Exploring Open Access Ebook Usage

BISG

Statement of Authorship

This white paper was prepared by the Book Industry Study Group (BISG) as part of the Andrew W. Mellon Foundation funded project, Understanding OA Ebook Usage: Toward a Common Framework. Primary authors are: Brian O'Leary (BISG) and Kevin Hawkins (University of North Texas). The project team, who contributed editing and improvements, include Charles Watkinson (University of Michigan), Lucy Montgomery (Curtin University/KU Research), Cameron Neylon (Curtin University/KU Research), and Katherine Skinner (Educopia Institute). Copyright for this white paper is held by BISG and licensed to the general public under a Creative Commons Attribution 4.0 International license.

Published 6 May 2019

How to Cite This White Paper

Brian O'Leary and Kevin Hawkins, *Exploring Open Access Ebook*, Book Industry Study Group, May 2019, https://doi.org/10.17613/8rty-5628.

Executive Summary

A recent research project led by the Book Industry Study Group in collaboration with KU Research, the Educopia Institute, and researchers from the University of Michigan and University of North Texas Libraries identified the challenges in understanding the usage of open-access (OA) scholarly ebooks, suggested some opportunities for resolving them, and created a framework for future action through community consultation. The project proposed the potential development of a "data trust" as a vehicle to manage the multiple data sets that are key to understanding OA ebook usage while respecting commercial and individual user concerns.

A data trust operates as an independent intermediary among industry stakeholders, compiling and analyzing data on behalf of trust members.[1] Members of a data trust for OA monograph usage data would agree to make their data available to others who are members of the trust. Members would access normalized data through a user-specific dashboard or interface, while the trust would provide benchmarking data in a manner that respects contributor confidentiality and privacy. The data trust could also allow certain anonymized data to be extracted, typically through an agreed-upon API, for independent analysis.

Comprehensive access to usage data has the potential to provide stakeholders with valuable strategic insight.

Comprehensive access to usage data for OA monographs has the potential to provide all stakeholders in scholarly communication—from scholars and their institutions to publishers, content aggregators and platforms, and research funders—with valuable strategic insight into how and where OA books are being used. The ability to benchmark and understand usage data in the context of wider patterns and trends depends on access to aggregate data from multiple stakeholders; individual parties are unlikely to have this kind of access. Furthermore, a data trust helps lower the cost in staff expertise and resources for individual stakeholder organizations to engage in data analytics.

Successful collaboration around data sharing requires thoughtful engagement with issues of trust between stakeholders, the development of shared technical standards, and the development of requirements for the validation of data and information. This is a classic collective-action problem. Its solution, therefore, requires the development of a trusted framework for coordination between all the relevant stakeholders. Our recommendations address these aspects of successful collaboration.

Relevant research and initiatives around OA ebook usage are currently conducted separately in the United States and Europe, by both for-profit and nonprofit entities. The HIRMEOS project in Europe (part of the broader OPERAS framework) has been

[1] Nic Suzor and Joanne Gray, "What is a data trust?", http://dx.doi.org/10.17613/gxa6-mg85.

particularly influential. Coordinating or connecting those efforts, as well as improving our understanding of needs in other regional markets, is a priority for future efforts. The key recommendations for future work are the following:

1. **Define the governance and architecture** for the data trust and articulate priorities.
2. **Create a pilot service** that implements the defined governance and architecture.
3. **Implement and extend relevant open-source technologies** across a base of stakeholders in the US.
4. **Develop personas and use cases that demonstrate who benefits** from OA monograph usage information and how a data trust can better serve their needs.
5. **Build engagement** across multiple markets.
6. **Better document the supply chain** for OA monographs.

This white paper provides detail on work to date and these recommendations.

A data trust operates as an independent intermediary among industry stakeholders.

Examining usage of OA monographs

Publishing of scholarly books (monographs) has long involved a range of stakeholders — authors, publishers, funders, vendors, libraries, and readers — with values and challenges to their viability that pre-date the digital age. These stakeholders are adapting to a landscape that includes online access, digital formats, and open-access (OA) possibilities. The new landscape is forcing a reassessment of strategic goals for all stakeholders.

Advocates for open scholarship suggest that an OA monographs will be more often downloaded, used, and cited than a comparable restricted-access title. Publishers need to demonstrate such impact to receive support for their OA publishing programs. Funders look for usage data to demonstrate return on their investments, and authors are eager to show evidence of additional reach and influence for their work.

