Metadata Challenges in Library Discovery Systems

Pascal Calarco
University of Waterloo

Lettie Conrad
SAGE, lettie.conrad@sagepub.com

Rachel Kessler
Ex Libris, rachel.kessler@exlibrisgroup.com

Michael Vandenburg
Queens University

Pascal Calarco, Lettie Conrad, Rachel Kessler, and Michael Vandenburg, "Metadata Challenges in Library Discovery Systems" (2014). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315642

Metadata Challenges in Library Discovery Systems

Pascal Calarco, Associate University Librarian, Research and Digital Discovery Services, University of Waterloo
Lettie Conrad, Executive Manager, Online Products, SAGE
Rachel Kessler, Product Manager, Primo Central, Ex Libris
Michael Vandenburg, Associate University Librarian, Queens University

Abstract

With discovery systems such as Summon, EDS, and Primo Central, patrons can search nearly all of their libraries' resources from a single platform. In order to create this experience, data from disparate sources must be normalized and unified into one index.

In this session, we discussed some of the metadata challenges facing each of the parties involved in library discovery; the library, the publisher, and the discovery system provider. Libraries must normalize their bibliographic records to make them compatible with the discovery system's schema. Publishers need to create mechanisms to regularly export records with meaningful metadata, and the discovery system provider must integrate metadata from these sources while ensuring the best possible user experience.

We also touched on the recent guidelines of the NISO Open Discovery Initiative. The guidelines include goals such as "to streamline the process by which information providers, discovery service providers, and librarians work together to better serve libraries and their users." The session will explore how these guidelines can be implemented along with some of the challenges and will include a discussion with the audience.

Introduction

With discovery systems such as Summon, EDS, and Primo Central, patrons can search nearly all of their libraries' resources from a single platform. In order to create this experience, data from disparate sources must be normalized and unified into one index.

In this session, we discussed some of the metadata challenges facing each of the parties involved in library discovery; the library, the publisher, and the discovery system provider. Libraries must normalize their bibliographic records to make them compatible with the discovery system's schema. Publishers need to create mechanisms to regularly export records with meaningful metadata, and the discovery system provider must integrate metadata from these sources while ensuring the best possible user experience.

We also touched on the recent guidelines of the NISO Open Discovery Initiative. The guidelines include goals such as "to streamline the process by which information providers, discovery service providers, and librarians work together to better serve libraries and their users." The session will explore how these guidelines can be implemented along with some of the challenges and will include a discussion with the audience.

For Publishers

SAGE sees our participation in library discovery services as critical to our success as a member of the scholarly communication supply chain. New resources at SAGE enable closer partnership with other members of this supply chain, to openly share our vision and product strategy, to listen to peers in other organizations, to understand their needs and priorities, and to collaborate toward solutions to share challenges—such as those we're all experiencing with the metadata required for discovery of scholarly content.

To optimize the visibility and performance of SAGE content in discovery systems, like Primo, we dedicate a good deal of resource to operationalize distribution of high-quality metadata. These come in the form of:

http://dx.doi.org/10.5703/1288284315642

- Content architecture—to ensure the full text of our content is well structured and fully marked up.
- Industry standards—it's important to SAGE that we're in compliance with the proper data standards.
- Systems development—for storage and delivery.
- Dedicated staff—SAGE has put a library discovery work group in place, made up of product and technical analysts along with reps from across the business.

This new group at SAGE recently conducted a SWOT of our metadata in order to crystalize our understanding of the challenges and changes we're facing. Here are the highlights:

- New content types—for SAGE, when we decide to add new types of content to our publishing programs, we struggle to establish metadata expertise, define process and develop our human and systems workflows.
- Manual versus automated—like many companies, these new workflows often begin by significant hand-wrought metadata, so we must eventually invest in systems development to automate these new process.
- Accuracy—keeping an eye on data accuracy along the way and developing appropriate QA routines that don't slow down publication.
- Industry standards—NISO and others are doing yeoman's work to establish clear protocols for metadata, but there isn't a published standard for every single metadata entity that we publish, so we sometimes find ourselves tripping a bit in the standards gaps.

For the most part, journals metadata is the most mature and the most automated, and has the benefit of the greatest number of industry standards. However, there are still moving targets, as the industry continues to evolve. Chief among these are around open access. Hybrid OA is a challenge for all members of the supply chain. So far, we've not hit on a great way to identify open articles in traditional subscription journals—largely because we've not arrived at a standard metadata element to indicate OA status and license types in our FT XML. Our prevailing journal markup protocol (now JATS) does not yet have an agreed upon identifier for OA articles or terms.

