Information Services & Use 43 (2023) 269–274
DOI 10.3233/ISU-230207
IOS Press

Reframing the returns on metadata investments through discoverability, usage, and analytics

Michelle Urberg[*]
Independent Consultant, Seattle, WA, USA

Abstract. The interactions among metadata, content discoverability, and content usage have been undervalued by publishers for years. Despite the regular exhortations by librarians to encourage publishers to develop a good metadata program, it remains very difficult for them to make a robust business case for a positive return on investment from metadata creation and maintenance. However, with the ever-growing body of open access research and the production of datasets, blogs, websites, and other unique research outputs, it is an ideal time to revisit how metadata directly affects discovery and usage in the scholarly communications ecosystem. Certain pieces of metadata make demonstrable differences for helping or hindering end-user engagement, can increase or suppress usage, and, therefore, can be one of the keys to driving the business case for supporting open access and non-traditional scholarly outputs. Better metadata is key to gathering robust analytics on user-engagement.

Keywords: Metadata, discovery, analytics, user engagement, open access

1. Introduction

It is hard to build a business case for creating a positive return on investment for metadata creation and maintenance. It is time consuming to manage internally and even more to make sure it is packaged and shared externally. As for correcting mistakes? How can one date error, an unidentifiable IP address, an incorrect DOI, or a title typo really affect anyone demonstrably in the scholarly communications ecosystem? The answer is this: it is impossible to measure at the micro-level. At the macro-level, however, the impact of metadata on research outputs can be felt by everyone in the scholarly communications ecosystem.

Metadata mistakes can render content inaccessible, undercount usage, misrepresent the usefulness of research, and hinder a content provider from building a use case to invest in a new series or develop a new product. The business case for better metadata is, moreover, notoriously elusive simply because it is hard to quantify the direct benefit of fixing small mistakes or ensuring that one piece of information is accurately captured in a data pipeline. Despite these challenges to building a case for better metadata creation and maintenance, several key pieces of metadata demonstrably work in tandem to produce better experiences for end users (i.e., researchers and readers). These pieces of metadata, when handled well,

[*] E-mail: maurberg@gmail.com.

will help to boost usage and ideally contribute to product investment. In light of the ever-increasing body of open access content and the growing interest in multi-modal research outputs (e.g., datasets, websites, video), now is an ideal time to revisit how to build a positive return on investment for metadata creation and maintenance. What follows is a literature analysis tracking how librarians and publishers have discussed metadata for discovery and usage, which dovetails nicely into the findings of a recent research study on book and book chapter discoverability. The scholarly communications industry has long acknowledged that better metadata makes libraries and publishing work better; this article proposes steps toward rethinking the role of metadata in business decisions [1].

2. A brief history of metadata ROI

The discussion about metadata and return on investment is not new to scholarly communications, but it has been couched in discourse about discovery or sales. Librarians in particular have been heralding the importance of metadata in content discovery for decades. Where librarians continue to focus on metadata as a driver of discovery, publishers have traditionally needed metadata to boost sales of content. Over the past twenty years, the interest by librarians in content discovery has gradually expanded into measuring that discovery by tracking content usage. Discovery and usage are, necessarily, cozy bedfellows: you cannot have one without the other. I argue that publishers lag behind librarians in this area; sales and usage are similar bedfellows to discovery and usage. BOTH discovery and sales are driven by metadata indexed in catalogs, platforms, and the open web. Where to begin then? While not exhaustive, here are a number of resources I have found helpful to track the relationship between metadata, discovery, and content usage in the scholarly communications industry.

Librarians have long been promoting better publisher metadata to improve discoverability. They have discussed the use of bibliographic metadata (title, author, date, page, publication, etc.) in MARC records, OpenURL link resolvers, and library discovery layers to ensure the highest level of discovery [2]. Certain fields in MARC records have also been found to enhance discoverability and bump COUNTER stats, such as the 505 (table of contents) and 520 (summary) notes fields [3]. More recently, librarians have broadened the focus of their investigations to study the impact of individual metadata fields on e-book and streaming video discovery [4].