These stakeholders face challenges identifying and aggregating relevant information from different platforms. Information about the impact of academic ebooks, especially OA books, is much more difficult to gather, analyze, and communicate than comparable information about online scholarly journals, for which publishing is dominated by a small number of publishers and infrastructure providers with a widely used system of stable identifiers (DOIs). A central issue is that book publishers do not use DOIs comprehensively or consistently. Stakeholders also encounter difficulties analyzing any collected data in ways that respect user privacy and communicating relevant information about usage to other stakeholders.

Stakeholders face challenges identifying and aggregating relevant usage information.

Despite the challenges, stakeholders are working to capture data and articulate the value of investments in OA monographs. As the number of published OA monographs grows, the need for data about their impact also increases. Because OA monographs are openly licensed, they can be redistributed widely, meaning users engage with the books across multiple sites and formats. Granular and comparable information on users and usage of OA monographs has the potential to support OA publishing by informing the acquisition, marketing, and sustainability strategies required to meet the new opportunities and demands of an evolving scholarly communication ecosystem.

Comprehensive access to usage data for OA monographs has the potential to provide all stakeholders in scholarly communication — from scholars and their institutions to publishers, content aggregators and platforms, and research funders — with valuable strategic insight into how and where OA monographs are being used.

If well-managed at a community level, OA monograph usage data could provide:

- Insight into the relative performance of individual books and collections
- Benchmarking and tracking of changes in patterns of use over time
- Information about subject-specific patterns of use for OA monographs
- The ability to map the communities engaging with OA monographs
- New tools for evaluating and communicating the value and performance of OA monograph publishing

These opportunities matter to organizations that publish monographs as well as those that host and distribute digital content or that provide metadata about monographs. OA is creating opportunities for monographs to reach new audiences, but new business models are requiring publishers and other stakeholders to articulate anew the value of investments in publishing and dissemination to new financial supporters of scholarly publishing and to old financial supporters in new ways. In this context, information about who is using content and how they are engaging with that content is increasingly important.

Usage data that supports diversity, quality, and impact for OA monographs must be comparable, trusted, granular, and appropriately benchmarked.

Capturing and analyzing this usage data presents a significant challenge. Data relating to OA monographs is generated at many different points within the digital landscape, and no single player has access to a complete picture of how OA monographs are being discovered and used. To provide useful information to stakeholders in monograph publishing, and to ensure the privacy and security of users, usage data must be gathered, cleaned, analyzed, and presented with skill and care. Even the largest players in the monograph space may not have staff with the technical and statistical background necessary to unpack complex relationships between OA status and patterns of use in a changing global context.

The ability to engage with usage data relating to large numbers of books and across multiple platforms in aggregate has the potential to generate beneficial network effects for all monograph stakeholders — that is, the more data that stakeholders share with one another, the more each benefits. However, direct comparisons between individual titles, publishers, and platforms must be approached with caution because naive quantitative comparison can hide many confounding factors. Association or correlation do not mean causation. At the same time, aggregate data has an important role to play in supporting benchmarking, as well as in helping stakeholders to understand the performance of an individual book, publisher, or subject area in the context of larger trends.

Delivering on the potential for usage data to support diversity, quality, and impact for monographs requires that it be comparable, trusted, granular, and appropriately benchmarked. Achieving this requires appropriate sharing of data across all stakeholder groups. This raises many challenges, which are discussed in detail below. As we note, the technical issues are largely solved problems. Now, we need to develop and agree upon goals for the stakeholder community and then select a set of systems that support achievement of these goals.

The monograph landscape in general, and the OA monograph landscape in particular, is characterized by a number of features that make the development of a community approach to the management of usage data both feasible and necessary. It is feasible because in book publishing, in comparison to journal publishing, there are no dominant players with interests significantly different from those of many smaller players. It is necessary because this diversity means that no one single player or small group is likely to act on its own to solve this problem for books.

Through research and discussion, comments on a widely-circulated discussion document[2], a summit held in December 2018[3], and further interviews and conversations with interested parties, the project team found that:

- A good deal of data is already available to those who want to study the impact of OA monographs. This data can be characterized as either available but in closed environments or available in open, accessible environments. Determining the best ways to access both types of data is an ongoing discussion.
- A number of data points and information sets about OA monographs are of interest to stakeholders but have never been compiled.
- The number of available data sets, whether in closed or open environments, dwarfs the data that is not yet available. The undeveloped options sometimes receive the greater share of time and attention even though significant data is already available.
- The data of greatest interest varies by audience (authors, publishers, funders, vendors, libraries, and readers). Across the several audiences, relatively little of the available data is being used widely or consistently.
- There are marketplace, privacy, and ethical concerns about use of certain data points: data about how OA monographs are being used may include sensitive commercial information as well as information about users that must be handled carefully in order to safeguard privacy.
- There is a standard already in place for gathering usage data in a consistent way: COUNTER. It does not provide some of the qualitative information about OA ebook usage that stakeholders want, and it has historically focused on measuring use within institutions. However, its governance group is willing and eager to adapt the standard to be more useful for OA ebooks and has already taken steps to do so with Release 5 of the COUNTER Code of Practice.

