convention so that a link resolver can create a link that directs a session to a specific article.

Getting back to Richard, let's assume he sees this article in Web of Science (Figure 3) and clicks on the purple 'Check for Full Text' button. Clicking that button essentially performs the same action as entering that long OpenURL we just looked at. The UW link resolver then:

- 1) parses the OpenURL for ISSN (0926-2644) and date (2011) elements,
- 2) checks those elements against the KB to determine that the UW does have the article licensed from the publisher (so SpringerLink is the target), and
- 3) uses the publisher linking logic to create the following article-level link: <http://www.springerlink.com/OpenURL.asp?genre=article&id=doi:10.1007/s10726-011-9233-y>.

Depending on how the link resolver is set up, the user's session might automatically redirect to this URL or it may be redirected to an intermediate display that offers the link to the user. In either case, Richard gets from Web of Science to the full text of the article in one or two clicks. *In order for link resolvers to function properly, citation metadata coming from a source must be accurate. In most cases, this data is originally coming from the publisher.*

Figure 3. Example showing a Web of Science search result

Certain data included herein are derived from the Web of Science® prepared by Thomson Reuters Scientific, Inc (Thomson®), Philadelphia, Pennsylvania, USA: © Copyright THOMSON REUTERS® 2013. All rights reserved

Scenario 3: Library discovery services

A library discovery service is essentially a search interface to pre-indexed metadata and/or full-text documents made available by a library. From the user perspective, it is similar to the Google search experience in appearance. A good discovery service provides a simple search that is comprehensive enough that it can serve as a good starting point for research. Discovery services differ from federated search applications in that discovery services don't search live sources. By searching pre-indexed data, discovery services return search results more quickly than federated search systems. Discovery services can include local collections in addition to licensed resources. And because most discovery services use OpenURL resolution, they are able to provide access to online content that the traditional library catalog may not.

Each discovery service has a different mix of content and can often be customized to include local content. However, most library discovery services consist primarily of content that has been historically available from libraries (e.g. books, journals, articles). One of the key points about discovery services is that they are comprehensive enough to serve as a 'one-stop shop' for basic research. As an example, OCLC WorldCat Local includes metadata for 681 million articles, 30 million digital items (from sources such as Google Books, Hathi Trust and OALster), 13 million e-books and 225 million print books.

"... discovery services return search results more quickly than federated search systems."

So let's see how our undergraduate (Brooke) uses the library discovery service. In Brooke's ENGL 210 class (English Medieval and Early Modern Literature), she learns about the Anglo-Saxon literary practice of *opus geminatum* (twinned work; a work consisting of a pair of texts, one in prose and one in verse). Her professor also mentions that paraphrase was often used as a literary device in this time period. Brooke is required to write a research paper on a topic of her choice which includes three peer-reviewed articles as background, and she decides she wants to research this practice.

295 Unfortunately, Brooke doesn't take note of the phrase '*opus geminatum*' so she starts at the library home page searching on terms such as 'twinned work', 'anglo-saxon', and 'paraphrase'. At one point, she searches using the terms 'paraphrase anglo-saxon literature' and gets the results shown in Figure 4.

