Building a Trusted Framework for Coordinating OA Monograph Usage Data

Statement of authorship

This report was prepared by KU Research as part of the Andrew W. Mellon Foundation funded project *Understanding OA Ebook Usage: Toward a Common Framework*. Primary authors were: Cameron Neylon, Lucy Montgomery, Nic Suzor, Joanne Gray and Alkim Ozaygen. The project team, who contributed editing and improvements, included Kevin Hawkins (University of North Texas), Charles Watkinson (University of Michigan) and Brian O'Leary (BISG).

5 November 2018

Problem statement

Stakeholders in monograph publishing are adapting to a landscape that includes online, digital, and open access possibilities, requiring a reassessment of their strategic goals. In particular, stakeholders face challenges in moving beyond reporting of financial transactions (print and e-book sales, fee-based licensing) toward capturing and articulating the value of investments in open access (OA) monographs in the context of users who engage with their books across multiple sites and formats. Granular and comparable information on users and usage of OA monographs has the potential to support these stakeholders in adapting their acquisition, marketing, and sustainability strategies to the new opportunities and demands of an evolving scholarly communication ecosystem.

However, data about how OA books are being used may include sensitive commercial information such as sales figures of print-on-demand copies, as well as information about users that must be handled carefully in order to safeguard privacy. In addition, there are also major issues of scale: the resources required to collect and analyse this data are often cost-prohibitive for individual stakeholders, and the possibilities of benchmarking and understanding usage data in the context of wider patterns and trends depends on access to aggregate data from multiple stakeholders, which individual stakeholders are unlikely to have access to. As such, a community approach to handling data about OA books brings potential for achieving economies of scale and for taking advantage of network effects, both of which help to address resource challenges facing individual stakeholders and to allow comparison and benchmarking to the benefit all stakeholders in the system. At the same time a widest possible community approach brings open access authors the benefit of better knowledge of how and where their work is visible and in use, thus raising the perceived value and acceptance of open access in general.

Successful collaboration will require thoughtful engagement with issues of trust, 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. To lay the groundwork for this, we develop a case for the need for action and a description of the landscape, and we propose a ‘community data trust’ as a way forward for the monograph community.

Introduction

Opportunity

Comprehensive access to usage data for OA scholarly 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 become visible and are being used. If well managed at a community level, OA book usage data could provide:

- Insight into the relative performance of individual books and collections
- Insights on dissemination strategies and intermediary platforms or networks
- Benchmarking and tracking of changes in patterns of use over time
- Information about subject-specific patterns of use for OA books
- The ability to map the communities engaging with OA books
- Justification for greater investment in OA book discovery
- New tools for evaluating and communicating the value and performance of OA monograph publishing
- Evidence for policy-makers and funders to inform policies on OA monographs

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* is increasingly important.

Capturing and analysing this usage data presents a significant challenge. Data relating to OA books is generated at many different points within the digital landscape, and no single player has access to a complete picture of how OA books are being discovered and used. In order to provide useful information to stakeholders in monograph publishing, and to ensure the privacy and security of users, usage data must be gathered, cleaned, analysed, and presented with skill and care. Even the largest players in the monograph space can lack 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 stakeholders share, the more each benefits. 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; all that is required is selecting a set of systems that support the consensus goals of a community of stakeholders. The hard problem is developing and agreeing on those consensus goals.

The monograph ecosystem makes a solution tractable and desirable

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 article publishing, there are no dominant players with interests significantly different to 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.

Diversity

There are a wide range of players engaged in monograph publishing. Unlike journal publishing, which is dominated by a small number of presses and infrastructure providers, monograph publishing is both distributed and highly diverse. The number of OA monograph presses is growing, whereas the number of journal publishers is dropping as a result of market consolidation (Neylon et al. 2018). This diversity means that, while the interests of individual stakeholders varies, OA monograph publishing is not dominated by a small group of players with radically different interests to the rest. For instance, in journal article publishing, large publishers possess exclusive access to large data sets relating to a broad range of topics and classes of journal. Given the breadth and depth of their own data, they might not see a benefit in sharing with their competitors. In the case of monograph publishing on the other hand, individual players are unlikely to have exclusive access to comprehensive data relating to the use of their titles. As a result, the benefits of collaborative approaches for individual stakeholders are more apparent.

Institutional connection

Teaching and research institutions are playing an increasing role in scholarly publishing, often through their libraries, and this is particularly prevalent for OA monographs. The 2018 Library Publishing Directory lists 156 library-based publishing programmes, located on four continents. These programs produced over 2,000 monographs from 2015 to 2018 (Library Publishing Coalition 2018). In addition, over 30% of university presses in North America now report to libraries, which are increasingly invested in their success. The engagement of libraries, with their broader role in supporting research communications, brings a new perspective to scholarly publishing, in particular a mission focused on maximizing dissemination of and access to published works. This is just one reason that usage data has become more important as a metric for measuring the success of OA publishing.

Community infrastructure

Not-for-profit and community-funded platforms are playing a key role in OA monograph discovery and distribution. Examples include hosting and preservation services such as the OAPEN Digital Library, Project MUSE, HathiTrust, and JSTOR Books, as well as discovery services such as the Directory of Open Access Books and the Digital Public Library of America. These content collections include OA books published by both commercial and not-for-profit publishers.

Multiple distribution pathways

In contrast to journals, which are most often hosted on a publisher's website, OA monographs tend to be hosted and distributed via third-party aggregation and distribution platforms, with content often distributed through multiple channels. This means that data on usage needs to be collected from multiple platforms and systems to provide a comprehensive view of a publication's reach.

Shared interests make a community solution viable and desirable

Monograph publishers distributing content through external platforms have a direct interest in obtaining comparable and usable usage data from these platforms, particularly for OA titles. In turn, platform providers have an interest in demonstrating to publishers that they provide a good service in terms of availability, accessibility, discovery, and use. All players have an interest in obtaining granular data that informs them of who is engaging with specific titles and where they come from.

The presence of a large number and diversity of important players in monograph publishing makes it unlikely that a single actor will unilaterally provide a global solution and highly unlikely that any such unilateral action would be adopted by a critical mass of players. Successful coordination requires an approach appropriate to a collective action problem and will involve building a framework that is sufficiently trustworthy to achieve engagement by a critical mass of players. The final section of this report details a proposal for coordination, with further background information found in Appendix A.

Challenges

In identifying how the community of stakeholders might proceed, it is helpful to lay out the challenges that need to be addressed. These challenges can be categorized using the PEST framework (Aguilar 1967; "PEST Analysis" 2018). The PEST framework, commonly used to support an environmental scan or market research, forms the basis of the *Open Scholarship Framework* proposed by the European open scholarship infrastructure consortium Knowledge Exchange (2017).

The framework categorizes challenges according to the 'arena' they take place in (political/regulatory, economic, socio-cultural or technological) as well as by the level of granularity at which they operate (macro, meso, or micro). While there are challenges for OA monograph publishing that arise across the framework, we will see that the critical issues in this domain lie particularly in the social arena at the meso (i.e., community and organizational) level. For that reason we tackle the categories slightly out of order, handling

the social challenges—how to organize a community to address challenges in the other arenas—last.

Political challenges

Within the PEST framework the ‘political’ arena refers primarily to regulatory and legal issues that affect the strategic environment.

Privacy of user data

Ensuring that the privacy of individual users is maintained is a real and legitimate concern for all stakeholders in the OA monograph space, from scholars and their institutions to publishers, content aggregators and platforms, and research funders. Clear mechanisms for protecting user privacy, demonstration of technical and regulatory competence, and appropriate sharing of sensitive data are necessary both to create the trust necessary for collaboration and to ensure compliance with institutional data privacy policies and with legal regulation.