For SAGE, we're seeing a range of challenges around article-level workflows. Most of our systems—and our indexing partners' systems—are organized around the traditional issue/volume model. So, ahead-of-print articles, hybrid OA articles, and any other new model publishing at an article level create challenges for assembly, storage, and distribution of journal metadata.

Metadata for e-books and e-reference have some unique challenges. First, we don't have the benefit of journals' consistency in content structures—instead, in this category, we're dealing with encyclopedias, dictionaries, handbooks, monographs, case studies, and others, none of which share the similar formats. Some have abstracts, some don't. Some have references, some don't. These diverse types of content demand a good deal of manual work and limit our ability to automate some metadata creation and delivery steps.

Since you can't really set your watch to e-books metadata, the indexing routines and practices of our discovery partners is also diverse and variable. This further undermines our ability to standardize and automate our own processes. Some indexers—like Google Scholar—just won't touch scholarly ebooks with a 10-foot pole due to the chaotic landscape of e-books metadata.

Not surprisingly, e-books and e-reference data standards are also variable. There is no JATS for books—would that be BATS?—so everyone along the supply chain is struggling and you might say "storming" toward more consistency routines and data protocols.

Finally, I want to touch on both data and video metadata—what we sometimes joke is still the "wild west" of metadata. I don't mean to oversimplify either type of content, but these

each have very similar patterns at the moment. If ebooks/e-reference metadata are diverse, then metadata for data and video content is pure anarchy. When we ask indexers and standards bodies for guidance on marking up these formats, we get a lot of head scratching—understandably so. Some things are becoming clear over time—such as the fact that both data and video assets require rich text-based content to make search and discovery remotely possible. But, we're still dealing with a void of standards for video transcripts or narratives to accompany datasets. And the published guidelines for standard identifiers—such as DOIs or ISBNs—are still lacking clarity in handling these new forms of scholarly output.

So what is a poor publisher to do with all these steep hills to climb? At SAGE, we believe it's vitally important for all vested parties to actively participate in the standards formation process. We are NISO voting members and active in committees such as the Open Discovery Initiative. We conduct original research into discovery and metadata practices, we publish whitepapers on these topics, in an effort to share our knowledge and continually learn from our industry peers.

We're also ramping up our internal metadata practices with more formal routines for generating and enhancing our content metadata. That Library Discovery WG mentioned earlier is now meeting more regularly to develop rubrics for assessing our data quality and compliance. We're also developing some internal standards and guidelines. For example, after a great deal of research conducted this year, we're putting together guidelines for all SAGE teams to use in our application and use of DOIs and other identifiers.

In general, though, we're practicing the art of flexibility, keeping our ears to the ground and working toward more agility reacting to developing standards and new protocols as metadata continues to evolve.

For Discovery Systems

The key challenge is to pull content from disparate sources and normalize it into a uniform database that contains records that are "useful" to library patrons. While "useful" is a vague term, the objectives of normalization can be broken down into three main goals:

1. *To make the content discoverable*—users must be able to surface the records. If the associated metadata is incorrect, for example, the record could be difficult to find.
2. *To facilitate delivery*—while many consider the primary function of a discovery system to be the enabling of finding relevant materials, the ability to actually reach the full text of that content is equally important. After all, most researchers will not find a citation useful if they cannot actually access the material. If the metadata is insufficient to allow users to access material via a link resolver or other means, the record is not considered "useful."
3. *To maintain a visually appealing interface*—content should be scannable to facilitate easy skimming of result sets. Additionally, metadata should be uniform in order to maintain a uniform, professional appearance.

The achievement of these objectives is often dependent on the quality of the metadata supplied by data providers. Below are three examples of poor quality metadata supplied to Ex Libris Primo by various providers.

Example 1

> *"Microsoft's antitrust fine—Sin of omission"*
>
> **Information in publication title field:**
>
> **[t][The economsst]**
>
> This obvious typo can affect all three of the above objectives and despite the fact that it is clearly a mistake can be difficult to catch among one billion records.

Example 2

> **"Valuation of mangrove services of Andaman and Nicobar Islands, India"**

Information in the date field:

date=1013

The full text of this article lists the date as 2013. These types of errors are relatively easy to catch as rules can be created to flag records with dates from before a given year, i.e., 1450, the year the printing press was invented.