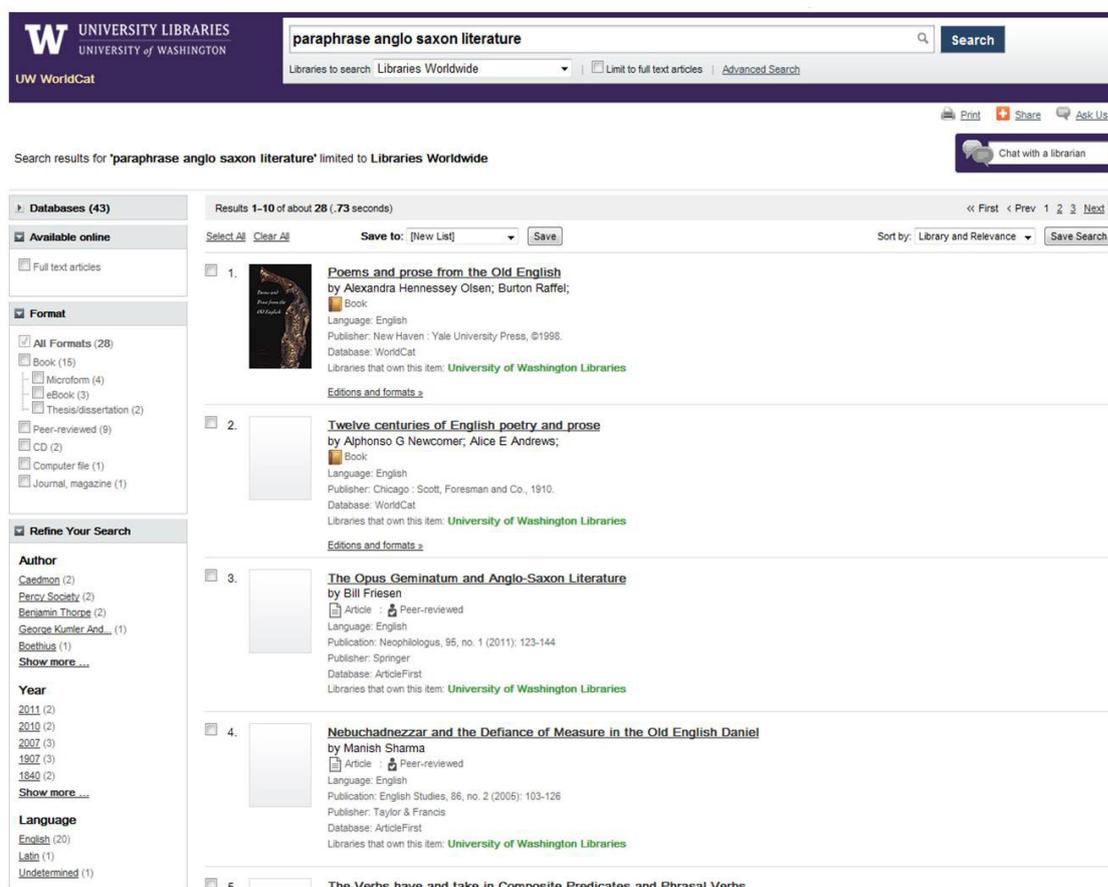


Figure 4. Example showing the result of a typical library discovery search

Typical of library discovery search results, she gets a mix of books and articles (one-stop shop). Brooke sees the third entry ('The Opus Geminatum and Anglo-Saxon Literature') and remembers that is the phrase her professor used. She clicks on the entry to see what is there and she gets a more detailed description of the article with a link offering her the full text (using OpenURL link resolution). She downloads the article for later reading.

Later, Brooke is in Google, remembers the phrase '*opus geminatum*' and searches in Google using that phrase. The results (Figure 5) include a Wikipedia entry for a specific instance of an *opus geminatum* ('Candidus of Fulda') which provides additional background information on the literature style. Other resources are listed (mostly articles and books) that might be useful for additional research. But note the second entry is for the Springer-published article that she found earlier. When she clicks on the link for the article, the session redirects to the article full text just as it does when an OpenURL call is taking her from a citation database or from the Library discovery service. This happens because the Library has profiled its IP ranges with Google, so that Google can pass the referring IP to the link target. As long as Brooke is on a campus workstation or has proxied her session, Google will recognize her as a University of Washington user. Since Google has the article metadata necessary to create an article-level URL (most likely an article DOI), Google can redirect the session to the article and because Brooke is already proxied, Springer serves the full text. This operation is completely transparent to Brooke, who probably thinks it is available 'for free' on Google. As with discovery services, a necessary element in this chain of events is accurate metadata from the publisher.