Competition and antitrust law

Our assessment is that competition and antitrust issues are not an immediate problem for a community approach to managing usage data, but any proposed solution or organization could add value by monitoring antitrust risks and identifying when competition and collusion issues might arise so as to aid the community in avoiding them.

Economic challenges

Cost of platform provision

The primary economic challenge in developing an interoperable ecosystem of online book usage data is the cost of building, operating, and continuing to develop a shared technical system for its provision. A solution has not been provided by the market because the stakeholders are too diverse and the available resources insufficient. With many small stakeholders in the market, some form of coordination is required to build such an infrastructure if it is to serve the needs of the monograph community. If market provision had been successful, it is likely that these costs would have been borne by a single player that unilaterally sought a market opportunity. It is noteworthy that this has not happened even in the relatively well-resourced journal space.

Coordination costs

Even if the resources to provide such a platform were found, there are additional ongoing costs related to the coordination of stakeholder actions. The social coordination challenges that arise from the diversity of players in the monograph community are discussed in more detail below. At this point we simply note that there are also direct financial costs to managing coordination amongst players, largely in the time of community members dedicated to establishing and managing standards and rules regarding usage data.

Technological challenges

Format diversity

OA book publishing is a site of continued experimentation and innovation. Publishers are actively exploring the possibilities of digital technology and OA distribution and are engaging with new approaches to supporting the costs of publishing OA books.

OA monographs are often made available in multiple ways: HTML, PDF, EPUB, and Mobi files are made available alongside print, and books are made available for viewing or download either in their entirety or as individual chapters. Each of these delivery methods and formats can be associated with different kinds of usage and therefore different forms of usage data.

Metadata consistency

Ensuring consistency of metadata remains a key challenge for OA monographs (Neylon et al. 2018). Incomplete metadata attached to books and individual chapters - for example, missing key words, a lack of consistency in the application of ISBNs and DOIs, and variations in spelling at a title level - create challenges for both capturing accurate usage data across a distributed landscape and for processing and analysing data at scale.

Data comparability

Usage data collected by diverse organizations applying different methodologies for identifying, capturing, and cleaning data is a consequence of the decentralized nature of OA monograph distribution and use (Neylon et al. 2018). The different approaches to capturing and handling usage data are compounded by diversity in the formats that data relates to: usage data may relate to an entire book, a single chapter, individual page views, etc. Careful handling of heterogeneous data is vital to supporting accurate interpretation and meaningful analysis, as well as to maintaining trust in claims about OA.

Social challenges

Defining usage data

A non-trivial issue in the development of a coordinated approach to the management of usage data is developing a community consensus on what type of usage is in scope. Would any system seek to capture download or viewing figures only? Would it also incorporate evidence of use arising from social media? What about creation and viewing of web annotations? What is required is a process by which specific forms of data are examined and considered for inclusion in any community-based approach to managing usage data.

Selecting data schema

Beyond defining what counts as usage data, another challenge is to define a shared schema for describing and collecting data from a range of sources. Scholarly publishing already relies on a number of differing schemas (ONIX vs MARC for bibliographic metadata, BISAC & BIC vs LCSH for subject headings). The difference between schemas and their

implementation details relate to the differing concerns of different stakeholders e.g. libraries (MARC and LCSH) as opposed to those of publishers or supply chain managers (ONIX and BISAC). Developing something that works for the diverse set of involved stakeholders requires either a resolution between schemas (a technical challenge) or a selection of one that will serve the needs of stakeholders (a social challenge).

Publisher diversity

An important feature of the OA monograph publishing space is its diversity and lack of commercial concentration (Prainsack et al. 2013). In contrast to the natural and medical sciences, which are now dominated by a handful of large publishing houses (Larivière, Haustein, and Mongeon 2015), humanities and social science (HSS) publishing is characterized by the key role that small and medium-sized presses play in supporting research and communication for HSS communities.

Organizations engaged in OA monograph publishing include large commercial players, university presses, library-based publishers, and independent scholar-led presses, all of which can have workflows that are either focused on delivering print books or intended for digital-first production. A consequence of this diversity is variation in the level of resourcing and in-house data-handling capacity available within individual publishers, as well as a lack of uniformity in workflows, metadata, and usage data across the monograph space.

Platform diversity

OA books can, by definition, be read freely and in most cases shared and re-distributed. In contrast to closed publications, access to them does not involve paywalls. Given use of open licenses, access often occurs on a website that is not controlled by the publisher. As a result, it is not possible to capture all of the usage data relating to individual OA monographs. However, there is evidence to suggest that discovery pathways for OA monographs are driving a significant proportion of use towards known platforms rather than to other collections of openly licensed content (Montgomery et al. 2018). We expect this trend to continue in the future.

Commercial concerns and risk management

Commercial and reputational concerns with sharing data about the uses of OA monographs are certainly real concerns for many stakeholders in scholarly publishing. Managing those concerns well is vital to enabling collaboration and realizing network effects. The risk that usage data will be misinterpreted—or that inappropriate comparisons will be made between titles or publishers, with consequences for the funding of researchers or presses—must be taken seriously. A thoughtful governance framework and effective mechanisms for engaging with stakeholders will be key to managing this risk.

Coordination

We noted above the financial costs of coordination. The social challenge of coordination is to provide a trusted organization or framework that provides a reliable and trusted means of addressing all the social issues discussed above. This means having governance that addresses the issues of stakeholder diversity, that provides trusted mechanisms for

managing and holding sensitive and confidential data, and that can manage the community discussions required to define scope and technical decision making. Addressing the coordination challenge is at the heart of tackling all the other challenges described above.

A landscape survey of monograph and scholarly usage data

In this section we examine some of the existing projects and organizations operating in the OA publishing space that are active in engaging with the challenges identified above. To do this we consider them in broad categories reflecting a simple life cycle for the collection of usage data for OA monographs: the services that host digital monographs, systems used to collect and process raw data for individual books, and organizations and services that support combining, comparing, and analysing usage information across books, across collections and over time periods.

A more detailed environmental scan of initiatives and developments relating to usage data for OA books is provided in Appendix B.

Hosting platforms

The platforms on which scholarly monographs are hosted fall into three broad categories. Some have been built specifically for books (such as the OAPEN Library, publisher platforms such as the one created by Open Book Publishers, Google Books, and Amazon) or built with a strong interest in books (such as OpenEdition and Project MUSE). Other platforms were built initially for journal articles with books being added later (JSTOR, Érudit). A related case is where books are hosted in institutional repositories, either by the publisher (e.g., UCL Press) or by authors self-archiving their work (“green OA”).

A limitation of these hosting platforms for the purposes envisioned in this report is that they have not always been built with the collection of usage data in mind. That means that collection of usage data is deferred to other systems, discussed below, which are often not designed with books in mind.

Usage data collection and processing

Systems for usage data collection fall into two main categories. First there are those built for general web-server infrastructure. These include Google Analytics, a widely used system for gathering usage and demographic data, and AWStats, an open-source alternative to Google Analytics that provides broadly similar, but generally less sophisticated, functionality. These systems can be useful for studying online usage of book platforms but are certainly not designed with this purpose in mind and raise a number of questions about user privacy.

The second class of systems was designed for scholarly content in particular. COUNTER provides a set of standards and requirements for processing server logs that was designed for libraries subscribing to electronic resources. This was originally focused on reporting subscription usage of journal articles but the scope has expanded over time; with versions four and five of the COUNTER Code of Practice there has been an effort to consider books

and OA content. Another example is the usage data collected by Open Book Publishers and systems being developed through the HIRMEOS project (see Appendix B).

The tight coupling between the mechanism of data collection and its initial processing tends to lead to an assumption that a choice for one precludes the other. For instance, COUNTER and AWStats are based on server logs, whereas Google Analytics uses on-page scripts. However, this does not mean that an approach similar to COUNTER processing to enable comparison of usage across platforms would not be possible for data derived from on-page script based systems.