Publishers often have to be cajoled in to investing in metadata, and, moreover, the writing on discoverability and metadata has not been directed toward scholarly publishers. Commercial publishers have been the primary audience of important industry studies on using metadata to improve disoverability and sales[5]. While keywords have been championed as tools to enhance marketing strategies in the academic sector, more evidence should be gathered specifically tracking the metadata needs of scholarly publishers[6].

Librarian evangelizing has brought discussions about metadata and discoverability more directly to scholarly publishers in recent years [7]. Publishers and other service providers in the industry have responded in kind to champion collaboration as a key tool to improve metadata for discovery [8]. When publishers and other service providers have turned their attention to improving metadata, DOIs (and associated metadata) are of particular interest [9].

Recent work on metadata and return on investment focuses on end-user engagement [10] and value-based assessment options for improving open access book usage [11]. Both librarians and publishers seem to be shifting their attention toward improving open access analytics rather than further enhancing the assessment of gated content. At least a couple of new dashboards have been released that track and analyze open access content creation and usage by country [12,13]. And, open access book usage is, for the first time, being used to improve access to open access book publishing models [14].

This short literature review highlights several things about metrics for scholarly communications:

- First, over the past decade, industry interest in metrics has been gradually changing in the scholarly communications landscape, from focusing primarily on tracking controlled (gated) access content usage to including the tracking of open access content.
- Second, along with this shift in tracking content usage, the parties interested in cultivating better usage data have gradually expanded from librarians to include publishers or publisher-adjacent service providers. These are anecdotal observations, of course, but in general, interest in content usage has increased across all stakeholder groups - librarians, publishers, service providers, researchers - in the scholarly communications ecosystem. Publishers in general still lag behind librarians in understanding the power of metadata to drive business decisions (librarians have been relying on COUNTER to drive purchasing decisions for decades already).
- Third, nearly all of the research focuses on the discovery and usage of traditional outputs (i.e., subscription-based books and journal articles). The next horizons will be to meaningfully measure the ever-increasing amount of open access content (which does include books and articles) and the research outputs adjacent to traditional book and journal publications (e.g., datasets, video content, websites). All stakeholder groups, and now research funders - another very recent group to come to the table - will have an interest in understanding the broadening usage of scholarly ouputs [15].
- Fourth, the literature to date does not highlight one piece of metadata as being the key to unlocking better return on investment. Instead, return on investment seems to accrue differently depending on who is tracking the content. That being noted, however, publishing workflows that create metadata inevitably shape the information experience of all stakeholders in the ecosystem. And, ultimately, a few key metadata fields, good indexing choices, and publisher commitments to improving content engagement are the best indicators of positive returns on investment from any metadata.

3. What metadata matters?: A study on books discoverability

Lettie Conrad and I recently finished a project, sponsored by Crossref, that explored the impacts of metadata on book and book chapter discoverability in Google Scholar [16]. We decided to use Google Scholar because its contents are publicly available, with fairly predictable search behavior and because we have a fairly high degree of confidence how it ingests and processes metadata. We started with the end user in this study, the person who is discovering the content. Why end users and not publishers or content providers? Two reasons: First, end users are the key to discoverability, one of the primary concerns in the discovery-usage relationship. Their experience (and ultimately usage) is driven by metadata. Second, any value-based engagement with scholarly content begins with end users.

Our study, which was as much about the metadata as it was about user engagement, examined the discovery of books and book chapters indexed in Google Scholar. We found that certain pieces of metadata matter more than others for enabling end-user discovery. When attempting to find a particular book, DOIs - perhaps the most important industry-accepted persistent identifier (PID) - matter, especially for the title-level searching, but DOIs alone are not enough. Titles (including subtitles) and author surnames are also key for finding books within a search string. Uncontrolled subject terms are less important for book-level discovery, but fields of study can be helpful with author disambiguation. Publisher names are the least helpful with discovery-oriented searching.

While books benefit the most from DOIs, titles, and author searches, these pieces of metadata do not offer book chapters a correlative boost in discoverability. The weak impact of DOIs, titles, or author metadata on chapter discoverability likely arises from a lack of systematic handling of metadata for book chapters and gaps in metadata standards to create connectivity between chapters and their parent books. Without end user discoverability, usage will be suppressed (or simply not tracked) and without demonstrable usage, it is inevitably harder to build a business case for promoting a series or a title.

