REPORT

OA Books Supply Chain Mapping

9 April 2021

Prepared in support of the Developing a Pilot Data Trust for Open Access Ebook Usage Project

Report Authors:
Michael Clarke
Laura Ricci

Funded by the Andrew W. Mellon Foundation

CLARKE & ESPOSITO

For more information about this report, contact:

Michael Clarke, Managing Partner
mclarke@ce-strategy.com
www.ce-strategy.com

Clarke & Esposito is a management consulting firm that focuses on strategic issues related to professional and scholarly publishing and information services. We advise universities and university presses, professional associations, learned societies, foundations, government agencies, research institutes, publishers, and other organizations involved in the creation, dissemination, and curation of professional and scholarly information. Please visit our website for more information: https://www.ce-strategy.com

Table of Contents

Acknowledgements 4

Overview 5

Methodology 7

Key Stakeholders and Intermediaries 8

- Content Funders 8
- Content Creators 9
- Distributors 10
- Content Platforms 11
- Sales Channels 13
- Catalogs and Indices 14
- End User Interfaces 16

Metadata Standards 17

- Product Identification Standards 17
- Distribution Feeds 19
- Usage Reporting 22

OA Metadata Elements 25

Gaps and Opportunities in the OA Books Supply Chain 27

APPENDIX A: OA Books Supply Chain Map 32

APPENDIX B: Index of OA Books Supply Chain Stakeholders 35

APPENDIX C: Stakeholder Semi-Structured Interview Guide 41

APPENDIX D: Pre-Interview Questionnaire 44

Acknowledgements

This report was developed in support of the Developing a Pilot Data Trust for Open Access Ebook Usage project. This two-year project is led by University of North Texas librarian Kevin Hawkins and conducted in partnership with the Educopia Institute, Curtin University, University of Michigan and the Book Industry Study Group. More information about this project can be found on the project website hosted by the Educopia Institute.

We wish to thank the many individuals who generously donated their time and attention in developing this report. We conducted 27 interviews with 32 individuals and many interviews were expansive and went far beyond allotted times. The information and perspectives we received from these interviews were essential to the report and the development of the Supply Chain Map.

Sincere thanks as well to Brian O'Leary of the Book Industry Study Group who provided comments and guidance throughout the development of this report, and to Christina Drummond of Educopia for facilitating the report's public comment period. Many thanks as well to those who took the time to provide feedback both publicly and privately.

This project was funded by the Andrew W. Mellon Foundation.

Overview

The supply chain for Open Access (OA) ebooks tells the story of an industry which is still in the midst of evolution. The scientific and scholarly publishing industry overall was one of the earliest to develop online and digital tools, with research articles and journals developing online editions starting in the 1990s (and digital editions dating back much further). With the decreasing role of print came the recognition that marginal costs are effectively zero for digital works (fixed costs are, of course, a different matter entirely) and the OA movement was born. But while journal publishers were quick to adopt digital distribution and adopt new business models, books have taken a different path. The reasons for this are simple: while outwardly these digital products may look alike and serve similar communities, structurally books are made with very different bones.

Sitting at the intersection of scholarly publishing and the global book trade, the scholarly monograph ecosystem arose from a confluence of factors. The books business has long been organized around the sale of printed, discrete units — one at a time. Each book takes money to stock and ship, so the supply chain's goal was to make efficient and informed predictions for how many copies would need to be in any one place at any given time. Local distributors arose around the world to serve particular markets, whether institutional or individual, and developed ancillary services focused on each different type of customer. Metadata standards emerged to help libraries accurately catalog and shelve books to make them easier to find. Feeds developed to provide instructions about how to sell and market a title, including all of the relevant information to boost interest in that title amongst potential buyers. Throughout all of this, success was measured in units shipped and copies sold.

Though ebooks had been around long before then, the success of the Amazon Kindle line of e-readers in the early 2010's created a surge of interest in the digital format. Those years gave rise to much commentary and handwringing with a similar premise: "is the print book dead?" As of today, the answer is still "no". But the market did change. Tools and standards were adapted to reflect the possibility of multiple formats and platforms, and the many new ways to sell and read a title. Publishers were able to create new direct-to-consumer or direct-to-institution sales models, and support simultaneous usage of the same copy amongst multiple people in the same library. Ebooks could be purchased in bulk, in large collections, or via aggregators and databases. Many of the limitations of the print access model disappeared. But two things did not: the driving force of book economics, which still largely rely on sales volume, and the continuing reliance of many in the industry on sales of printed books. In this sense,

the emergence of OA models for scholarly monographs represents a more fundamental shift than the evolution from p- to e-. Whereas the move to digital distribution required new methods and models, it did not upend the fundamental units of measure: copies sold (or in the case of aggregators, copies licensed).

One might be tempted to think of the OA books business as essentially the same as OA journals, just a longer format. And while the OA books business has borrowed many of the practices and protocols of journals, and often serves the same customers, books are distributed through a different set of suppliers and intermediaries, each with their own competing interests. These suppliers can also be very powerful, and benefit from information asymmetry about which books reach which customers. OA journals advocates are often wary of a field they feel is tilted towards large suppliers like Elsevier and Springer Nature. Publish books, and one quickly feels the gravitational pull of Amazon.

As compared to journals, monograph publishing is also resource-intensive: the production of one title takes far more on a unit basis than the production of a journal article. Each book must reach its audience as a discrete unit, whereas the scholarly article can rely more heavily on a journal's established brand. This difference results fundamentally different approaches to distribution. Journals (even OA journals) are often distributed on a small number of platforms (often only one or two) and the primary distribution platform is almost always that of the publisher. Books favor a multichannel strategy, where each title can be accessed on a larger number of platforms and formats and in many cases the publisher does not even offer a platform of its own.

These different approaches to distribution have led to the adoption of different metadata standards. They have also led to different challenges when tracking usage data. OA monographs create challenges for the supply chain in both directions of travel. One direction is "downstream" as metadata flows from the publisher through a supply chain that was not designed for OA titles. The other direction of travel is the flow of usage data back "upstream" to the publisher (and ultimately to authors and funders) — a flow that is increasingly important but for which there is insufficient infrastructure and standards.

Clarke & Esposito was engaged by the Mellon-funded "Exploring Open Access Ebook Usage" project team to examine the flow of information, in both directions (downstream and upstream), across the still emergent supply chain for OA monographs. Our remit was to closely examine each step in the supply chain and identify gaps, challenges, and opportunities. Participants in this supply chain span the for-profit and not-for-profit sectors; their dedication

to OA varies from full commitment to incidental contact; and they communicate with one another in differing languages, with varying specificity. Nonetheless, many had valuable insights. Though many of the processes for OA book distribution are new, we found they have been carved from past experience and are in continual refinement. This is a supply chain during a time of transition and the result is a map of the current moment in time, which captures both the legacy of the traditional (centered around print and paid-access) distribution system and the emergence of new approaches, new players, new practices, and new standards.

Methodology

Clarke & Esposito conducted a series of semi-structured interviews with representative stakeholders and subject-matter experts embedded throughout the OA monographs supply chain. Interviewees spanned a number of organizations:

- **Publishers** representing a spectrum of commercial, not-for-profit, library, and scholar-led presses whose catalogues include OA monographs
- **Academic librarians** active in funding and supporting the discoverability of OA monographs
- **Aggregators** of scholarly monographs
- Other **platform providers** and **online distributors** that facilitate the delivery of content to end users
- Other bodies leading inter-industry initiatives to further the viability of OA monographs

In our research we focused on the supply chain as it exists in *intentional* form, meaning we defined it as comprising those organizations that publishers, and their agents, are explicitly interacting with. As many OA books carry permissive CC licenses which allow onward sharing, some titles might also appear on community-driven platforms and tools (such as ResearchGate, or preprint servers like SSRN). We consider these latter examples part of the informal distribution of titles, and not a formally documented part of the supply chain.

Topics of discussion were developed in consultation with the Exploring Open Access Ebook Usage project team and included the value derived from various stakeholder interactions, metadata and other information flow, gaps in the supply chain, pain points resulting from incomplete or missing information, and recommendations for improvements and remediation of problems. (A template for these semi-structured interviews is given in **Appendix C**.) Interviewees were also provided with a brief questionnaire prior to the interview (**Appendix D**),

which they were free to share with the appropriate technical colleagues, to capture general information about metadata standards and identifiers currently in use.