Analytics tools provide a range of advantages and sophisticated tooling but often have opaque functionality, with design choices made for the consumer web and not the scholarly community. Server-log processing, on the other hand, is complex but flexible. That flexibility however comes with a requirement for standardization if data is to be compared across sources. This raises problems that are economic (standardization and certification have costs), social (coordination is required to agree standards and this difficult across differing user communities) and political (what level of regulation is appropriate) beyond the technical issues of systems and implementation.

Collecting and combining data for analysis

There are a range of organizations and projects collecting data from multiple sources. Of these the most significant is IRUS, a growing consortium working to support the collection of usage data from institutional repositories, which provide hosting for only a portion of published OA monographs. Other efforts such as KU Analytics are generally focused on delivering a service to a specific group of members. HIRMEOS is amongst the most general technical efforts looking to aggregate usage and other data from a set of project members.

To our knowledge there is no coherent effort to generate a combined set of usage data across the scholarly publishing space, either for books or for journals. This is in part the result of the challenges described above as well as a symptom of the challenges of coordination more generally.

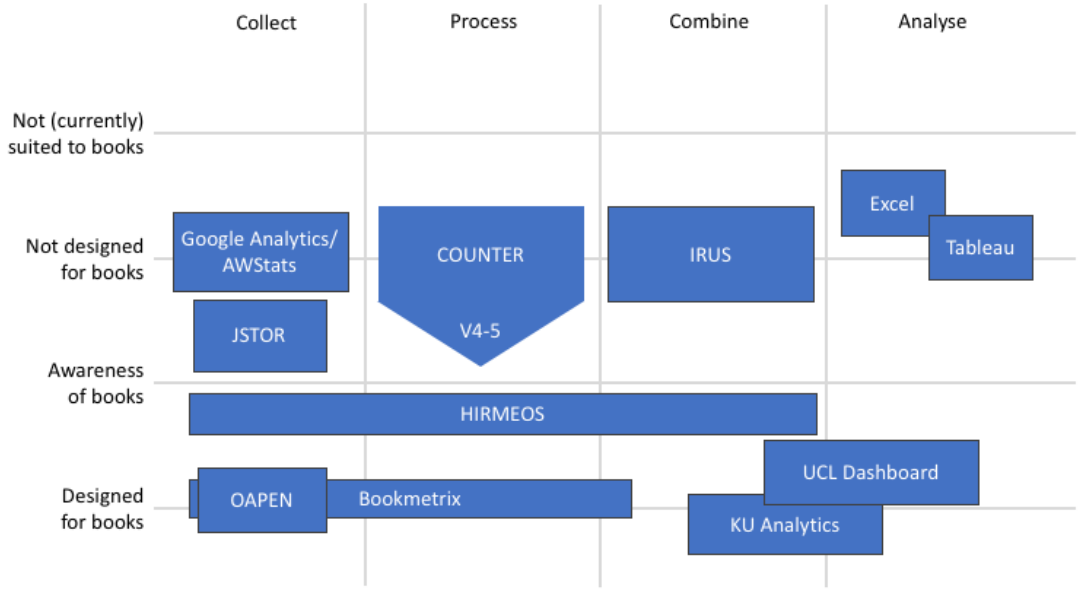

Figure 1. Various tools and projects and the relative usefulness for books at different stages of the usage analytics pipeline. The positions are intended to be indicative, not precise.

What is missing? Coordination.

Important investments have been made over the past decade in the development of platforms, protocols, and networks capable of supporting open scholarly communication. Although the focus of the library, publishing, and research funding communities was initially on infrastructure for journals, increasing attention is now being paid to the needs of OA monograph communities as well.

Tools and services with the potential to support OA monograph stakeholders in making the most of usage data already exist. However, existing projects reflect the focus of specific stakeholder communities, as well as the limitations of the national and regional funding sources supporting them. As a result, technical solutions relevant to OA monograph data are spread across a number of different projects. IRUS-UK, for example, has successfully created economies of scale through the centralized collection and processing of usage data for institutional repositories and is partnering to expand this approach beyond the UK. However, all of the IRUS projects (including those operating in the US and Australia) currently focus on capturing and processing data associated with participating institutional repositories. Capturing usage of content across multiple platforms is not yet something that IRUS is able to do. As an additional example, JSTOR provides publishers with information about the use of both open and closed books that are hosted on the JSTOR platform and has worked with publishers to explore ways to ensure that usage data that it provides to monograph publishers is useful and informative (Montgomery et al. 2017). But as with IRUS, JSTOR's efforts do not address the challenge of understanding the uses of books hosted in diverse locations.

Perhaps the greatest progress towards the development of federated, community-driven approaches to OA monograph infrastructure is being made as a result of European Commission funding. The OPERAS and HIRMEOS projects are both working to ensure that

OA monographs are effectively integrated into the European Open Science Cloud—not through the development of new technologies, but through conscious efforts to build and support OA publishing communities (OPERAS) and the coordinated application of existing technologies to the challenges of OA scholarly books (HIRMEOS). However, both OPERAS and HIRMEOS are funded to develop *European* infrastructure to support open scholarly communication. Addressing the needs of communities outside Europe is beyond the current scope of these projects. Furthermore, the technical infrastructure project HIRMEOS has not yet moved beyond a pilot phase, and the tools being developed by HIRMEOS to capture, manage, and display usage data are still a work in progress.

What is missing is a truly international network for the coordination, management, and reporting of OA book usage data. This network would include both human and technical capacities. A key feature of this network would be experts with the capacity to help members navigate the technical, ethical, and legal challenges associated with ensuring the benefits of OA book usage data.

What do we propose?

We posit that the strategic goals of many stakeholders in OA monograph publishing can be met through a collective approach to data analytics that engages with the international nature of publishing and scholarship. Specifically, we recommend a knowledge commons model, under which participants agree to share standardized usage data with an independent organization—a data trust. The data trust would compile members' usage data and provide regular analytics on behalf of members. This model is a pragmatic approach to providing advanced analytics to stakeholders in the sector: it engages the widest possible range of stakeholders, not by requiring substantial technical expertise from all players but by providing high-quality networks. It generates a knowledge commons that benefits from positive network effects without requiring complete transparency and data sharing from participants. Most importantly, by placing the management of this data in the community rather than third party hands, it addresses concerns over autonomy and security of data.

A data trust for open access monographs

This report makes the case that creating a knowledge commons for sharing OA monograph usage data represents an important opportunity to support innovation and growth in OA monograph publishing. As publishers continue to experiment with business models and dissemination practices, industry-level longitudinal, geographical, subject, and format-specific data and analytics can inform strategic decisions on publishing and dissemination. Furthermore, by tracking and analysing global readership of OA monographs, current understandings of the impacts of monograph publishing can be improved. Sharing industry data will allow OA monograph publishers to provide better metrics to authors and will improve awareness of both the social and private benefits of OA monograph publishing more broadly.