Example 3

"Book review corner"

Information in the resource type field:

<cto:doctype>cp</cto:doctype>

("cp" stands for conference proceeding!)

This document is clearly a book review and not a conference proceeding. Checks can be performed to catch some errors, i.e., does the title of the article contain the words "book review?" However, many mislabeled resource types go unnoticed if there are no obvious cues in the metadata.

Another area where discovery providers find it challenging to normalize data is authors. In addition to authority challenges, a standard has not yet been set for the format in which the data should be delivered. Simple issues such as punctuation and spacing can be easily fixed through normalization. However, how the names themselves are reported in the xml structure, can be more problematic. In the following example, you can see three different ways of providing the same information.

Provider 1

<author>

Lei Yang, Yajuan She, Shihua Zhao, Shihai Yue,

Qian Wang, Aiping Hu, Wei Zhang

</author>

Provider 2

700 1 {{[a][Hu, Aiping]}}

700 1 {{[a][She, Yajuan]}}

700 1 {{[a][Wang, Qian]}}

700 1 {{[a][Yang, Lei]}}

700 1 {{[a][Yue, Shihai]}}

700 1 {{[a][Zhang, Wei]}}

700 1 {{[a][Zhao, Shihua]}}

Provider 3

<preferred-name>

<ce:initials>L.</ce:initials>

<ce:indexed-name>Yang L.</ce:indexed-name>

<ce:surname>Yang</ce:surname>

<ce:given-name>Lei</ce:given-name>

</preferred-name>

<preferred-name>

<ce:initials>Y.</ce:initials>

<ce:indexed-name>She Y.</ce:indexed-name>

<ce:surname>She</ce:surname>

<ce:given-name>Yajuan</ce:given-name>

</preferred-name>

The more the fields are broken down, the easier it is for discovery providers to understand what they are receiving. Provider 3's data, for example, contains no ambiguity regarding which is the first name and which is the last or where one name begins and the other ends. This makes it easier for discovery providers to normalize the data and reduces the chance of parsing errors. An author name, which is normalized incorrectly, can lead to issues with all three objectives listed above. Ex

Libris is beginning to index ORCIDs within Primo Central records, which should disambiguate author names as ORCID's popularity rises.

It would also be exceedingly helpful to discovery services if a standard cataloging unit was created for discovery. The best way to explain this point is by means of an example.

The journal *Mass Communication and Society*, volume 6 issue 4, includes an umbrella article called, "Book Reviews," which unsurprisingly includes many book review subarticles. Provider 1 sent a record for the entire umbrella article with a start page of 453 and end page of 461. Provider 2 sent records for the individual subarticles. If we take the subarticle, "American Television News: The Media Marketplace and the Public Interest," the start page is 457 and the end page is 458.

If a user discovered the record from Provider 2 but had access to full text from provider 1, the linking could very well fail. The OpenURL sent from Primo to the link resolver would include the start and end page for the subarticle. These values would then be sent in the TargetURL to Provider 1's platform. However, since Provider 1 indexes the entire umbrella article and not the subarticles, Provider 1 would, accordingly, expect the start and end pages of the entire article and not the subarticle, thus causing the link to fail.

To work around this issue, Ex Libris created a feature called "Source to Target Matching" for its link resolver, SFX. This functionality allows libraries to define the preferred Target to be the provider of the source record if full text is available to the library from that provider. This will minimize the number of failed links that occur as a result of the above issue since it is more likely that Primo will send metadata that will result in a successful link if the Target is the same as the Primo Central data provider.

These few examples are merely a glimpse into the challenges faced by discovery systems when attempting to integrate content from many resources. In general, the solution to overcoming these challenges can be broken into three directions:

1. Rigid standards
    a. Discovery systems create their own standards to normalize data regardless of how it appears when initially received.
    b. We rely on the industry to set standards and on publishers to abide by these standards to minimize the amount of manipulation needed.
2. Cooperation with data providers—of course, positive relationships with providers is critical. These relationships encourage data providers and discovery providers to tweak their own processes to better serve the actual data.
3. Technological enhancements—technology can both help to improve data quality and provide solutions for dealing with data problems that cannot be easily solved, as was the case with the book review cataloging issue above.