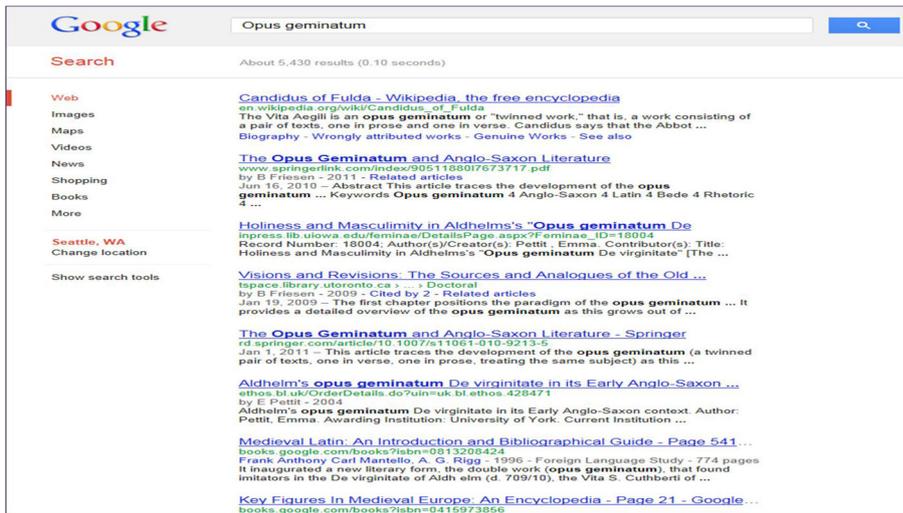


Figure 5. Example showing the result of a Google search

Google and the Google logo are registered trademarks of Google Inc, used with permission

In discussing these examples with staff at one major publisher, one of the 'Aha!' moments I witnessed was showing the entry for the same article in several different systems (e.g. WorldCat, Summon, Web of Science, Google, EBSCOhost) and pointing out that the data they create in their local systems is distributed across at least a dozen services affecting hundreds of thousands (if not millions) of potential users.

"... one of the 'Aha!' moments ..."

How it goes wrong and what you can do about it

The examples discussed so far are success stories. In working with library systems and troubleshooting bad links, I've seen classes of metadata-related errors that are in control of the publisher. In terms of article-level link resolution, many failures have to do with the use of incorrect ISSN (including use of ISSN for earlier or later titles). The incompleteness or inaccuracy of other metadata elements can also cause problems. An example that comes to mind is when publishers provide metadata feeds that include different types of resources (maybe a small number of theses/dissertations or e-texts along with articles) but the genre for all of their content is specified as an article. Genre is an important element in OpenURL resolution as the linking logic may look for different elements (ISSN vs. ISBN) depending on the genre indicated in the OpenURL. However, if it's an element that is set in a template, the content provider may be unknowingly sending inaccurate information (by calling everything an article, for example).

The National Information Standards Organization (NISO)'s Project Iota² has done quite a bit of research into the completeness of OpenURLs and even though the report is aimed at link resolver providers, publishers and content providers who want to know more about OpenURL elements may find this report useful. Publishers who need guidance as to what and how to send data to KB providers (and similarly to discovery services) should review the work of KBART.³ And if publishers want to provide better metadata, one tactic to take is to educate line staff on the importance to the end user of the work they do on a daily basis.

"... many failures have to do with the use of incorrect ISSN ..."

Summary

There is a perception that libraries only make use of MARC records. It is true that, historically, MARC records have supported an important access tool (the library catalog), but these days it is about a lot more than just MARC records (and it is about a lot more than just the library catalog). Any source that supports OpenURL (including Google) can potentially provide access to library-licensed publisher content. The metadata that supports these access methods is supplied by the publisher. In addition, metadata accuracy is about more than just correct transcription. It is about having a thorough understanding of the

297 standards being used to transmit metadata and making sure that metadata which is sent out to the world follows standards *and* accurately reflects the resource being described. Library cataloging departments have never been able to do it all and in this era of declining staff resources, we need to rely on the work of publishers and content providers even more than we have in the past.

Acknowledgment: Thanks to Mirjam Kessler who co-presented and also made the trip possible through her time, energy and patience.

“... we need to rely on the work of publishers and content providers even more than we have in the past.”

References

1. Ward, J, *Persona Development and Use, or, How to Make Imaginary People Work for You*: http://libraryassessment.org/bm-doc/ward_jennifer.pdf (accessed 23 Aug 2013).
2. IOTA: Improving OpenURLs Through Analytics: <http://www.niso.org/workrooms/OpenURLquality> (accessed 30 August 2013).
3. KBART: Knowledge Bases and Related Tools working group: <http://www.uksg.org/KBART> (accessed 30 August 2013).

Article © Steve Shadle

Steve Shadle, Serials Access Librarian
University of Washington Libraries, Seattle, WA 98195-2900, USA
Tel: (206) 685-3983; Fax: (206) 543-0854; E-mail: shadle@u.washington.edu

To cite this article:

Shadle, S, How libraries use publisher metadata, *Insights*, 2013, 26(3), 290–297; DOI: <http://dx.doi.org/10.1629/2048-7754.109>