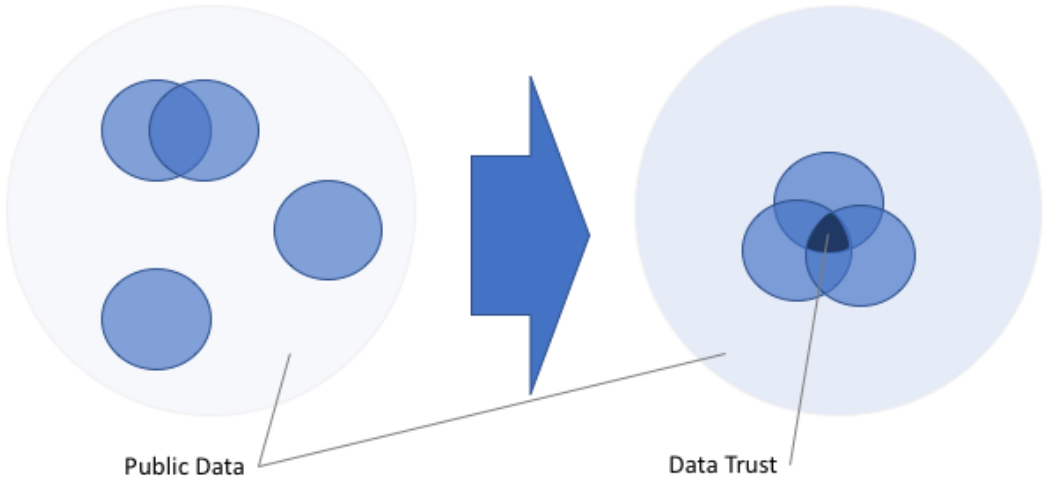

Figure 2. The creation of a data trust. Currently there is some specific sharing of data within projects or groups and a small amount of public data. By creating a data trust, it is envisioned that diverse actors will share a specific subset of data within a common (but not public) data store.

We propose that OA monograph data sharing should occur through a 'data trust' model. Under such a model, members of the trust agree to make OA monograph usage data available to an independent intermediary. The intermediary—the data trust—compiles and analyses data on behalf of trust members and carries out benchmarking. Members can access the analysed data through a user-specific dashboard. The data trust may also provide an API through which members can extract anonymized, structured data in order to conduct their own analysis. The data trust operates as a not-for-profit organization, functioning only to facilitate industry data sharing for the benefit of members. Funding will be required to establish the data trust and to set up the technical and governance structures. On-going operations may be partially or fully funded by member contributions. The core principles and contractual terms of the Trust for OA Monographs are provided below.

Additional detail relating to the broader concepts and considerations that inform these principles are provided in Appendix A.

Principles of a data trust for open access monographs

The Data Trust for Open Access Monographs will be guided by the following core principles: SECURITY, USEFULNESS and FAIRNESS.

SECURE. The Trust will ensure the security of information that is proprietary and will respect the privacy interests of readers by sharing aggregated, unidentified data only. Data will be stored and secured according to industry best practices, and the Trust will be transparent with members regarding data storage, security, and management policies and practices.

The Trust shall not operate in any manner which requires members to disclose, to the Trust or a third-party, their passwords or similar credentials used for collecting data (e.g., for accessing Google Analytics or similar services); in all cases the Trust must, without prejudice to the need for audits, provide members with the means to collect data relating to their OA monographs using facilities chosen and operated by the members themselves or by third parties they have chosen, and to control the disclosure of any data thereby obtained.

USEFUL. The Trust will create value for members by collecting, analysing, and sharing industry data to support strategic decision-making and innovation in OA monograph publishing. The Trust will seek to provide longitudinal, geographical, subject-area, and access information (for example, the type of platform through which the monograph was accessed and the number of chapters accessed). Data will be collected and reported at regular intervals—for example, monthly or quarterly—as determined by Trust members.

To ensure the collection of comparable data, the Trust will establish and facilitate the adoption of standardized data collection and cleaning protocols, potentially through the adoption or extension of existing technical systems. By standardizing data collection, the Trust will create new useful data for members and enhance the utility of existing data sources.

FAIR. All trust members will supply the same type of data and will have equal access to aggregated data and analysis. Limitations on data access will ensure the Trust will not compromise members' commercial interests and advantages. Data access will occur through a user-specific dashboard and through an authenticated, permissions-based API. Each member will have access to book-specific data for books represented by their organization. Only the Trust will have access to book-specific data from all members.

Contractual terms

Roles

- "Trust Members" are typically publishers, libraries, funders, or other organizations interesting in learning about the usage of OA monographs.
- "Trust Data Providers" are typically organizations that aggregate OA monographs and make them available for download (e.g. JSTOR, MUSE, Amazon).
- "Trust Administrators" are the people who work for the new non-profit trust organization and have the job of aggregating/cleaning data and making it available to the Trust Members.

Supply of data

- Data will be supplied to the Trust by Trust Data Providers with the express consent of Trust members.
- All members will agree to supply the same data under a common set of protocols that will include details of collection, any cleaning processes, and reporting.

Access to data

- Only the Trust will have access to all data.
- Trust members will have equal access to aggregated data and analysis.
- Trust members will have access to title level data for titles owned or represented by their organization only.
- Access to aggregated and individualized data will be provided through a closed platform—a data portal with a user-specific dashboard—or through an authenticated, permissions-based API.

- The Trust will review and, with the permission of Trust members, grant requests for limited access to the Trust's data by researchers.

Use of data

- The Trust will have the right to collect, aggregate, and report data to Trust members.
- The Trust will provide an API through which members can extract unidentified, structured data in order to conduct their own data analysis of aggregated data.
- Members will agree to use the data for internal decision-making and analysis only.
- No member will have the right to license, sell, or provide for commercial purposes access to aggregated data or reports.
- The Trust will not have the right to license, sell or provide for commercial purposes access to the Trust's data.
- Members will retain ownership of their organization's data.

Membership and termination

- Uniform legal agreements will be required for Trust membership.
- With notice, members will be able to leave the Trust and opt to have their data removed from both aggregate and individual data portal access.
- As the Trust grows, any new members must meet the eligibility criteria developed in full consultation with Trust members.

Governance

- The Trust will operate as a not-for-profit organization (which may or may not form a new legal entity).
- The Trust will adhere to competition laws and data regulations in jurisdictions in which it operates (the jurisdiction in which the Trust operates and stores data will be determined by founding members).
- Policy-making and policy modification by the Trust will occur with the participation of Trust members.
- The Trust will implement a policy for auditing members' data collection, cleaning, and reporting practices in order to ensure the quality of the data and adherence to Trust policies.
- The Trust will implement a policy permitting members to audit the Trust to ensure adherence to Trust policies.
- The Trust will provide a low-cost dispute-resolution mechanism.
- The Trust will implement a graduated sanctions policy applied in response to Trust policy violations.
- The Trust will implement a policy specifying liability for data loss or data breaches.

Appendix A

A data trust for industry data sharing

A data trust is a knowledge commons model for data sharing which can empower and create value for its members. The access provided by the trust does not depend upon the opaque and internal decisions of a firm operating for profit in a competitive market. Under a data trust arrangement, participants that share data create and control the resource, governing access according to mutually agreed principles. The data trust creates a new resource for under-resourced market participants—it is an economic resource that is shared for mutual benefit.

Building a data trust for industry data sharing is a collective action challenge. It requires encouraging a diverse range of actors to share commercially sensitive information about their assets and operations. It also requires the sharing of information in a form that is standardized and comparable so that the analysis of that information is accurate and useful.

A successful data trust must be mutually beneficial for all participants. Enabling collaboration at this scale requires a carefully calibrated system of incentives and investment in providing value back to participants. If carefully designed to meet the needs of participants, a data trust is an institutional model with the potential to facilitate sustainable industry data sharing and to improve opportunities for analytics-led innovation and growth.

How do you design a sustainable industry data sharing trust that will provide members with opportunities for analytics-led innovation and growth?

Knowledge commons research shows that governance design is critical to the sustainability of a commons. This is because a commons requires cooperation by a variety of actors, many of whom will be competitors (Frischmann, Madison, and Strandburg 2014, p.477). Sustainable cooperation among competitors requires trust, a sense of community, and reciprocity (Hess and Ostrom 2007, pp.43-44). The data trust must be designed and governed in a way that engenders these three properties among its members.