Clarke & Esposito used the findings from these interviews to create a schematic map of the supply chain for OA books, which is included in this report's **Appendix A**. In our documentation of the supply chain we separately captured the two primary directions of information flow: distribution (in which metadata and content flows "downstream" until a title reaches the end user) and usage reporting (in which data flows "upstream" and is gathered into structured reports). This report serves to provide explanatory documentation for both components of the map, and describes in greater detail the key gaps, challenges, and opportunities expressed by interviewees.

A preliminary version of report findings was made available for public comment on the Exploring Open Access Ebook Usage project website from 16 October 2020 – 9 April 2021. The project team and Clarke & Esposito also shared preliminary findings in conference presentations and industry interest groups during that time. Comments and clarifications recommended during the public comment period have been incorporated into this final report.

Key Stakeholders and Intermediaries

Perhaps not surprising for a creative industry in a period of innovation, every business or organization within the supply chain for OA books brings their own unique service and approach. Even so, to document the open access supply chain, our first task was to establish a few general categories and definitions. We have organized the supply chain into seven broad categories, each of which can include multiple subcategories (and sub-subcategories). Suppliers can also appear in multiple places on the supply chain, based on the services they provide.

CONTENT FUNDERS

The creation of an OA monograph can begin even before the first word is written, through stakeholders responsible for the funding of original research, or institutional support of the researcher.

- Some authors and works may have the monetary support of **funders**, who may then set publication policy (and hold authors to OA publication mandates for their work). In this category we would include both funders in original research (such as the European Research Commission, Gates Foundation, etc.), and also contributors to pooled OA

funding schemes such as Knowledge Unlatched (in their fundraising capacity) and the Toward an Open Monograph Ecosystem (TOME) initiative.

- Particularly in the latter circumstance, **libraries** can often play the role of funders, or else play an important support role in funding coordination on behalf of the institution.

Sitting upstream from the author, funders and libraries can have a strong guiding force on the OA model for publication. Funders are themselves often identifiable in specific metadata elements. It is important to note, however, that the order of operations is not always linear in the case of OA funding — in some cases funders can make OA publication a pre-requisite for accepting a grant but OA funding can also appear at any point in the publishing process, and sometimes many years after initial publication.

CONTENT CREATORS

The supply chain becomes more tangible with the introduction of persons and entities responsible for the original creation and production of OA monographs.

- **Authors** are the prototypical content creators, although authorship is comprised of many activities: one can be a primary author, one of many co-authors, and, for compilations of works, a primary editor or one of many chapter authors. As noted before, authors may choose OA publication at the behest of their own funders, or they may need to sign off on OA publication if a new funding source is found to "flip" an existing book's status. If there are multiple authors, there may be multiple funding mandates operating upon the same title.

- **Publishers** are the entities who handle the business of producing the book and sending it on its journey through the supply chain. Even the simple term "publisher" encompasses a diverse array of subspecies: from university presses, library publishers, large commercial publishers, to medium commercial publishers, scholarly society publishers, small commercial publishers, and even small publishers run out of a spare university departmental office. It is a testament to the increasing awareness of OA books publishing that we know of examples of many different types of publisher who are engaging in this emerging supply chain.

While many providers have the capability of introducing or generating new metadata at later points in the process, one rule of thumb seems to hold true: it is important for the publisher (as

the steward of the work) to be responsible for as much of the business-critical data as possible. This means if a title is OA, it should be clearly indicated as such by the publisher — there is great danger in someone downstream extrapolating an OA price point, let alone a license type, if not explicitly stated.

Though we separate them into two different categories within the supply chain map, content creators and funders are united in their interest in observing how the work being produced is accepted into the world, and are generally the most important consumers of information about performance and usage. Thus, while they are the starting points in the supply chain for distribution, they are the ultimate ending point of the reverse supply chain about usage information.

DISTRIBUTORS

Given the variety of content creators, distributors serve the books industry by packaging and normalizing metadata and content across several (sometimes thousands of) content creators, and reformatting the metadata to meet the unique specifications of myriad downstream suppliers.

- Most **distributors** have their origins in the print book, and later the "traditional" (paid-access) supply chain. Their primary role is to provide publishers with access to downstream retailers in both B2B and B2C markets. Some examples of this type of distributor are Ingram Academic, Hopkins Fulfillment Services, and Longleaf Distribution Services. Traditional distributors may handle OA, but mostly as a service to publishers working to get OA titles onto the same platforms as other traditional titles. Their systems and processes are based around the concept of price points and units of sale.

- Given the challenges for traditional players to handle zero-price products, additional **OA distributors and services** have emerged specifically to address this gap in the supply chain. These distributors are distinguished in their focus on OA platforms and repositories. This role may be driven more by necessity rather than profit motives, as these providers often play multiple roles in the supply chain and do not consider distribution their primary value proposition. Examples here are Knowledge Unlatched (who we note are also involved as funders and as the creator of an OA ebooks platform), and Unglue.it. The work of OA distribution is also sometimes organized within pilot OA initiatives, such as TOME or the Sustainable History Monograph Pilot (SHMP, which is led by the University of North Carolina Press and its Longleaf Distribution Services).

Over time, as traditional and OA distributors establish uniform processes and capabilities, we expect the line between these two subcategories to blur, and in fact for purely OA distributors to focus on other ways to deliver value. One important example of this is the additional role the OA distributor plays in monitoring usage. Given the nascent marketplace for OA books, distributors like Knowledge Unlatched are taking a more active role to gather and consolidate usage reporting for titles (or are actively investing in projects to do so).

Traditional distributors, on the other hand, speak the language of book sales, and where they typically do provide robust dashboards for reporting purposes (like Ingram iQ, a service for Ingram publisher clients), these dashboards typically do not incorporate usage. This is because historically, there is little need for the traditional distributor to monitor engagement with that title once the sales transaction is complete.

CONTENT PLATFORMS

To ensure the broadest possible usage, many book publishers favor a multi-platform strategy, meaning a publisher will send (or engage its distributors to send) the same title to many different platforms to increase its potential exposure and usability. We define content platforms as the site where content is aggregated and hosted and linked to, waiting for discovery by the end user.

- **Publisher-specific platforms** (sometimes operated via a third-party platform provider like Silverchair or Atypon), are a way for a publisher to directly control and monitor the access and usage of its ebook titles in one location. These platforms — a few examples of which include SpringerLink, Taylor & Francis Books Online, Cambridge Books Online, and myriad others — are typically based on institutional access, in which case OA books co-exist alongside paid-access titles. This benefits both kinds of titles by, helpfully, giving the library patron more content to search and use, and increasing the visibility of paid-access content to users discovering OA books off the web. Given the awkward user experience this can create (paid-access titles require institutional authentication to view; OA books do not), some publishers have created fully OA platforms (examples being Ubiquity Press, the platform for the University of California Press's Luminos Imprint, and Manifold from the University of Minnesota).

- Because not every publisher has the appetite to develop its own platform, **ebook aggregators** license and consolidate titles from many publishers into one combined database, where they are then purchased by institutional customers. The most well-

known examples of these are EBSCO eBooks, ProQuest Ebook Central, JSTOR, and Project MUSE, although additional smaller players exist for particular markets and disciplines. These aggregators serve primarily a sales function, though some have begun to carve out sections of their platform specifically for OA books (such as OA Books on JSTOR and MUSE Open Monographs). Publishers who host and aggregate content for other publishers (such as Oxford University Press Scholarship Online, DeGruyter, or Michigan Publishing's Fulcrum) also fit into this category. These publisher/aggregators, like other aggregators, often combine OA and paid-access titles for greater discoverability and convenience.

- **Consumer ebook platforms** offer titles for an individual's use and access, and do not actively support institutional or library integration. These are primarily retail businesses such as Kindle, Kobo, and Nook, which are not focused on the scholarly monograph ecosystem per se but are nonetheless important as they will be familiar and well-used by individual readers. OA books on these platforms are typically added to a user's library after they conduct a zero-dollar "purchase," remaining immersed in the language of paid-access books.