Our findings further the research and analysis in the history of metadata ROI. We know particular pieces of metadata drive discoverability and therefore are a factor in end-user engagement. We also know that with engagement comes usage and the opportunity to demonstrate how particular types of content drive value for a publisher. Book chapter discoverability in our study highlights a clear gap in the scholarly communications industry with creating and using standardized metadata. Key pieces of metadata are either not indexed or are subsumed in the parent book, making chapters hidden from the end user. Moreover, book chapters are a kind of proxy for other kinds of outputs in the ecosystem, including datasets, video, websites, blogs, or any other form of scholarly communication that is not a gated book or an article. Like book chapters, these outputs likewise lack an industry-agreed-upon best practice to allow them to simultaneously stand alone with a unique record and be linked to parent metadata. For improved user engagement with book chapters, data sets, or other types of publications other than the book or the article, our industry needs to plan to assign standard metadata to these outputs that promotes discoverability.

4. Planning for increased engagement

The key pieces of metadata discussed in the books study - DOIs plus titles and authors - should be core to production and distribution workflows regardless of output type. We know that the experience of metadata can be frictionless. When metadata is doing its job, you will likely not see what it is doing. Moreover, when metadata is detailed and accurate, it is part of a data-rich pipeline for everyone in the ecosystem. When considering whether your metadata is maximizing your content usage, the data pipeline, friction, and points of engagement should be examined in detail. These questions can help you get started:

- What are your key research outputs and how do they relate to versions of record or main publications? Are you providing metadata to discover them together as well as separately? Do you need metadata for books or articles, as well as for data sets or videos? The reality of creating metadata for different formats and keeping content linked is using relational types of metadata such as: “IsPublishedIn” or “IsCitedBy”. How do your schemas and DTDs [17] accommodate that currently? Data Cite has a great set of information regarding linking of data to other outputs [18].
- Where do you find friction in your system that slows the flow of metadata for discovery? Is it in creating item-level metadata? Is it in working with vendors? Is it in getting good information back about your content usage? Focus on what you can control inside of your organization and what you send out.
- What metadata are you providing across channels? Is it sensitive to the needs of users of each channel, while still being robust? The research study discussed in this article showed that while a DOI does not significantly boost discoverability, there is a correlation between the DOI and discoverability. Publishers who do not assign DOIs also tend to provide limited metadata on their websites.
- What information is missing in your pipeline to make business decisions about investing in metadata? What data do you need to make product decisions? Can you tell if key metadata fields, especially titles, are indexed correctly in your distribution channels?

- What data needs to flow back to your organization? Robust analytics and usage data, for example, can be key to making business decisions, but publishers need to have an easy way to access this usage data.

5. Concluding thoughts

Based on this review of literature regarding discoverability, usage, and metadata, more work can still be done both tactically to improve discoverability and usage, and strategically to research what different pieces of metadata contribute to a robust picture of end-user engagement. From a tactical perspective, discoverability is still the first problem to resolve, because discovery drives usage. If your DOIs, titles, and author statements are well-maintained, then your next task will be to establish a channel strategy and routine distribution. This becomes a foundation for developing ROI metrics for metadata based on content usage data. Better metadata will provide the way forward to generate trackable usage in platforms, and ideally better product development decisions.

From a strategic perspective, publishers could, for example, replicate the books study directly by analyzing end user experience of open access content or indexed data sets across distribution channels. Other metadata fields can be added to a study, such as ISBNs, ORCIDs[19], or RORs [20]. Based on the success of these queries, effectiveness of metadata can be tracked across these channels. With the right metadata a publisher should even be able to understand end-users: What do you want to know about your users? Where are they located and what content are they interested in? Better metadata is key to gathering robust analytics on user-engagement.

Acknowledgement

Many thanks to Lettie Conrad for reading and commenting on a draft of this article.

About the author

Michelle Urberg is an Independent Consultant with Data Solve LLC and the Client Success Manager at LibLynx LLC. She holds a PhD in Music History and an MS in Library and Information Science. Her work can be found at the Humanities Commons: https://hcommons.org/members/murberg/ and at ORCID: https://orcid.org/0000-0002-2748-8. Email: maurberg@gmail.com.