In her analysis of voluntary communal management of common pool resources, Elinor Ostrom identified that sustainable commons share particular design features that provide certainty for participants about the use and governance of the resource (Ostrom 1991). Certainty about how a resource is used and governed engenders trust and supports sustainable collective action (Cox, Arnold, and Villamayor Tomás 2010, p. 2). In the context of data sharing, this principle is illustrated by the results of a global 2017 survey in which '25% of respondents from health care industries said they are likely to share data with competitors, compared with 19% of respondents from other industries' (Ransbotham and Kiron 2017, p.12). The authors of the survey concluded that in the health care sector, established patient privacy and data regulations encourage industry data sharing. Structured governance is important for engendering trust and supporting resource sharing.

According to Ostrom's analysis, a sustainable commons must have clear boundaries. This requires defining the resource, who may use it, and who is excluded from using it. Defining who is included and who is excluded from the commons provides participants with assurance that the benefits derived from their contribution will not be unfairly enjoyed by those that have not made a similar contribution (Ostrom 1991, p. 91). Once boundaries are defined, establishing a sustainable commons requires clear rules for using and contributing to the commons that are tailored to fit the specific resource. Rule-making and rule-

modification should be devised with the participation of users of the commons. The use of the commons should be actively monitored in order to ensure compliance with commons rules and boundaries, and there should be low-cost dispute resolution mechanisms and graduated sanctions applied in response to boundary or rule violations (Ostrom 1991, Cox *et al.* 2010, p. 10). A sustainable commons also requires that the self-governance of the commons is respected by external authorities. Finally, for complex common resources, governance of the commons should occur on multiple levels—for example local, regional and national—in order to successfully manage the different and interconnected components of the resource system (Ostrom 1991, p. 101).

Ultimately, both the governance and technological layers of the data trust must be tailored to the specific requirements of the data resource and its members.

European union guidance on business-to-business data sharing

In 2018, the European Union released a report providing initial guidance on private sector data sharing. This guidance identifies important issues that require consideration when designing an organization that shares industry data and has informed the principles and contractual terms proposed in this document.

The report provided the following five principles for business-to-business data sharing arrangements (European Commission 2018):

> a) *Transparency*: The relevant contractual agreements should identify in a transparent and understandable manner (i) the persons or entities that will have access to the data that the product or service generates, the type of such data, and at which level of detail; and (ii) the purposes for using such data.
>
> b) *Shared value creation:* The relevant contractual agreements should recognize that, where data is generated as a by-product of using a product or service, several parties have contributed to creating the data.
>
> c) *Respect for each other's commercial interests:* The relevant contractual agreements should address the need to protect both the commercial interests and secrets of data holders and data users.
>
> d) *Ensure undistorted competition:* The relevant contractual agreements should address the need to ensure undistorted competition when exchanging commercially sensitive data.
>
> e) *Minimized data lock-in:* Companies offering a product or service that generates data as a by-product should allow and enable data portability as much as possible. They should also consider, where possible and in line with the characteristics of the market they operate on, offering the same product or service without or with only limited data transfers alongside products or services that include such data transfers.

The report also specified issues for consideration when preparing or negotiating data sharing contracts (European Commission 2018):

1. What data shall be made available?
2. Who can access and (re-)use the data in question?
3. What can the (re-)user do with the data?
4. Define the technical means for the data access and/or exchange
5. What data do I need to protect and how do I protect it?

6. Include rules on liability provisions for supply of erroneous data, disruptions in the data transmission, low quality interpretative work, if shared with datasets, or for destruction/loss or alteration of data (if it is unlawful or accidental) that may potentially cause damages.
7. Define rights of both parties to perform audits on the respect of the mutual obligations.
8. What is the intended duration of the contract? What rights to terminate the contract? What notice to be given to your partners?
9. Agree on applicable law and dispute settlement mechanisms.

Industry data sharing

Current industry data sharing practices suggest data sharing is emerging as an important driver of innovation. A 2017 data and analytics study by MIT found that highly innovative organizations share data internally and externally at higher rates than other organizations (Ransbotham and Kiron 2017, p. 10). In 2015, German car companies BMW, Daimler, and Volkswagen collectively purchased a business that owns a digital mapping technology that facilitates traffic jam and road condition video sharing between drivers (Ransbotham and Kiron 2017, p. 11). Through the jointly owned business, the car companies share data collected from sensors in their vehicles and are developing services that deliver real-time alerts to drivers. As Sam Ransbotham and David Kiron (2017, p.11) note, the customer base and data that each company brought to the partnership has allowed the companies to collectively create a more valuable service for their customers than if they had each attempted the project individually.

Industry data sharing occurs commonly in the field of medical research. Within the pharmaceutical industry, for example, clinical trial data sharing is recognized as important for supporting scientific discourse, innovation and discoveries (Francer, 2015). In this field, a common model for data sharing is via an independent intermediary. For example, Johnson & Johnson works with the the Yale School of Medicine's Open Data Access (YODA) project (Yale University Open Data Access Project 2018). YODA is responsible for deciding when to share data that is requested by third parties. In this industry setting, organizations seeking to share data must carefully navigate issues relating to patient privacy and commercial incentives. The high rate of data sharing in this sector highlights that data obtained through large financial investments, with a potentially high commercial value and subject to significant regulatory requirements, can be shared successfully.

Appendix B

This appendix provides an overview of initiatives that are relevant to the development of a trusted framework for the coordination of OA usage data. This overview is organized according to the PEST framework, whereas in the main text initiatives are organized according to the stage of usage data production that they relate to. This environmental scan covers more initiatives than in the main text, but does not claim to be comprehensive.

Political interventions

Political interventions relevant to coordinated action on OA usage data frequently come from outside the OA monograph community. They include those that impact OA monograph production dissemination and use such as government policy towards OA and other regulations relating to, or informing, the distribution of funding. They may also include measurement efforts with implications for future policy—for example, the Open Science Monitor will inform policy analysis in the European Union for the open science agenda. Another example is proposed frameworks to define rules of engagement such as the Principles for Open Scholarly Infrastructures (Bilder, Lin, and Neylon 2015).

Other regulatory issues are not specific to publishing or monographs but have significant implications for collecting and sharing data. These include privacy regulations, most recently the EU General Data Protection Regulation, as well as regulations that enable state agencies to obtain data such as the Patriot Act in the USA, and those governing other nations' security agencies. Privacy regulations create obligations to safeguard data, while security regulations can create an ethical duty to refrain from collecting or destroy it. Other general regulatory frameworks include the European Digital Single Market draft guidelines for private sector data sharing (European Commission 2018) and antitrust/competition regulation in various jurisdictions.

Some of these are discussed below.

Regulation within (or of) the scholarly community

OA mandates and their evaluation

Much of this discussion about how best to capture and manage usage data for OA books is occurring in the context of significant, but as yet not fully defined, policy initiatives relating to OA more broadly. In the UK, the four Higher Education Funding Bodies have signalled their intention to move towards an open access requirement for monographs submitted to the post-2021 Research Excellence Framework (REF).. a. Whilst setting the direction of future policy for the REF after next, the four funding bodies recognise that there will be legitimate reasons why some monographs cannot be OA. The new initiative by European funders, 'Plan S', includes reference to OA books, although it notes that achieving 100% OA will not be feasible on the same timeframe as for journal articles ("Plan S: Accelerating the Transition to Full and Immediate Open Access to Scientific Publications" 2018). Within the context of the shift to OA and the complex changes in financial flows that this implies, an ability of stakeholders to demonstrate the value created and return on investment for OA monograph publishing will be crucial. In addition, it is likely that future government policy and funding requirements will be tied to the provision of high-quality bibliographic metadata into open systems.

Principles, manifestos and proposed governance requirements

There are a wide range of scholar-led statements, manifestos, and other documents that might be considered relevant. Amongst these are the Principles for Open Scholarly Infrastructures (Bilder, Lin, and Neylon 2015), which set out principles for the governance of systems such as the Data Trust for Open Access Monographs presented in this document.