- Most of these platform types are dependent on, or adjacent to, paid-access distribution of ebooks, and have incorporated OA books as an innovation. **OA platforms and repositories** have no underlying infrastructure for the buying and selling of books, and are intended to host exclusively free or OA content.[1] These can include well-known centralized OA platforms such as the OAPEN Library, the Internet Archive, Hathi Trust, and Project Gutenberg, but also institutional repositories where researchers or librarians can deposit content (possibly under the formal operations of a library publishing unit, which are a common source of OA ebooks especially in Europe). While such repositories have fewer ties to the primary channels of book distribution, they play a valuable role in book discovery, and have oftentimes been amongst the most innovative in ensuring robust indexing of OA titles within search engines and other discovery services later on in the supply chain.

As the place where content is hosted and access is granted, content platforms also play the critical role of gatekeepers for many individual usage events — they hold the files, and serve up the ebook to the appropriate person (and, particularly with paid-access models, they can also

[1] Sometimes an OA platform will refer the user to a bookseller or publisher website for purchase of a print or paid-access copy, but any resulting transaction typically takes place away from the OA platform itself.

check IDs at the door). This makes platform providers the natural progenitor of usage reporting, in that they are responsible for consolidating usage events into structured reports, and making those reports available to customers and suppliers.

SALES CHANNELS

Whereas platforms are simply the place where books are aggregated and hosted, the *sales channel* is the intermediary which — stemming from the traditional book supply chain — transacts with the end user or institution to provide access (along with any other post-sale services such as invoicing and records).[2]

- **Library sales channels** focus on sales of books in multiple formats (including print) to institutional library customers, following B2B purchase and access models. Primary examples of this type of sales channel are EBSCO's GOBI and ProQuest's OASIS. While "OA sales" to libraries is a bit of an oxymoron, these library sales channels are relevant in this context because they may still facilitate the purchase of paid-access versions of OA titles (including and especially print). Library sales channels also provide help with catalog records, integration into library management systems, and can aid selection of titles with customized profiling and selection services. It is worth noting that, for the sales channel, a customer's usage of a particular title is of interest only insofar as it underpins purchase activities, for instance in the service of demand-driven acquisitions (or DDA, where patron usage can trigger an automatic purchase) or evidence-based purchase models.

- Similarly, **consumer ebookstores** (and by this we mean primarily Amazon, although providers such as Kobo, Apple Books, Google Play, Scribd, and eBooks.com continue to serve portions of the market) are the B2C storefront serving both the scholarly and the trade book markets. As the level of understanding of OA is much lower amongst the general lay reader, OA titles may appear in these channels as simply "free" (alongside, of course, any paid-access versions). While it is entirely possible (even likely!) that a provider like Amazon is able to capture and act on the usage information generated by its Kindle e-reading platform (including popular highlights, etc.), we did not see evidence

[2] We note that in many cases, particularly for large publisher platforms and aggregators with robust institutional sales teams, platform providers may conduct business directly with customers and bypass the sales channel entirely. In this case, it is incumbent on the platform to provide the same kind of post-sales support as would be expected by the sales channel.

that they formally report this kind of usage data to suppliers. Rather, the number of users downloading a book is expressed as a number of zero-dollar "sales".

While we draw distinct boundaries between B2B and B2C sales channels in this report, the distinction does have the potential to break down in some contexts. The Library Acquisition Patterns Report published by Ithaka in 2019[3] found that, while GOBI was the predominant acquisitions channel for print books amongst academic libraries, Amazon came in a close second. Print books, however, do not require an access model. For ebooks, Amazon's platform does not have a B2B-appropriate access method, and they hold no meaningful presence amongst library purchases.

CATALOGS AND INDICES

These are platforms which host metadata and promote discovery of particular titles to end users. Indices play an important role in synthesizing lots of metadata sources — they are where many metadata streams are combined, and records are de-duplicated and "FRBR-ized" (mapped according to IFLA Functional Requirements for Bibliographic Records) into a more orderly structure.[4]

- At the library level, the **Library Management System** is where a librarian can manage their institution's locally curated catalog of titles. The Library Management System (examples of which can be Library Service Providers like ExLibris' Alma, the open source FOLIO project, or OCLC's WorldShare Management Services, but which also can include more conventional integrated library systems (ILS) like those by SirsiDynix or Innovative Interfaces, alongside associated library administrative modules) also play an important role in consolidating usage and expenditure reports across many platforms, providers, and content types, for the use of library administrators. In many cases, this usage is gathered in service of future purchases, for example to renew subscriptions or re-stock a DDA program. With OA books, the library might also be making decisions to fund or subsidize the publication of OA titles (which makes the librarian a stakeholder in content creation).

[3] Available at the following link: https://sr.ithaka.org/publications/2019-report-library-acquisition-patterns/
[4] See Barbara Tillett, "What is FRBR: A Conceptual Model for the Bibliographic Universe," Library of Congress Cataloging Distribution Service, Revised February 2004, available at: https://www.loc.gov/cds/downloads/FRBR.PDF

- **KnowledgeBases** are library-agnostic, global content indices (which, for simplicity, we also use as an umbrella term to include any accompanying union catalog). The most well-known examples of these are OCLC's WorldCat KnowledgeBase and Alma's Central Knowledge Base. KnowledgeBases can contribute to Library Management Systems by providing supplementary metadata to catalog records, and the Library Management System and KnowledgeBase together underpin the index for the library's Discovery Service. Depending on their needs and preferences, librarians may incorporate OA titles into their catalog and discovery service alongside their other holdings to make them discoverable by their patrons, typically by loading a whole platform or collection.

- **Third-party content indices** are more specialized types of products that promote metadata curation and discovery. Relevant examples include the Directory of Open Access Books (or DOAB, which indexes qualifying OA books), and Unpaywall (which indexes the OA status of most publications with a Crossref DOI, including OA books, though its primary use case is journal articles). Not only are such databases important channels for discovery, but they can also generate new metadata about a title to aid in evaluation and selection (such as Clarivate Analytics Book Citation Index or Digital Science's Dimensions, which track citations for books across many titles, regardless of OA status). KnowledgeBases, for example, often rely heavily on metadata pulled in from such sources, and many catalogs trade and exchange metadata between one another.

- **DOI Registries** play a special role as a content index — and within this subcategory, we focus on Crossref in particular as integral to the OA books supply chain. Crossref is the primary issuing body for book and chapter DOIs (Digital Object Identifiers), or persistent and unique URLs which resolve back to a particular title. As the DOI proliferates across platforms, catalogs, and indices, Crossref then tracks usage, or "resolutions", for that DOI. The number of DOI resolutions then become their own data point about a particular title (albeit a nuanced one, as we discuss later in this report).

- Beyond the specialized indices and catalogs for scholarly use, **web-scale search engines** crawl vast swathes of information across the web to serve up useful links to users, and as such play a critical role in connecting content to audience, even when that audience is outside the technical and physical confines of a library ecosystem. Between them, Google Scholar and Google have generated vast amounts of traffic for open access content in the journals space, and books platform providers have in recent years taken steps to ensure books content is equally well-indexed and included within relevant results.

Where content creators are the primogenitor of a great deal of ebook metadata, catalogs and indices are the only other stakeholder who typically creates additional metadata. Many catalogs and indices make it their business to better describe and organize titles, and provide the supply chain — particularly those interested in discovery — with the necessary information to place the ebook on the right "shelf," or in the correct search result within a database. Every catalog or index organizes information according to their own methodology (and increasingly using automated and machine-driven means), and as such presents a downstream partner or user with a different way to search for the information they need.

END USER INTERFACES

End users are the final step in distribution, where a reader interacts directly with their chosen title. This is where the process reverses and usage is generated, thus initiating the flow of usage reporting back through the supply chain.

Every title will naturally have a different set of end users as its audience, so we do not generalize them here. We do differentiate between the different tools that readers may use to generate the usage that we measure:

- If the end user uses an **online ebook viewer**, they interact with the book within a webpage, or on an internet-connected device. In this situation, the user's activity is visible (and trackable) to the platform or server. It is possible to capture what portions of a book are read and when a reader accesses features such as annotation or highlighting. Examples in this category include books in browsers (e.g. Chrome, Firefox, Safari) as well ebook viewer software built into a platform provider's website.

- "Online" usage can also trace back to **ebook viewing apps and devices**, to the extent that those apps and devices can send data back to their parent platform. Kindle apps and devices are a primary example of this: though the user is not explicitly using a web browser and may even be working in "offline" mode, the platform is still capable of sending usage data about reader behavior back to Amazon whenever the device syncs.