References

[1] This article is developed from the author’s presentation given at the 2023 NISO Plus Conference on February 14, 2023: M. Urberg, The Impact of Metadata on Research Outputs, available from: doi:10.6084/m9.figshare.22196875.v1, accessed September 30, 2023.

[2] S. Shadle, How libraries use publisher metadata, *Insights: The UKSG Journal* **26**(3) (2013), 290–297. doi:10.1629/2048-7754.109, accessed September 30, 2023.

[3] C. Sassen and K. Harker, Cataloging that Works: How to Make E-books Findable, University of North Texas Libraries, UNT Digital Library, 29 June 2014. Available from: https://digital.library.unt.edu/ark:/67531/metadc306050/, accessed September 28, 2023.

[4] N. Trujillo, E. Radio and M. Walker, What metadata matters?: Correlation of metadata elements with click-through rates for e-books and streaming video in the academic library catalog, *Journal of Web Librarianship* **14**(3–4) (2020), 86–99. doi:10.1080/19322909.2020.1850390, accessed September 30, 2023.

[5] The Nielsen Company, The Importance of Metadata for Discoverability and Sales, 2020. Available from: https://www.booksonix.info/wp-content/uploads/2021/09/Nielsen-Metadata-Marketing-Report.pdf, accessed November 24, 2023.

[6] T. Carpenter, Enriching Book Metadata is Marketing in the Digital Age, *The Scholarly Kitchen*, (December 7, 2017). Available from: https://scholarlykitchen.sspnet.org/2017/12/07/enriching-metadata-is-marketing/, accessed November 24, 2023.

[7] S. Bull and A. Quimby, A renaissance in library metadata? The importance of community collaboration in a digital world, *Insights: The UKSG Journal* **29**(2) (2016), 146–153. doi:10.1629/uksg.302, accessed September 30, 2023.

[8] J. Kemp, C. Dean and J. Chodacki, Can richer metadata rescue research? *The Serials Librarian* **74**(1–4) (2018), 207–211. doi:10.1080/0361526X.2018.1428483, accessed September 30, 2023.

[9] J. Kemp, Metadata and discoverability: A use case overview, *Information Services & Use* **38**: (2018), 81–84. doi:10.3233/ISU-180004, accessed September 30, 2023.

[10] M. Urberg, Creating return on investment for large-scale metadata creation, *Information Services & Use* **41**(1-2) (2021), 53–60. doi:10.3233/ISU-210117, accessed September 30, 2023.

[11] E. Farrell, E. Poznanski and C. Watkinson, Building a framework together for assessing OA book publishing models, *Commonplace* **1**(3) (2021). doi:10.21428/6ffd8432.185d3532, accessed 28 September, 2023.

[12] J.P. Diprose, R. Hosking, R. Rigoni, A. Roelofs, T. Chien, K. Napier et al., A user-friendly dashboard for tracking global open access performance, *The Journal of Electronic Publishing* **26**(1) (2023). doi:10.3998/jep.3398, accessed September 30, 2023.

[13] Open Access Book Usage. University of Michigan Press 2023. Available from: https://ebc.press.umich.edu/impact/, accessed September 28, 2023.

[14] C. Watkinson, The Good, Bad, and Ugly in Open Access Humanities Monographs, NISO Humanities Roundtable 2023. Available from: https://www.niso.org/niso-io/2023/07/good-bad-and-ugly-open-access-humanities-monographs, accessed September 28, 2023.

[15] The Nelson Memo 2022. Available from: https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf, accessed September 30, 2023.

[16] L.Y. Conrad and M. Urberg, With or Without: Measuring Impacts of Books Metadata, Zenodo 2023. doi:10.5281/zenodo.8145260, accessed September 30, 2023.

[17] See: https://wikipedia.org/wiki/Document_type_definition, accessed September 30, 2023.

[18] DataCite Metadata Schema 4.4 202. Available from: https://schema.datacite.org/meta/kernel-4.4/, accessed September 28, 2023.

[19] The Open Researcher and Contributor ID (ORCiD) is a global, community-led registry of persistent identifiers for researchers, see: https://orcid.org/, accessed November 24, 2023.

[20] Research Registry Organization (ROR) is a global, community-led registry of open persistent identifiers for research organizations, see: https://ror.org, accessed September 30, 2023.