The Data Trust proposal is broadly in line with the Principles for Open Scholarly Infrastructures.

General regulations impinging on monographs

There are a range of general regulations and political initiatives that are broadly relevant to monographs because they either relate to digital and technical platforms or have implications for the collection of data.

The European Digital Single Market is a significant policy initiative that seeks to create greater interoperability across the European Single Market. These policy changes touch on copyright, interoperability, and consistency of charging for digital services. Another area of regulation more generally relating to privacy are the ethical issues raised by the increasing ability of security services to obtain private data. For many players it is preferable not to collect or keep personally identifiable data at all in the current environment.

As we noted in the main document there does not seem to be any specific risk of infringing antitrust and competition regulations. Nonetheless, maintaining a watching brief, potentially in collaboration with other relevant organizations and communities, may be of value.

Economic interventions

Economic interventions have mainly taken the form of funding initiatives and experiments seeking to support sustainable models for scholarly monograph presses and/or to encourage a transition towards open access. The earliest amongst these was the Wellcome Trust's commitment to provide funding for book processing charges for OA books arising from funded research. Funders have also directly supported platform development including OAPEN, DOAB, and other relevant providers.

Non-funder initiatives have had some success, with examples such as crowdfunding (Unglue.it) and the donor model of small presses such as Punctum Books. Other presses have directly developed alternate models based on a freemium approach (Open Book Publishers, OpenEdition) or direct subsidy (UCL Press). Finally, collective funding models, with Knowledge Unlatched (KU) being the most prominent example, have made significant progress.

KU was established as a not-for-profit company in 2012 with the goal of enabling a more effective monograph ecosystem. It sought to create a sustainable route to OA for book-length publications by helping libraries to share the costs of OA for professionally published scholarly books (Montgomery 2015). KU provides a coordination service to libraries and publishers: libraries pay into a fund, and KU uses the fund to pay publishers an agreed fee for OA monographs selected by the library community. In 2016, after 3 successful pilot rounds, KU operations transitioned from the UK-based not-for-profit into a German-based limited liability company (GmbH), a subsidiary of the privately-owned company Fullstopp ("Knowledge Unlatched – Scaling Up In 2016" 2016; Knöchelmann 2018). In addition to coordinating consortial funding for OA monographs, KU also provides publishers with services to support the dissemination of OA books and has expanded to include journals in its programmes.

The not-for-profit (KU Research) continues to operate as a separate organisation with an independent board ("Frances Pinter Steps down from KU Research | The Bookseller" n.d.).

Social interventions

As we have noted throughout this document, the greatest challenges for the establishment of a data trust for OA books relate to the social issues involved in collective action. The major interventions in this space today are based in Europe. Other groups that are relevant include various press and industry associations as well as conferences that organize relevant discussions. The Library Publishing Coalition provides an example of a group that does both of these things. Here we discuss the largest specific intervention, one focused on community building around books and usage data: the OPERAS network.

OPERAS, OPERAS-D and HIRMEOS

The OPERAS (Open Access in the European Research Area through Scholarly Communication) project was established with funding from the European Commission's Horizon2020 research and innovation programme. Its goal is: *"to coordinate and pool university-led scholarly communication activities in Europe, particularly in the Social Sciences and Humanities, in view of enabling Open Science as the standard practice."* (OPERAS, 2018).

In October 2018 OPERAS partners include 38 members from 15 countries within the European Economic Area. The project is led by a nine-member core group, and coordinated from France by OpenEdition and Huma-Num.

The OPERAS-D ('design') project was launched in January 2017 in order to address the long-term requirements of e-infrastructure and community building of the OPERAS network. The 18-month project involved a design study defining governance models, the identification of key challenges, principles, and opportunities that need to be addressed in the development of technical infrastructure, and a road-map for the long-term development and sustainability of the OPERAS network. A report on the digital visibility of OA monographs was carried out as part of the OPERAS-D project (Neylon et al. 2018). The report explores the extent to which OA monographs published and made available by OPERAS members are visible to the communities that might make use of them, investigating OA monographs from six repositories/publishers located in six different countries. Variability in metadata quality is identified as a key challenge for both the visibility of OA monographs and for the aggregation and analysis of usage data.

OPERAS is also the coordinating organisation for the HIRMEOS (High Integration of Research Monographs in the European Open Science Infrastructure) project. This project, also funded by the European Commission, commenced in 2017 and is a 30-month sister project of OPERAS-D. In the context of social interventions the role of OPERAS and OPERAS-D is enabling the coordination and organisation that underpins the preparation of the proposal and the organisation of work in the HIRMEOS project.

Technological interventions

A wide range of technological interventions have emerged to tackle challenges of capturing, cleaning, and reporting usage data. These include publisher-specific initiatives focused on monographs (like Bookmetrix from Springer Nature), coordinated efforts to tackle the costs of managing usage data arising from specific types of platform (like IRUS), and community-negotiated standards intended to support consistency in reporting of usage data for electronic resources (like COUNTER). These and other projects discussed below engage with an important piece of the usage data puzzle.

COUNTER

The COUNTER project describes itself as providing "the standard that enables the knowledge community to count the use of electronic resources". COUNTER developed as a mechanism for the provision of trusted, comparable data to libraries on usage of subscription journals. This was necessary because of the potential for a non-standardized means of usage counting to allow journal publishers to 'pad' usage numbers. The COUNTER Code of Practice seeks to reduce this problem by providing rigorous and extensive guidelines on the collection and processing of usage data.

COUNTER was developed with a focus on journal articles and on subscription usage. However as a trusted system, the COUNTER Code of Practice has expanded to cover a wide range of electronic resources, including OA content. Version 5 of the Code of Practice, which comes into force in January 2019, has been updated to include more features relevant to books. Importantly, COUNTER v5 allows for license type identification in usage reports.

COUNTER is widely seen as a sector-wide standard for the collection, processing, and reporting on usage data for scholarly content. It has also attracted criticism, primarily related to the cost of obtaining COUNTER certification for usage data, and the technical complexities of conforming with the Code of Practice. The costs of certifying compliance have lead to a situation in which many publishers and platforms report 'COUNTER compatible' data. This usually means providing reports in the format specified by COUNTER but without providing compliance certification. This raises issues for trust and for consistency in the use of this data.

IRUS

IRUS-UK (Institutional Repository Usage Statistics) is a UK-based statistics aggregation service for institutional repositories. The service processes raw usage data from member repositories, consolidating those data into COUNTER-compliant statistics. As a member of COUNTER, IRUS-UK undergoes COUNTER auditing to ensure the quality and consistency of its data cleaning and reporting.

IRUS-UK seeks to provide a nationwide view of repository usage, to provide opportunities for benchmarking, and to act as an intermediary between UK repositories and other agencies. The initiative was established by the UK infrastructure provider Jisc following the conclusion of the Publisher and Institution Repository Usage Statistics project, PIRUS2, in 2011. The IRUS-UK service is funded by Jisc and is being developed by a consortium that involves

Jisc, Cranfield University, and Evidence Base. In October 2017 there were 131 institutional repositories participating in IRUS-UK.

In addition, IRUS-UK provides services to the Netherlands-based OAPEN Digital Library. Pilot implementation of national instances of IRUS in the US and Australia is also underway.

HIRMEOS

As noted above High Integration of Research Monographs in the European Open Science Infrastructure (HIRMEOS) is a project to develop a common layer of added value services on top of existing e-publishing platforms for OA monographs. The 2-year, €1.99 million project began in 2017, with funding provided under the European Commission's Horizon2020 user-driven e-infrastructure development program. Coordinated by CNRS (France), HIRMEOS involves 8 European partners.