- If using an **offline ebook viewer**, on the other hand, the user downloads the book and reads it using third-party software with no further connection to the platform, like Adobe Acrobat, or Bluefire Reader. This viewer is "offline" in that, even though there might be an internet connection, there is no further tracking (or if there is, that tracking is not part of the ebooks supply chain).

These different types of usage can enable different types of reports when aggregated by the platform (which as noted receives and aggregates the relevant data about use). The limits of reporting from an online ebook viewer are dictated by what information the platform provider is set up to capture, and the boundaries of privacy. Reporting on a book taken to an offline ebook viewer is more limited. The platform knows only that the book has been downloaded; beyond that the user's behavior is unknown.

Metadata Standards

Each of the providers above interacts with other bodies in the supply chain through a complex handoff of metadata and content, each transaction requiring a different language to get the necessary information across. We have endeavored to catalog where various information standards are in use to send information both upstream and downstream, and we note below where particular data elements undergird the OA books supply chain as a whole.

PRODUCT IDENTIFICATION STANDARDS

The most critical standard is the one which defines "the book" itself. There is inherent difficulty in establishing unique product identifiers for digital titles when they can be available in multiple formats and across a variety of platforms (and can also, in many cases, have a physical manifestation floating around the world as well). NISO's E-Book Metadata Working Group's preliminary E-Book Bibliographic Metadata Requirements[5] acknowledged the overall difficulty of applying a single standard identifier for each title, and in fact recognized multiple identifiers required for every book. There were two such identifiers in particular which our research found to be critical to the operation of the OA books supply chain:

ISBN

The ISBN was the most common and most critically important of the unique book identifiers across nearly every provider we spoke with. With very few exceptions[6], the ISBN is a required piece of metadata and fundamental to the operations of suppliers and intermediaries.

This is about as far as the agreement goes. In practice, ISBNs are assigned very differently by each publisher — particularly when it comes to OA titles. There is general agreement already

[5] Available here in its first draft https://www.niso.org/standards-committees/ebmd; the commentary period ended on August 2, 2020.

[6] The notable exceptions are public access books on fully-OA platforms such as Project Gutenberg, which are likely not within the category of scholarly monographs, and monographs posted to institutional repositories, which tend to interact very little with the rest of the supply chain.

(which is codified into the supply chain) that print and digital books should each be assigned separate ISBNs. When a title is produced OA from its inception, the OA ISBN is simply the digital ISBN (with sometimes multiple digital ISBNs if the publisher wants to differentiate EPUB formats from PDF, and so on).

Complication arises when a title is born paid-access and becomes OA later in its life, or when a title is available simultaneously as a paid-access book in some channels and as OA in others. We found no uniformity amongst publishers as to best practices. Some publishers establish a new ISBN for the OA version of a title, which either replaces or co-exists with the paid-access version (depending on the publisher's sales practices). Other publishers simply re-use the same ISBN, which requires overwriting or backing out its previous pricing and license data. Both approaches carry some risk that there will be confusion between the title's two "modes" (paid and OA), and the very real potential that an entity downstream will not get the crucial update about the title's sales status.

DOI

Digital Object Identifiers, or DOIs, are meant to be unique identifiers assigned to a particular piece of content. For OA ebooks this can be either the full book or book chapter, depending on what the publisher wishes to make discoverable. While not all DOIs are issued by Crossref (DataCite is another popular registry), Crossref is widely used as the source of DOIs for research publications and has pre-existing relationships with the journals divisions of many OA books publishers.

Uptake of DOIs is less uniform throughout the books ecosystem as compared with journals (where they are nearly universal). There are logistical challenges to their implementation: where a publisher does establish a book-level DOI for a title, that title may appear on multiple platforms. To deal with this, Crossref enabled Multiple Resolution, which ties together all locations where a digital is hosted and ensures a single view of all metadata and citations.

However, not all book publishers establish DOIs for their titles in practice, and so some platforms and aggregators take it upon themselves to establish brand new book DOIs. This can result in the same title having multiple DOIs, each leading to a different platform (which acts somewhat antithetically to the DOI as a "unique" identifier). In order to resolve this issue Crossref also offers Co-access for ebook DOIs, which provides a "last resort" to link DOIs and allow the user to see the many different platform options for each title.

In practice, neither Multiple Resolution nor Co-access are a complete solution, as currently implemented, as they cannot overcome messy handling within the supply chain of both book and chapter-level DOIs. The DOI simply does not hold the same cachet to most players as the ISBN. This makes it a challenge to handle monographs within value-added services designed to track the performance of a particular title, when those services are based on the DOI (such as Altmetrics and Unpaywall).

Other Identifiers

Just about every intermediary in the supply chain assigns their own internal unique identifier for each book to relate the various formats. Some of these unique identifiers have risen to the level of an industry standard shared by multiple intermediaries, such as the OCLC Control Number (or OCN) and Library of Congress Control Number (LCCN). But these identifiers were typically important to particular steps in the process, and none of these rose to the serve the same end-to-end utility.

This speaks both to the importance of both the DOI and the ISBN in the supply chain for OA ebooks, and the difficulty of integrating a new identifier to solve the challenges posed by an industry built on a multi-format, multi-platform model.

DISTRIBUTION FEEDS

Things are no less complicated when it comes to the standards around distribution feeds. Most standards were developed without the other standards in mind, and as a result there are no clear crosswalks between ONIX and MARC, or KBART and ONIX, or KBART and MARC. In the case of OA books, there is the real risk, if each standard evolves capabilities before the others, that it will lead to metadata that are “lost in translation”.

ONIX

ONIX is the international standard for communicating book industry product information. It is the primary language of sales channels, because it includes fields for any and all pertinent metadata for pricing and marketing. ONIX feeds are primarily developed by the publisher or content creator and shared with distributors, retailers, and platforms where a book is exposed to customers.

In building new elements of the supply chain, one must typically start upstream to build the containers, before moving downstream to the receptacles. This can be a challenge — as we’ve noted, publishers are a diverse lot and some continue to make heavy use of manual processes (relying on downstream intermediaries to perform the transformation and automation that

makes for an efficient supply chain). It seems everyone's ONIX file is different; one provider called it "a format not a standard." Many fields are optional, and each publisher and platform implements these fields in their own way (some publishers' feeds even require staff to populate relevant metadata fields by hand).

OA license information is possible to incorporate into ONIX feeds, but our interviews uncovered varying levels of success. Providers need to support OA-specific data points, which sometimes requires development work, for both the entities sending and receiving the feed. In the meantime, as one interviewee told us, "OA has many publishers working ***around*** the metadata feed" (meaning, generating a separate manual feed for OA titles because their standard ONIX feed doesn't contain the right information). This may simply be a product of time and place; several publishers we spoke with expressed a preference to incorporate pertinent OA metadata into ONIX (as the standard allows), rather than compile this data manually. What these publishers lack (so far) is simply the time and resources to update their systems.

KBART

KBART is the connective standard which allows hosting platforms to share information about titles to a central KnowledgeBase (which then provides catalog records to a Library Management System, and supports discovery via a library's Discovery Service).

While KBART does allow individual titles to be flagged as OA, the organization of a KBART file can play a more important role in facilitating discovery for publishers and platforms who have both OA and paid-access books. Some hybrid platforms establish separate collections in their KBART feeds, exclusively comprised of OA books, so that libraries can incorporate those OA titles regardless of whether they buy anything else on the platform. In other cases, a platform may offer bundled collections of titles for sale, and will include OA titles in that collection's holdings file (though of course not the collection price). Where such a collection is incorporated a librarian will be able to integrate both OA and paid-access titles into their holdings with one step. Both options have the effect of handling OA titles as a single, holistic collection — there is very little designed into this process to support title-by-title selection activity for OA titles.

MARC Records

MARC is the format used to load metadata into the Library Management System and populate the library catalog. As such it is a foundational format for title discovery by library users, and for holdings management. There have been recent changes to the MARC standard which allow for designation of OA status in the MARC record — but this change needs to be adopted by those cataloguing the records to truly take effect.

Librarians are rarely themselves performing original cataloging or contributing metadata about OA titles to a Union Catalog or KnowledgeBase about OA titles. This is a more common practice when acquiring books on a title by title basis, especially in print, and there is little motivation to perform this time- and resource-intensive process on titles the library has not paid for.