The monograph landscape is characterized by a number of features that make a community approach to managing data both feasible and necessary.

[2] Cameron Neylon, Lucy Montgomery, Nic Suzor, Joanne Gray and Alkim Ozaygen, "Building a Trusted Framework for Coordinating OA Monograph Usage Data," http://dx.doi.org/10.17613/36hw-gs17.

[3] "Exploring Open Access eBook Usage: Toward a Common Framework," http://dx.doi.org/10.17613/fpcz-gp24.

- Use cases for OA monograph discovery, access, consumption, and engagement have not been widely or fully developed. Relevant use cases, when developed, must be mapped against the needs of audiences identified above.

- Significant work is being done outside of North America, but real-time, interactive coordination with European and other international efforts has been inadequate.

- Among North American and European participants in the OA monograph value chain, there is general support for the concept of a data trust. However, there is significant debate about how to build such a trust — specifically, whether its governance and operation should be centralized, federated, or distributed.

- With respect to creating a data trust, agreements are needed in at least three areas: standards for data exchange, where and how data is stored and managed, and how analytics will be built on top of that data.

A data trust requires agreements on standards for data exchange, where and how data is stored and managed, and how analytics will be built.

The initial discussion document, the summit, and this white paper were planned and executed to help achieve five high-level objectives:

- Strengthen relationships between stakeholders in OA ebooks, both inside and outside North America

- Develop understandings among stakeholders regarding their different perspectives and goals regarding the measurement of engagement with OA ebooks

- Identify key impediments to aggregating, analyzing, and communicating information about OA ebook engagement

- Identify and prioritize activities stakeholders might engage in to lower those barriers

- Stress-test the idea of a data trust to promote cross-stakeholder collaboration, including consideration of its potential form, function, and governance and business models.

The research project also led to the following:

- An informal network of stakeholders in OA ebooks who are willing to align some of their efforts to collaboratively pursue joint goals. Several steps have been taken to build and strengthen that network.

- Recommendations for projects whose scoping documentation can be developed and refined in the near term by stakeholders to pursue. A number of these will include and build on initiatives that are already under way in Europe.

Detailed recommendations are provided in the next section.

Recommendations

These recommendations build either on existing efforts or proposed projects whose results would catalyze support for and the effectiveness of a data trust.

1

Define the governance and architecture for the data trust and articulate priorities.

Discussions identified three possible architectures for the data trust: centralized (a single repository with analysis and reporting tools built in), federated (a set of repositories with interchanges that may be tightly or loosely defined), and decentralized (data is broadly distributed, with standards set to facilitate interchange and analysis).

The choice of architecture affects priorities and governance. For example, a federated model would require a focus on standards and software rather than the repository service. The choice of architecture also leads to definition of near-term and mid-term steps for realizing a data trust.

To make these decisions, we recommend convening an advisory board composed of representatives from all parts of the OA monograph value chain. The advisory board should commission a focused discussion draft of governance and architecture alternatives, distributed to a wide audience, followed by a governance summit of key stakeholders for deliberation. The final decision for structure and priorities would rest with the advisory board, though we still imagine the governance model being refined during a pilot period for the data trust (see recommendation #2 below).

2 Create a pilot service that implements the defined governance and architecture.

The initial discussion document included an outline of contractual components defining supply of data, access to data, use of data, membership and termination, and governance. A pilot effort would provide a range of stakeholders with an opportunity to implement the governance model and priorities outlined in recommendation #1 above.

The pilot would require involvement of personnel with advanced technical skills, experience managing large (and evolving) data sets, and the ability to engage actively with a wide range of stakeholders. Familiarity with OA issues, even beyond those specific to monographs or book content, would help ground this effort to improve data collection, analysis, and reporting around OA monographs.

However designed, a data trust depends on core principles of security, usefulness, and fairness. Any pilot or implementation must be conducted in an open, inclusive, and balanced way across all stakeholders. The pilot effort would work to test the effectiveness of standards and verify the interoperability of data across multiple sources and uses, consistent with the principles established for governance and architecture.