HIRMEOS is focused on the development of the technical capacities and services needed to ensure that OA monographs are effectively integrated into the European Open Science Cloud. Additional services, including for the collection, management, and communication of usage data, are being added to five OA monograph publishing platforms as part of the HIRMEOS project. The capacity of platforms to ensure effective identification of OA monographs, interoperability, annotation, and entity fishing are also included in the work plan.

The specific technical goals of the HIRMEOS project is to prototype a set of services capable of integrating OA monographs into the open science knowledge system. This includes the implementation of standard identifiers for documents (DOI), authors (ORCID) and funders (funder registry); development of a centralized certification and badging service to indicate the peer-review status of OA books; crosslinking texts in order to support entity fishing and annotation; and the development of a series of APIs that will allow publishers to gather data on readership, altmetrics, and citations from a variety of websites, as well as a widget for the display of these metrics (HIRMEOS 2017).

During the pilot, the services being developed by HIRMEOS are being applied to the platforms of five established OA monograph players: OpenEdition, OAPEN, University of Gottingen Press, EKT, and Ubiquity Press. This group includes both individual publisher platforms as well as OA monograph aggregation platforms.

A key feature of the approach to usage data being developed within the HIRMEOS project is its federated model. Rather than depending on the centralized aggregation and processing of usage data, HIRMEOS is developing standardized software that can be deployed by individual publishers. The API will allow publishers to collect altmetrics, citations, and readership information from different platforms (e.g., Google Books and OpenEdition). The project will also provide publishers with a Javascript widget, which will allow all of the metrics relating to an individual book to be displayed on a publisher's website, and an analytics dashboard for use by publishers. The project is focused on an implementation that assumes that this usage data will be made public and open. It seems probable that this approach will be too radical for many in the broader OA Monograph community at this stage.

KU Open Analytics

Knowledge Unlatched (KU) has provided usage statistics as part of its offering to participating libraries. This usage data is collected from a range of sources and presented in documentary reports. More recently KU GmbH has developed KU Open Analytics, an online report-generation tool that manages the collation, collection, and filtering of data to generate specific reports for participating publishers and subscribers. At present, the tool incorporates usage data from JSTOR and the OAPEN Library for books made OA via KU. KU hopes to expand the tool to incorporate more data sources in the future.

Bookmetrix

Bookmetrix is a project of Springer Nature to gather and present data on the usage and impact of the books published by its imprints including Springer, Palgrave Macmillan and Apress. Originating as a collaborative project with Altmetric.com to manage social-media data for books, it aims to provide book and chapter-level data on citations, mentions, downloads via SpringerLink, and reviews for monographs. Bookmetrix currently only provides information for books published by Springer Nature imprints.

Platforms

OAPEN

A key European initiative in the OA monograph space is OAPEN. The platform takes its name from the 30-month (2008-2010) EU funded project that supported its establishment: Open Access Publishing in European Networks (OAPEN). The goal of the project was to achieve a sustainable publication model for academic books in the humanities and social sciences and to improve the visibility and usability of high-quality academic research in Europe. At the conclusion of the project, OAPEN continued its operations under a newly formed foundation.

OAPEN deals exclusively with OA, peer-reviewed books and operates two platforms: the OAPEN Library, a central repository containing more than 5,000 fully downloadable OA monographs, and a discovery service for OA books, the Directory of Open Access Books (DOAB).

OAPEN makes COUNTER-compliant usage data available to member publishers and organizations using data prepared by IRUS-UK.

JSTOR OA Books

The JSTOR digital library was established in 1995 in order to help university and college libraries to cope with the storage challenges associated with an ever-increasing volume of scholarly literature. Originally focused on digitized back-issues of scholarly journals, JSTOR now provides access to current journal issues as well as primary sources and scholarly books.

In October 2016 JSTOR launched its OA Books platform; JSTOR now hosts more than 3,000 OA books and lists 52 participating OA publishers in the books section of its website. In contrast to OAPEN, JSTOR hosts both OA and non-OA titles by the same publisher and

provides usage data on both OA and closed books. This makes it possible for individual publishers to compare usage of open and closed titles on the same platform. JSTOR has also been active in collaborating with OA monograph publishers to explore what usage data captured by the JSTOR platform might reveal as well as in exploring the needs of OA monograph presses in relation to reporting of usage data (Montgomery et al. 2017).

MUSE Open

Project MUSE provides access to more than 600 journals published by more than 250 university presses. Project MUSE, which operates as a unit of Johns Hopkins University Press, began incorporating journals from other university presses in 2000. In 2012 electronic books published by members of the University Press Content Consortium (UPCC) were included in the Project MUSE interface.

In 2016 JHU was awarded a $938,000 grant by the Andrew W. Mellon Foundation to support the development of an OA monograph initiative: MUSE Open. MUSE Open, which launched in 2018, provides access to approximately 600 fully OA monographs in PDF format via the Project MUSE interface. The forthcoming release of a further 300 OA titles in HTML5 format has also been announced.

Google Books

Google is relevant both as a platform on which books are hosted, but also as a source of data about those books. The platform provides information on downloads, views, and has also been recently used to obtain data on citations from books. Google is also a source of broader usage information through Google Analytics. This service is widely used by publishers and other stakeholders to track usage, and valuable data can be collected this way. That data needs analysis to be compared with that from other sources (since, for example, it is not directly comparable to COUNTER usage data). The Google Books API has also been used for a range of purposes including tracking citations from books. The Google Books platform can also provide additional usage information.

Projects and organizations providing relevant metadata

Accurate information about the title, licensing status, and availability of OA books is necessary if usage data is to be gathered and managed efficiently and at scale. In this context, centralized sources of information about the licensing status of OA books is important. So, too, are protocols for sharing and harvesting this data. The initiatives below are presented in this context.

DOAB

The Directory of Open Access Books (DOAB) is maintained by the OAPEN Foundation and is based at the National Library of the Netherlands. As of October 2018, DOAB provides a searchable index and links to the full text of 12,968 academic peer-reviewed books and chapters from 282 publishers. DOAB covers multiple subject areas and uses specific criteria for the inclusion of books in its directories. All books listed in DOAB have an OA license, and collaborating publishers are screened for their peer review policies.

DOAB supports the OAI-PMH protocol for metadata harvesting; service providers and libraries are encouraged to harvest DOAB metadata for inclusion in collections and catalogues.

OpenAIRE

OpenAIRE is an initiative funded by the European Commission that promotes and supports open science in Europe through the provision of dedicated expertise and training and through the development of technical infrastructure that harvests research outputs from connected data providers. In October 2018 OpenAIRE has harvested metadata relating to approximately 21 million publications and 606,153 data sets (OpenAIRE Monitor n.d.).

OpenAIRE aims to establish an open and sustainable scholarly communication infrastructure responsible for the overall management, analysis, manipulation, provision, monitoring, and cross-linking of all research outcomes. OpenAire uses OAI-PMH in order to collect data on OA publications and data sets; OpenAIRE compliant repositories are harvested once a week. Repository data is cross-linked with information provided by research funders as well as to data contained in the European Commission's R&D project database Cordis.

Altmetrics data providers (ImpactStory, Plum Analytics, Altmetric.com, and Crossref Event Data)

While not strictly usage data, various indicators and data under the general banner of ‘altmetrics’ and their providers and systems are also relevant and may be of interest to a community gathering around book impact data in the broader sense. Altmetric.com and Plum Analytics are competing providers that collect a range of social media and other attention data including news coverage, policy references, and references from Wikipedia. Altmetric.com is a Digital Science portfolio company that has remained largely independent, although it is increasingly integrated into the set of offering from Digital Science that are wrapped together in the Dimensions product.