Instead, there is a greater reliance on MARC records freely available from platforms and providers (which can be of varying quality). Whereas some libraries do pay for records from intermediaries for paid-access content, paying for higher-quality MARC records may be difficult to justify within the library budget when titles are managed as large collections, the library does not own the books being cataloged (which raises concerns about preservation), and the content is otherwise free.

The MARC standard itself has many variations, and is limited in the amount of information it can capture about works in multiple formats and iterations. The BIBFRAME cataloging standard has been explored as one alternative to address some of these deficiencies by operating on a more robust linked data model. It was rarely mentioned amongst interviewees, however, and it appears it is not yet active in the marketplace, nor supported by the supply chain.

OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[7] standard enables a different method of discovery, without requiring manual cataloging via library channels. Originally developed by the Open Archives Initiative, OAI-PMH enables platforms and repositories to encourage harvesting and indexing of their content, and promotes discovery via web services, citation indices, and other databases.

In practice, this standard was developed primarily to serve institutional repositories and platforms where all content is open access. It has been under-utilized and indeed may have limited application for mixed-model platforms and aggregators where some content continues to be available on a paid-access basis. However it is a valuable standard for intermediaries tasked with capturing and indexing OA content, as a means to retrieve information and promote discovery of repository content in a simple and efficient manner.

[7] For more information, see https://www.openarchives.org/pmh/

USAGE REPORTING

COUNTER Standards

As noted above, the platform provider typically holds the responsibility of aggregating and reporting on usage of the titles accessed by the end user. In portions of the supply chain serving the library market, most platforms capture this information in COUNTER-compliant ways. The release of COUNTER 5 attempted to address the central challenge in consolidated ebook reporting across platforms by introducing a new "Unique Title Request" measurement. This is intended to allow comparison of usage reports regardless of whether the book content is served by as a series of chapter files or as a full-book PDF (one of the main shortcomings of COUNTER 4 reporting). COUNTER 5 also allows for separate reporting for gold OA and paid-access content.

In practice, publishers who have implemented or are working to implement the COUNTER 5 standard find there is still room for improvement: an independent review of the COUNTER 5 standard published in October 2020[8] suggested there remains a potential for providers to interpret the standard differently across platforms, and gaps within the definition of "Open Access" for books (for example, books that are "free to read", or those which contain OA chapters). Some providers are still in the implementation process for COUNTER 5, and so continue to report in COUNTER 4. Regardless, COUNTER has taken note of the need to improve its standards to support books hosted across platforms.

Despite this progress, it's likely that a meaningful portion of OA ebooks usage will remain opaque – not because of the format, but because of the access model. The COUNTER 5 standard (like all previous COUNTER standards) is designed to *provide librarians with information about usage by their patrons*. This leads to certain decisions about how to report usage if the patron is outside the library network or does not have their device tethered to the library network for off-campus use. A COUNTER report on an OA collection is likely to underestimate the amount of usage generated by a particular institution, because unlike access-controlled content, users outside the university network are not prompted to authenticate (which few librarians may be concerned about, given the resources do not cost budget dollars and there are many more costly demands on the budgets for time and attention). Reports generated for publishers, on the other hand — which are not explicitly required for compliance with the COUNTER standard, but which typically follow the COUNTER

[8] See "COUNTER Release 5: An Independent Review" by Dave Jago, Mark Sandler, and Alicia Wise, available at https://www.projectcounter.org/wp-content/uploads/2020/10/COUNTER-report-Oct-2020-for-public-release.pdf

format (because that's what many platforms support) — will show such usage as anonymous "guest access".

A further challenge with usage report for OA books is the problem of robots. Whereas COUNTER reports are measuring usage of authenticated users on a particular network, OA content is, by definition, available on the open web. "Guest access" totals will often include high volumes of automated access from robots and crawlers moving through the web. Platform providers are obliged to set up filters to strip these crawlers from their usage reporting (and COUNTER provides an official list of bots which *must* be excluded), but again here is where each provider might be challenged to keep pace with implementation in a constantly changing environment. These seemingly minor details can substantively impact the way reports from two different platforms might be compared against one another, even when both are technically following the same standard.

HIRMEOS

It is also important to note that not every platform will have the interest or capability to accommodate the COUNTER 5 standard in the first place. Retail and consumer channels have no use for it, and indeed there is no set standard for usage reporting back to the publisher from these providers. Open platforms without a library sales component, which are freely available directly by end users, might choose to establish COUNTER 5 capabilities, but there are many who will not.

The High Integration of Research Monographs in the European Open Science Infrastructure (HIRMEOS) project, funded by the OPERAS group,[9] considered a new standard, and has produced Open Source code that allows publishers to aggregate usage and alternative metrics for OA publications.[10] While this was an important step towards a presentation of usage data more appropriate in the context of OA books, there is still work to be done, as many platforms don't yet allow for programmatic harvesting of usage data.

DOI Resolutions

As noted above, Crossref plays an important role in the supply chain for OA ebook usage, and are perhaps the only example of a provider who is able to provide some insight into title usage despite not hosting the content itself. Where a title or title chapter has been assigned a DOI,

[9] More information about HIRMEOS is available here: https://www.hirmeos.eu

[10] Open Book Publishers provides an excellent example of how this source code might be implemented, including details about the sources and methods in use, at https://www.openbookpublishers.com/section/84/1.

Crossref will keep track of how many times traffic flows through that DOI URL. A count of DOI resolutions is then reported to publishers (or whoever that DOI is registered to). Again, as there are sometimes multiple DOIs for the same title, some of which are registered to different parties, there is still no guarantee that the recipient of such a report will truly be able to see all resolutions for that title. It is also the case that DOI resolutions will not capture all the usage even to titles that have registered DOIs. That is because books can be accessed on most platforms without recourse to a DOI. A user browsing to a book on a publisher or aggregator platform, for example, may never encounter a DOI.

Non-Standardized Usage Reporting

Platform providers may have insight into more user activity than what is captured in a COUNTER report via Google Analytics or similar business intelligence tools hooked onto the platform's back end. There are a few issues in being able to make this kind of analytics data available to others: First is the important question of privacy around specific end users. Google might in fact know that a user is associated with a particular institution, but it goes against library principles and (in some cases) laws and regulations to incorporate this into a report to that institution without the user's consent. Second is an epistemological limitation: Google can only tell the platform what happens on the platform. The data gathering stops when a user downloads the book to read in an offline reader (or even loads that file into an online reader hosted by a different provider). As with COUNTER reports, incomplete information around end user behavior can lead to unreliable or even misleading metrics. When it comes to the full universe of usage, even Google doesn't know everything.

Challenges and Limitations in Usage Reporting

In our interviews, we found this potential for "incompleteness" and inconsistency leaves many publishers hesitant to conduct robust analysis of usage, despite acknowledging its importance. It is difficult to have confidence in an analysis where some as-yet-unknown amount of data is not available and that which is available raises nearly as many questions as it answers. Compared to a sales report, where one unit shipped looks the same in every channel, usage captured via different standards requires a higher degree of interpretation and qualitative insight, and there is hesitation to draw conclusions from partial and inconsistent reporting.

There is also the challenge of extrapolating from usage data to arrive at higher-level value measurements (such as outcomes). In a previous iteration of this project's research on the OA Books Supply Chain[11], an important finding was the wide range of value that might be derived

[11] "Mapping the Free Ebook Supply Chain: Final Report to the Andrew W. Mellon Foundation", published in June 2017; a copy is available at http://hdl.handle.net/2027.42/137638

by different users of an ebook. There are plenty of instances where it is more important to understand what that book is used *for*. If an OA monograph is chosen for classroom use, does usage tell the whole story? There are many measures of engagement which, in addition to usage, might become relevant to the supply chain's concept of "success", when sales are no longer the central driver.

Usage, finally, is an activity that happens in time. This is relevant because titles can become OA at many points in their life cycle — they can be born OA or they can be made OA later in life. The OA version of a book might replace the paid-access version — one ISBN disappears for sale, another appears for free — or the two versions may continue to co-exist (whether by accident or on purpose). Alternatively, a paid-access ISBN might suddenly become an OA ISBN. There are many choices a publisher might make when interpreting that use – for example, should usage be segregated between when the title is OA versus when it is paid-access? How should the publisher interpret usage of the paid-access version of the title already purchased and available in a library's holdings? These are complex decisions, and perhaps a good opportunity for community discussion and normalization around best practices.