In short, discovery providers face many challenges as a result of having to unify data from disparate sources. Some of these issues are easy to solve and other are more difficult. We rely on our relationships with data providers and the industry as a whole in order to provide our users with the level of service they expect.

For Libraries

Discovery related challenges associated with metadata have been apparent to libraries since well before the advent of the current generation of discovery layers. At Queen's University, prior to implementing a discovery layer, LibQUAL and other feedback consistently showed that our users ranked the ability to find information resources highest in terms of their expectations, but lowest in overall perception of services delivered. By 2010, user expectations had become informed by their experience with the tools they interacted with daily on the open web, and the rigid application and interpretation of library metadata in the traditional OPAC was increasingly seen as a barrier to access to information.

To help address this issue, Queen's implemented Summon in summer 2010. It was a deliberately

streamlined implementation taking approximately eight weeks, and this early feedback ranging from very positive to very negative, prompted the formation of a discovery layer assessment project:

> "SUMMON = AMAZING!!!!!"
>
> "As a graduate student heading into my 'research & paper writing' year I am pleased to see the efforts being spent on making the library tools as user-friendly and intuitive as possible. Thanks for the investing in this area of Queen's infrastructure!!"
>
> "Not only is Summons an idiot version for searching, it doesn't work."

The two primary goals of the discovery layer assessment were to:

1. Investigate how students, faculty, and library staff are using Summon to determine its impact at Queen's.
2. Recommend best practices for incorporating Summon into our broader suite of research tools, and evaluate the role of a discovery layer at Queen's.

As part of the assessment, in 2011-12 we worked with the University's Office of Institutional Research and Planning to develop and conduct a survey for students, faculty, and library staff. Much of the feedback from this can be tied to metadata challenges with the discovery layer.

Undergraduate Student Feedback

Undergraduate feedback was quite positive. Where we saw negative feedback, it generally wasn't about the ability to find articles but about problems getting from the discovery layer to full text. Undergraduate feedback included the following comments, highlighting common points raised by this group:

> "It's on the front page and always just finds what I'm looking for with zero effort on my part."
>
> "Easy to use. It brings up relevant information and is very helpful in finding academic sources to complete course assignments."
>
> "There have been numerous times that full text of an article is not available online even if it is indicated that this is the case."

Looking at issues of the sort raised in the final quote, we found that the problem was often that records didn't contain the metadata needed to generate an OpenURL that would successfully link to full text. Underlying problems include insufficient metadata, incorrect metadata, and inconsistent application of metadata—all issues raised by my copresenters.

In many cases it simply boiled down to the way that different vendors interpret the OpenURL standard. In 2011, Serials Solutions stopped depending solely on OpenURL and began linking directly to full text for many providers. This has resulted in a significant decrease in broken links, but removes the users' option to choose between multiple providers in the OpenURL resolver where we subscribe on more than one platform.

Graduate Student Feedback

Graduate student feedback was also generally positive, but more critical of the structure of search results. A reoccurring theme in graduate student feedback was the request to improve relevance. Graduate student feedback included the following comments:

> "The option to type in keywords without having to modify them by using asterisks and symbols I'm not familiar with and tend to forget . . . is simply superb."
>
> "Improve the relevance function because sometimes relevant articles don't appear near the top of the search results."
>
> "Irritating to have done a search which results in lots of hits, only to find that many of them are just citations which are not in the library's collection."

The underlying metadata challenge in the second comment is that of creating a unified index from records with vastly differing levels of quality. When records with full text indexing appear high in the results list, but users don't see their search terms in the metadata that's displayed in on

screen, it reduces their confidence in the discovery layer. It's also difficult for librarians and other staff at public service points to explain these results to users, and having to fall back on an explanation that the term must be somewhere in the full text isn't very satisfying. Thankfully we've seen significant improvements in relevance ranking in since 2011.

The last comment here points to a metadata related challenge we face about whether to continue subscribing to A&I indexes, and if so how to integrate them into the discovery layer when they don't really respect the "limit to resources outside of the library" filter.

Postgraduate Student Feedback

Postgrads had positive feedback about the interdisciplinary nature of the discovery layer, and how it provides a good general starting place for research.

They also pointed out a metadata challenge around comprehensiveness. Postgraduate student feedback included the following comments:

> "I like having a centralized search tool that searches a range of source material. With so many discipline specific and complementary journal sources to choose from, as well as printed material, its often hard to start anywhere but a general search."
>
> "It would be helpful to know how comprehensively it searches, so I could get a sense of whether I am missing information out there on a topic."