Plum Analytics was acquired by Elsevier in 2017 from EBSCO; Elsevier has similarly integrated many of their services aspects into Scopus. Plum does not provide social media metrics for monographs. However, their history with EBSCO means that they include usage data from those sources, as well as usage data from GoodReads, Mendeley, and WorldCat holdings.

ImpactStory is a free offering from a not-for-profit organization that provides a set of attention and other data, largely sourced from Altmetric.com on the works registered in a user’s ORCID profile. Other relevant data sources include the Open Syllabus Project, which tracks incorporation of works in class syllabi, and LENS, which tracks references from patents to scholarly works, although the latter is largely restricted to references to journal articles.

A new resource is Crossref Event Data, which tracks a wide range of ‘events’ relating to objects with DOIs. This includes social media mention and Wikipedia references and has the potential to be both a future resource of open data and potentially a platform through which data can be shared. However, this is currently restricted to objects that have a Crossref DOI, which remains a minority of books.

Bibliographic and status data

Crossref

Crossref is a global provider of Digital Object Identifiers for scholarly publishers. Incorporated as the Publishers Interlinking Resource Association, it is an industry association of publishers acting as a registration agency for the International DOI Federation. As one of the major DOI registration agencies, Crossref is a key player in this space. Crossref also provides Crossref Event Data, a service which tracks ‘events’ that relate to DOIs (see above). This is an important infrastructure that can be leveraged or used by the monograph stakeholder community.

Unpaywall

Unpaywall is a free service and data source provided by the not-for-profit ImpactStory organization. Its main focus is in providing OA status data for objects with DOIs. It is technically feasible to expand this to include objects with ISBNs. The Unpaywall data is obtained by searching for and analysing objects that are held in thousands of repositories worldwide. This means that it additionally provides a mechanism for identifying the location of OA copies of research objects. This mechanism is of value in the design of a usage data aggregator as a means of identifying the various repositories that hold a copy of an OA book. Similar data could also potentially be made available through Crossref, as noted above, although this would presumably be limited to objects with Crossref DOIs.

General Bibliographic Data sources

For completeness we include a brief survey of general bibliographic data sources that are relevant to the indexing and discovery of scholarly books.

The main providers of bibliographic data are Clarivate (through the Web of Science products), Scopus and related Elsevier products, PubMed and Europe PMC, and Microsoft Academic. All of these are proprietary data sources with some limitations on re-use. The Microsoft product provides consistently good coverage of books compared to the others, with some variability across disciplines. It additionally has the advantage of being free to access and re-use (with some costs for usage of the APIs involved). Other important catalogue services include WorldCat, other library catalogues, and DOAB (see above).

Google Scholar is not included in the above list because its terms of service prohibit the systematic gathering of any large scale data from it, which makes it not useful in the design and implementation of broader systems. It is still relevant as a discovery systems, and Google Books is a provider of information on books provided through that platform.

Appendix C - References

Cox, M., Arnold, G., and Tomás, S.V., 2010. A Review of Design Principles for Community-based Natural Resource Management. *Ecology and Society*, 15 (4).

Aguilar, Francis Joseph. 1967. *Scanning the Business Environment*. 1st THUS edition. Macmillan.

Bilder, Geoffrey, Jennifer Lin, and Cameron Neylon. 2015. "Principles for Open Scholarly Infrastructures-V1." https://doi.org/10.6084/m9.figshare.1314859.v1.

Cox, Michael, Gwen Arnold, and Sergio Villamayor Tomás. 2010. "A Review of Design Principles for Community-Based Natural Resource Management." *Ecology and Society* 15 (4). https://doi.org/10.5751/ES-03704-150438.

European Commission. 2018. "Guidance on Private Sector Data Sharing." Digital Single Market. 2018. https://ec.europa.eu/digital-single-market/en/guidance-private-sector-data-sharing.

Eve, Martin Paul. 2018. "On the OA 'Mandate' for Books in the Third REF and the Worry over Trade Books." Blog. Martin Paul Eve. April 4, 2018. https://www.martineve.com/2018/04/02/the-OA-mandate/.

Francer, Jeff. 2015. "Biopharmaceutical Industry Strongly Committed to Responsible Clinical Trial Data Sharing to Improve Public Health." July 23, 2015. https://catalyst.phrma.org/biopharmaceutical-industry-strongly-committed-to-responsible-clinical-trial-data-sharing-to-improve-public-health.

"Frances Pinter Steps down from KU Research | The Bookseller." n.d. Accessed October 26, 2018. https://www.thebookseller.com/news/frances-pinter-steps-down-ku-research-703916.

Frischmann, Brett M., Michael J. Madison, and Katherine J. Strandburg, eds. 2014. *Governing Knowledge Commons*. Oxford ; New York: OUP USA.

Hess, Charlotte, and Elinor Ostrom. 2007. *Understanding Knowledge as a Commons: From Theory to Practice*. Cambridge, Mass: MIT Press.

HIRMEOS. 2017. "HIRMEOS Metrics Services Specification."

Knöchelmann, Marcel. 2018. "Knowledge Unlatched, Failed Transparency, and the Commercialisation of Open Access Book Publishing." *Impact of Social Sciences* (blog). October 3, 2018. http://blogs.lse.ac.uk/impactofsocialsciences/2018/10/03/knowledge-unlatched-failed-transparency-and-the-commercialisation-of-open-access-book-publishing/.

"Knowledge Unlatched – Scaling Up In 2016." 2016. *Fullstopp* (blog). March 11, 2016. http://fullstopp.com/2016/03/ku_scaling-up-2016/.

Larivière, Vincent, Stefanie Haustein, and Philippe Mongeon. 2015. "The Oligopoly of Academic Publishers in the Digital Era." *PLOS ONE* 10 (6): e0127502. https://doi.org/10.1371/journal.pone.0127502.

Library Publishing Coalition. 2018. "Library Publishing Directory 2018." https://librarypublishing.org/wp-content/uploads/2017/03/LPC_LPDirectory2018.pdf.

Montgomery, Lucy, Cameron Neylon, Alkim Ozaygen, and Tama Leaver. 2018. "Getting the Best out of Data for Open Access Monograph Presses: A Case Study of UCL Press." *Learned Publishing* 31 (4): 335–44. https://doi.org/10.1002/leap.1168.

Montgomery, Lucy, Alkim Ozaygen, Frances Pinter, and Neil Saunders. 2017. "Exploring the Uses of Open Access Books via the JSTOR Platform," November. https://hcommons.org/deposits/item/hc:16603/.

Neylon, Cameron, Lucy Montgomery, Alkim Ozaygen, Neil Saunders, and Frances Pinter. 2018. "The Visibility of Open Access Monographs in a European Context: A Report Prepared by Knowledge Unlatched Research." KU Research. https://doi.org/10.17613/M6156F.

Ostrom, Elinor. 1991. *Governing the Commons: The Evolution of Institutions for Collective Action*. Unknown edition. Cambridge ; New York: Cambridge University Press.

"PEST Analysis." 2018. *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=PEST_analysis&oldid=863592174.
"Plan S: Accelerating the Transition to Full and Immediate Open Access to Scientific Publications." 2018. Science Europe. https://www.scienceeurope.org/wp-content/uploads/2018/09/Plan_S.pdf.
Prainsack, B, M Caball, S Louca, and R Pochet. 2013. "Open Access: Practice, Opportunities and Challenges." Vienna COST Symposium Report.
Ransbotham, Sam, and David Kiron. 2017. "Analytics as a Source of Business Innovation." *MIT Sloan Management Review* (blog). 2017. https://sloanreview.mit.edu/projects/analytics-as-a-source-of-business-innovation/.
Yale University Open Data Access Project. 2018. "Yale University Open Data Access Project." 2018. http://yoda.yale.edu/welcome-yoda-project.