OA Metadata Elements

Many of the most common standards described throughout this supply chain were not designed with OA content in mind. Pricing and "salability" play such a foundational role for many providers that they are not yet able to support titles without a price point. Some publishers and platforms have attempted to work around this restriction by pricing titles at $0.01, but this is clearly not a sustainable approach for the long term, nor is it sufficient to handle the many nuances and manifestations of OA licensing.

Our conversations of sales channels and distributors revealed general disagreement about how to overcome this price requirement. Some are working to allow a $0.00 price point (or are already accepting $0.00 titles into their system today). Others consider it more accurate for pricing to be indicated as "undefined" (which is often the same pricing status used in early metadata for Not-Yet-Published titles). Some are asking upstream providers to filter out OA titles from distribution feeds until they can agree on an approach (which creates the potential for unintended sales activity when a paid-access title has been flipped to OA).

Outside of the environment of the sales channel, pricing metadata becomes less of a factor. The patron at a library has no concept of how much the library paid for that book. Of course, OA

books are more than simply “free”. HIRMEOS has provided foundational guidance to identify other key elements that are mission critical to the creation and management of OA book titles across platforms. Necessary elements include those pertinent to *how* a title is OA, and *why*. This will include the title’s license type — whether CC-BY, CC-BY-NC, CC-BY-NC-ND, and so on. Other key identifiers support the requirement to report information about a title back to funders and stakeholders, and include the persistent identifiers Funder ID (for the funding source) and ORCID ID (for the author). In the interests of preservation, HIRMEOS also recommended use of the DOI as a persistent URL directing users to that title (the challenges of the DOI as a unique identifier have been explained in detail above).

These metadata elements are powerful and important indicators for the supply chain, and in (in the case of license type) dictate what an intermediary and user is allowed to do with the content, so there is already progress in adding these fields to be machine readable in ONIX and MARC standards. (KBART focuses more on collection-level detail and has an indicator which simply differentiates “paid access” titles from “free”.) We have previously described how both ONIX and MARC standards can vary wildly in implementation. Progress is steady, but unfortunately it needs to happen one publisher at a time. We bear in mind as well that these metadata elements are important to STM and HSS academic publishing, but the supply chain for scholarly books often sits on the same train tracks as the retail market. In retail there is no concept of OA, and no perceived value in doing extra work to incorporate OA books (let alone their metadata elements). It may be the case that retailers like Amazon or Kobo will never incorporate detailed licensing information into their systems. Some publishers get around this recalcitrance by ensuring that license information is ensconced in metadata feeds to these channels for display in human-readable format. Nevertheless, the result is that in practice there is currently no single way of flagging in all channels whether and how a book is OA.

The current supply chain also makes the fundamental assumption that monographs are always OA or paid-access at the title level. In practice, there are situations where an author contributes only a monograph chapter, but that author’s funder insists on the chapter being OA (even when the book is not). Most systems and standards don’t contemplate different business models for different chapters (think about that platform provider who only serves the book as a single, complete file!) and in practice this may be fairly rare. It is nevertheless clear in our conversations with several publishers that such titles are unlikely to ever go away.

Gaps and Opportunities in the OA Books Supply Chain

Our conversations with stakeholders from beginning to end of the OA books supply chain revealed many ways in which information can be lost or processes can break down. The print book industry has had hundreds of years to develop a language for its supply chain to share business-critical information with all necessary stakeholders. The digital book industry was able to adapt this language for its own purposes and establish parallel operations, handled by many of the same organizations. OA books operate on a fundamentally different model, and require further innovation to articulate the most important facts about a book in a way that is understandable to all parties.

Through our interviews with stakeholders across the supply chain, Clarke & Esposito has identified five key findings related to the role of metadata and usage throughout the OA books supply chain:

1. **Journal-based standards and models are a poor fit for OA books**
   When viewed on a website, there is very little to distinguish a book from a journal. Though very different product types, in practice they often share certain formats (particularly PDF and XML), the same platform, and at times the same producer. The journals ecosystem has also evolved much more quickly to accommodate OA business models, and when speaking about OA books providers often borrow the same terminology.

   This becomes even more apparent as books are increasingly "chapterized" in their delivery, and those chapters begin to take on a life of their own (multi-author contributed volumes are colloquially referred to as "journalbooks"). With the increasing prevalence of OA mandates across all publication outputs, it is likely there will be more book content which is made OA at the chapter level. This is not something easily articulated in book metadata standards and is a challenging exception for the supply chain to handle.

   The need to capture more and better information about individual chapters has led to attempts to retrofit certain systems to make chapters look more like articles. (It also leads to the train of thought that books are somehow "behind" journals.) The industry has room to improve and normalize the management of chapter-level DOIs. There is also a value in making sure indices and catalogs have access to this chapter-level information, which can be critical in connecting the right content with its intended reader.

That said, the books supply chain will never be the journals supply chain. Most of the intermediaries, and in fact a good many content creators, have a need for the book as a single complete unit. A better conclusion is that the industry is in need of product-appropriate standards for capturing chapter-level metadata, including OA metadata, and making use of it in downstream systems. Many interviewees articulated the need for those service providers with their foundations in journals metadata — Google Scholar, Altmetrics, Crossref, Unpaywall, or many other data aggregators and indexes — to adapt data capture in a way that more accurately reflects the many ways that books are constructed differently.

2. **The supply chain is built for paid access and incentives are aligned for paid access – not necessarily for OA**

   It took time for publishers to evolve workflows from "print first" to "digital first", so it is reasonable to expect similar growing pains to move to "OA first." But just as digital has not fully replaced the printed work, OA is likely to co-exist for some time (perhaps a very long time) with paid-access works. In fact, one of the primary differences between the evolution of OA between books and journals is that, with books, the print copy often remains an important companion product, and it is reasonable to expect this version will be continue made "for sale" (even if on a purely Print-on-Demand basis). This means the traditional supply chain is likely to remain relevant and important for many OA books publishers.

   The inherent challenge, however, is that the starting point for the print supply chain is built around *a sales process* — and there is little incentive for some key intermediaries to participate if there is no money changing hands. Each step in the process has an incremental cost, for which distributors and sales channels typically expect to be compensated. Without compensation, there can be little business justification to update the technical infrastructure to accommodate OA metadata, invest in marketing, or provide any other service which incurs incremental costs. Even within those publishers and suppliers working to develop a robust OA program, we heard lingering uncertainty about the role of sales and marketing personnel, including whether to set "targets" for OA books with no true sales activity, especially when the organization has only a nascent idea of how to interpret usage patterns in a way that defines "success".

   It is tempting for OA publishers and advocates to foster new supply chain participants that can bypass traditional players and thereby elide many of these "square peg in round hole" problems. However, so long as the print book persists, end users continue to look to sales channels for both paid access and OA books, and publishers themselves continue to publish

with mixed models, it seems doubtful that the traditional supply chain can be bypassed entirely. If this is the case, OA publishers and other stakeholders will need to continue to explore ways to keep intermediaries interested, and for intermediaries to recognize new ways to gain benefit from the flow of OA content.

3. **Distribution processes are complex, and do not easily handle changes to a title's OA status**
   Publishers are enthusiastic metadata sharers. For any title, the flow of books metadata about a title begins before the book is even published. Pre-publication metadata is sent to sales channels months in advance of publication, for title marketing and so to help print-based sales channels predict inventory. Publication metadata completes the title record and establishes the book as "for sale" in its various formats. Later updates can overwrite an older price point, or indicate that the title is out of stock. Though the documented map creates a static representation of information transfer from one stage to another, it also bears in mind the adage, "The map is not the territory". We were struck in our findings as to how dynamically data flow, and how constantly. What we've uncovered is a complex plumbing system, where information flows through the data pipes at steady speed (and throughways can be clogged, and sometimes bypassed altogether). The flow of return information is constant in response, and as difficult to separate out to its source as the water from the hot and cold faucets.

   A book can become OA at any point in this process. A book originally conceived of as paid-access might receive OA funding just before its "for-sale" date. A title might exist for years before being "unlatched" and made OA. No common understanding seems apparent about what to do about this transition, or the data that are already flowing through the supply chain. Some publishers will issue a completely new ISBN and deprecate the data associated with the old ISBN. Some publishers will refresh all of the data for the original ISBN; that OA title now has a "sales history" (which is historically valid, though confusing for reporting purposes). In either approach, not all vestiges of the previous version will disappear. Where channels who can't support the $0.00 price point, the OA pricing over-write may fail, and inappropriate sales activity may continue.