3

Implement and extend relevant open-source technologies across a base of stakeholders in the US.

The pilot service (recommendation #2) could be either followed or augmented by having multiple interested OA monograph publishers implement a set of open-source technologies. This kind of collaborative approach—having different parties install and test the same infrastructure—is one that University of Minnesota's Manifold project, for example, has taken as a path toward obtaining both buy-in and near-term validation.

These tools are increasingly available. The HIRMEOS project, centered in Europe, offers a basis on which to build community and sustainability for a data trust. The initial discussion document, the summit, and discussions with stakeholders confirm the promise that HIRMEOS brings to the North American market.

Work required need not start from scratch. HIRMEOS and Open Book Publishers have already established links to various external sources, including OAPEN, JSTOR, and Unglue.it. Further, HIRMEOS has set up a mechanism for maintaining and continuing to develop these data links. In a similar way, Jisc's JUSP and IRUS services can be seen as data trusts whose experience can shape future efforts.

We recommend developing ongoing partnerships with other organizations that have solved or are closer to solving these data-access and data-sharing challenges. The architectural components of a data trust (recommendation #1) and the pilot service (recommendation #2) can help identify potential partners whose capabilities can solve a technical, operational, or data challenge.

4 Develop personas and use cases that demonstrate who benefits from OA monograph usage information and how a data trust can better serve their needs.

Participants at the summit identified personas and use cases as core components of effective design of a data trust and related tools. These use cases can provide an effective filter in defining the data types, analytical tools, reporting tools, and export features that are part of any data trust. The value of both usage data and qualitative data on the impact of usage can be framed using personas as a filter.

Several use cases can be developed based on the US and European experiences represented in the development of the initial discussion document, the summit, and this white paper. These provide important reference points for the initial design and implementation of the data trust.

With some priority, solicitation, and development of personas and use cases outside of the North American and European markets is also needed (see recommendation #5 below). As scholarly communication grows more global (and digital), the utility and sustainability of the data trust will be measured at least in part by its ability to meet the needs of those in emerging markets.

5

Build engagement across multiple markets.

The recommendations proposed here focused on bridging stakeholders in the US and Europe. These initial efforts offer important building blocks, but they do not fully reflect the wider universe of OA monograph creators, publishers, funders, and consumers.

To address this gap, we propose work to convene focus groups whose membership includes representatives outside of North America and Europe to seek understanding and clarity on the minimum return that each organization would need in order to participate fully in a data trust. Outreach must also be structured to create use cases that capture what may be unique aspects of the OA monograph ecosystem outside of North America and Europe. Scholarly publishing in other markets may also require different features from the data trust.

6 Better document the supply chain for OA monographs.

As noted elsewhere in this white paper, OA monograph publishing involves authors, publishers, funders, vendors, libraries, and consumers. Each of these stakeholders has a role to play in making OA monographs discoverable, relevant, accessible, and consumable.

A supply chain is a network organized to create, manage, and distribute a product. Often represented by the steps taken to deliver a product to an end user, it is typically optimized for cost (e.g., cheapest alternative), speed (e.g., fastest to market), or value (e.g., quality or exclusive access).

The traditional book industry supply chain uses price as a proxy for value, and investments (including author payments) are based on expected sales. OA monographs have no comparable proxies, and their value is tracked using separate measures and data repositories. As a result, useful data about OA monographs gets lost or underused.

To identify pain points and understand where such data gets "lost", participants in the summit identified an opportunity to create a map of the supply chain for OA monographs. Building on the work of an earlier project[4] that demonstrated the complexity of this issue, this map will help define a value-based model that supports discovery, access, and consumption of these titles. It will also identify any gaps in the existing framework, bringing a wider mix of stakeholders to the table.

[4] Charles Watkinson, Rebecca Welzenbach, Eric Hellman, Rupert Gatti, and Kristyn Sonnenberg, "Mapping the Free Ebook Supply Chain: Final Report to the Andrew W. Mellon Foundation," http://hdl.handle.net/2027.42/137638.

Acknowledgements

We are very grateful to the Andrew W. Mellon Foundation for supporting this work, the many people who responded to our online survey or commented on our initial working paper, the participants in the New York summit, and important stakeholders who were willing to be individually interviewed. We are particularly grateful for the collegiality and engagement of members of the HIRMEOS project and colleagues at UKRI and JISC in the UK. Since publishing is a global activity, engaging in a complementary way across national boundaries is particularly important in supporting a more holistic understanding of engagement with open access books.