It's difficult to gauge the scope of the index in our discovery layer since search result don't indicate when a well-known resource related to a topic has not been indexed. With the Open Discovery Initiative, we hope to see better and more open relationships between discovery layer vendors and the information providers whose resources they represent so that information about resources we subscribe to isn't kept out of discovery layers because of competition between vendors. Another issue with comprehensiveness is our inability to get metadata for many resources we subscribe to, particularly e-book and multimedia packages. This can have less to do with competitive practice, and more that producing quality metadata is an afterthought, especially for vendors of new and emerging formats.

Faculty Feedback

Faculty feedback was similar to the students, but more critical. They noted the difficulty in being able to limit results to particular formats. Faculty feedback included the following comments:

> "It is much faster for me to find information using Summon."
>
> "Summon is great for cross-disciplinary research."
>
> "I find useful information using Summon, but not all papers that I find are readily available online."
>
> "Often I'm looking for an author and year of publication (e.g., Smith, 2005) and Summon turns reviews or articles that its Smith, 2005. Just give me Smith, 2005 please!"
>
> "Sorry, I don't know what Summon is."

The metadata challenge in the fourth quote is one noted by Rachel in her presentation—that of being able to distinguish between articles and reviews where we can't count on high quality metadata being available in the index. Although it isn't represented in our survey comments, another issue we've noticed with faculty is that their knowledge of the literature in their field makes them more aware than students of when they're facing metadata issues in the discovery layer. They are more likely to notice when articles from the most recent issue of a key title in their field are missing from results lists, making them more likely to dismiss the discovery layer and go to subject specific databases. The metadata issue we face here is with the connection between publishers, aggregators, and discovery layer vendors. An important element of the cooperation recommended by the ODI is the timeliness of information sharing between content providers, discovery layer providers and libraries, and standardization of the methods of

sharing. No one should have to screen scrape records to populate an index.

Finally, one of the main differences between student and faculty feedback was that many faculty hadn't heard of the discovery layer. In many cases this was because it hadn't been promoted by library staff, who were finding the transition from the OPAC to the discovery layer challenging. For many this is because they were expert users of the OPAC and familiar with our ILS and cataloging standards, but not as well-versed in electronic resource management and e-resource troubleshooting.

Staff Feedback

Whereas most library staff felt comfortable helping users experiencing problems with our OPAC, many reported that they were often unable to explain what was happening when students and faculty came to service desks to report issues they had while using the discovery layer. Staff feedback included the following comments:

> "Where I do find Summon to be useful is as a broad discovery tool."
>
> "It's a great first step to help me figure out more targeted, sophisticated searches."
>
> "Rarely used for teaching or reference work—unpredictable, lacks precision, confusing links to resources—not a pleasure to show students."
>
> "There are a lot of mysteries in the functioning of this software in our environment."

Whereas library staff are generally well versed in the way metadata is applied in the search results of our OPAC, many of our library staff don't feel that they possess expert knowledge of how content in the discovery layer is indexed or how relevancy is determined, and as a result don't feel as confident in their ability to play the role of an expert user with Summon.

One of the most common issues that has been raised was that they are unable to determine what is and isn't indexed and how frequently new content is added to the index. The tools provided by our discovery layer vendor to show users what content they've indexed are not user friendly, and don't give an indication of how quickly new content is added after publication. Like faculty, library staff have been most likely to avoid the discovery layer when they know we subscribe to content and are aware of new articles that should be available in Summon, but are unable to find them in results lists. If ODI recommendations for consistent and transparent methods of content exchange are respected, it should be possible to have new content indexed in all discovery layer platforms as soon as possible once it is published, which would ameliorate this issue.

Library staff recommendations for improving the discovery layer include the following comments:

> "Better indication of why results are being retrieved."
>
> "Better indexing, particularly when harvesting records from QCAT. Serials Solutions MARC records are sometimes quite minimal with no subject headings."
>
> "Get MARC records in QCAT (and hence Summon) for e-books faster than currently is the case."

These comments are quite relevant to the issues being addressed by the recommendations of the Open Discovery Initiative, and we are hopeful that if all parties follow those recommendations, that many of the metadata related barriers to effective use of our discovery layer will be removed, making it an effective tool in our broader suite of information resources.