   It is dangerous for downstream providers to make assumptions about what to do in these situations — just because an OA title exists does not mean the paid-access version is an accident. But failure to remove an erroneous paid-access title can violate the publisher's intended license, and lead to unhappy customers. This means it is critical for the supply chain as a whole to work well from end to end, in full recognition of both business models, and with the understanding that no two titles are made OA alike.

4. **Existent standards and practices are not yet firmly established for OA monographs, and some are insufficient**

   Ebook distribution relies on numerous standards, most of which were not designed to be interoperable. Adding to the confusion, OA monographs require the introduction of new data elements, and there are currently varying levels of uptake. Publishers greatly preferred to set the correct OA metadata indicators which would then flow downstream — inputting this data via manual intervention if necessary — but the truth is that intermediaries have sometimes implemented standards in different ways (with some platforms issuing their own DOIs, and others differing in their use of "$0.00" and "Undefined" price points). It will take time and some persuasion before OA metadata elements are present as a standard and consistent element for most publishers and suppliers engaged in OA book distribution.

   Even more problematic are the standards for sharing information back "upstream". The library-focused supply chain is primarily built for the COUNTER standard, which presupposes an institutional view of usage. The intent of the standard is to show a librarian what and how an institution's patrons are using resources. COUNTER was not initially intended as a lens for authors and publishers to gain insight into both cross-institutional and extra-institutional access. COUNTER 5's handling of book usage, though an improvement over the COUNTER 4 standard, still leaves room for idiosyncratic platform-level decisions which makes it difficult to compare reports across providers.

   This has led to the perception amongst many of our interviewees that there is still work to be done to develop "best practices" around analyzing OA ebook usage. Over time, as the market grows, funders' and authors' expectations will grow more sophisticated. Publishers know they will need a better way to articulate success for particular works, in the competition for quality research to publish**.** This will require working with platforms and partners to better refine the current usage reporting standards and practices in place.

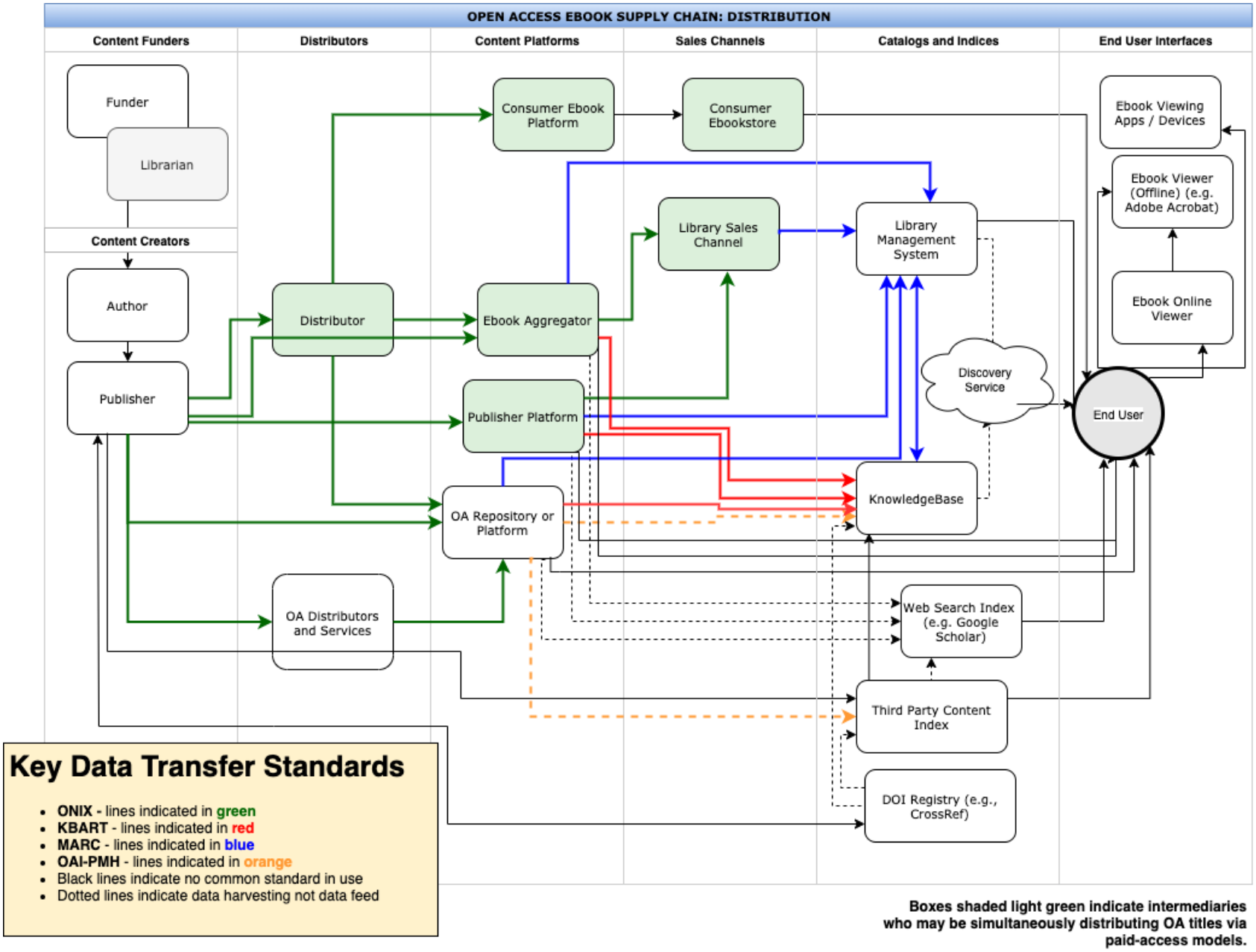
5. **The large and growing number of platforms that deliver OA books to end users creates challenges for usage reporting**

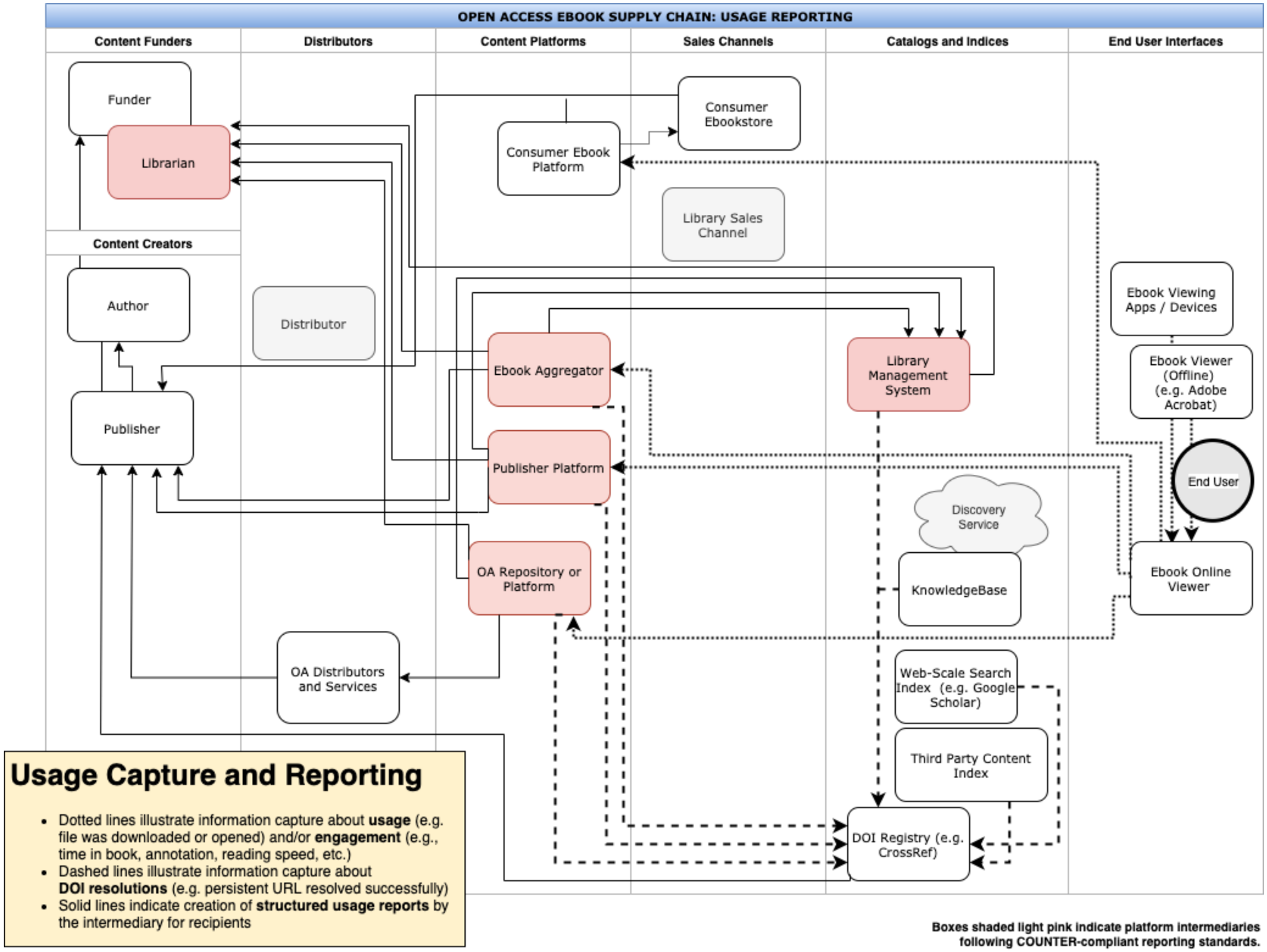
   The most striking aspect of the supply chain map overview, in our opinion, is the extent to which users have opportunities to come across the same OA title on different platforms. They might go to the publisher's platform or an ebook aggregator and come across the title there. Google or Google Scholar might direct the user to a version of the book via an OA aggregations platform, like OAPEN or Knowledge Unlatched's Open Research Library. They might encounter the book outside of the formal supply chain via a scholarly collaboration network like ResearchGate, Mendeley, or SSRN or via an institutional repository. Or they

might be on their phone, hitting Amazon's "buy now" button on a title to read in their Kindle app while they commute.

While some of these platforms will be willing to supply usage data back upstream to publishers, others will not. Paid-access books share the same problem, but the stakes are higher for OA titles. Whereas even infamously data-stingy platforms such as Amazon's Kindle do provide publishers with sales data, such data is less relevant when the book is OA. The sheer complexity of the monograph ecosystem means that a complete picture of usage is unlikely to ever be possible. But even an incomplete picture can be valuable. The many stakeholders we spoke with consistently indicated that any improvements in visibility, even if imperfect, are welcomed.

APPENDIX A: OA Books Supply Chain Map


OPEN ACCESS EBOOK SUPPLY CHAIN: DISTRIBUTION
Content Funders
Distributors
Content Platforms
Sales Channels
Catalogs and Indices
End User Interfaces
Funder
Librarian
Content Creators
Author
Publisher
Distributor
Consumer Ebook Platform
Consumer Ebookstore
Ebook Aggregator
Library Sales Channel
Library Management System
Publisher Platform
OA Repository or Platform
OA Distributors and Services
Discovery Service
KnowledgeBase
Web Search Index (e.g. Google Scholar)
Third Party Content Index
DOI Registry (e.g., CrossRef)
End User
Ebook Viewing Apps / Devices
Ebook Viewer (Offline) (e.g. Adobe Acrobat)
Ebook Online Viewer
Key Data Transfer Standards
ONIX - lines indicated in green
KBART - lines indicated in red
MARC - lines indicated in blue
OAI-PMH - lines indicated in orange
Black lines indicate no common standard in use
Dotted lines indicate data harvesting not data feed
Boxes shaded light green indicate intermediaries who may be simultaneously distributing OA titles via paid-access models.


OPEN ACCESS EBOOK SUPPLY CHAIN: USAGE REPORTING
Content Funders
Distributors
Content Platforms
Sales Channels
Catalogs and Indices
End User Interfaces
Funder
Librarian
Consumer Ebookstore
Consumer Ebook Platform
Library Sales Channel
Content Creators
Author
Distributor
Ebook Aggregator
Library Management System
Ebook Viewing Apps / Devices
Ebook Viewer (Offline) (e.g. Adobe Acrobat)
Publisher
Publisher Platform
End User
Discovery Service
KnowledgeBase
Ebook Online Viewer
OA Repository or Platform
OA Distributors and Services
Web-Scale Search Index (e.g. Google Scholar)
Third Party Content Index
DOI Registry (e.g. CrossRef)
Usage Capture and Reporting
Dotted lines illustrate information capture about usage (e.g. file was downloaded or opened) and/or engagement (e.g., time in book, annotation, reading speed, etc.)
Dashed lines illustrate information capture about DOI resolutions (e.g. persistent URL resolved successfully)
Solid lines indicate creation of structured usage reports by the intermediary for recipients
Boxes shaded light pink indicate platform intermediaries following COUNTER-compliant reporting standards.

APPENDIX B: Index of OA Books Supply Chain Stakeholders

Open Access Books Supply Chain Map - Stakeholder Index					
Category	**Stakeholder**	**Role in OA Supply Chain**	**Example(s)***	**Inputs**	**Output**
Content Funders	**Funder**	Organization contributing funds to the author or publisher for the OA publication of a title	European Research Council UKRI Wellcome Trust	Usage Reporting	
	Librarian	Can play a coordinating role for institutional funding for OA titles; also responsible for promoting selection and discovery of titles for patrons and monitoring institutional usage reports	Libraries or organizations contributing to pooled funding schemes, e.g. Knowledge Unlatched or TOME	Usage Reporting	
Content Creators	**Author**	Original creator of the book; can have multiple forms of authorship (e.g. a title can have a single primary author, multiple co-authors, a primary editor, chapter authors, etc.)		Usage Reporting	

Open Access Books Supply Chain Map - Stakeholder Index					
Category	**Stakeholder**	**Role in OA Supply Chain**	**Example(s)**	**Inputs**	**Outputs**
Content Creators (cont'd)	**Publisher**	Entities who produce the book and initiate its distribution via the OA books supply chain; under this category there are business ranging from large to small; commercial, university press, and other not-for-profit; fully OA to hybrid (e.g. publishers of both paid-access and OA titles)	Springer Nature Taylor & Francis Oxford University Press DeGruyter InTechOpen Open Book Publishers Brill	Usage Reporting	ONIX Content Files
Distributors	**Distributor**	Responsible for packaging and normalizing metadata and content files from many publishers for distribution to various downstream B2C and B2B retailers; "Traditional" distributors are those with a primary focus on paid-access business models	Longleaf Distribution Services HFS Ingram Academic	ONIX Content Files	ONIX Content Files
	OA Distributors and Services	Responsible for packaging and normalizing metadata and content files specifically in service for OA content and initiatives	Knowledge Unlatched Unglue.it	ONIX Content Files	ONIX Content Files

Open Access Books Supply Chain Map - Stakeholder Index					
Category	**Stakeholder**	**Role in OA Supply Chain**	**Example(s)**	**Inputs**	**Outputs**
Content Platform	**Publisher Platform**	Content delivery platform dedicated to the content of a specific publisher; any paid-access content on the platform is typically sold on a B2B basis to libraries and institutions. OA content is frequently hosted alongside these paid-access titles.	SpringerLink Taylor & Francis Online Cambridge Books Online MIT Press Direct BrillOnline	ONIX Content Files Other formats (non-standard)	KBART MARC Usage Reporting Other formats (non-standard)
	Ebook Aggregator	Content delivery platform which aggregates titles across multiple publishers. Typically serve B2B markets (institutions and libraries) although some aggregators host OA-specific collections on their platform. We include in this category publisher platforms which incorporate content from other partners.	EBSCO eBooks ProQuest Ebook Central JSTOR Project MUSE University Press Scholarship Online DeGruyter.com Fulcrum	ONIX Content Files	KBART MARC Usage Reporting Other formats (non-standard)
	Consumer Ebook Platform	Content delivery platform which provides access to titles after purchase or download by individuals. OA titles may appear here after "purchase" by the individual account owner	Kindle Kobo Nook Google Play	ONIX	Other formats (non-standard)

Open Access Books Supply Chain Map - Stakeholder Index					
Category	**Stakeholder**	**Role in OA Supply Chain**	**Example(s)**	**Inputs**	**Outputs**
Content Platforms (cont'd)	**OA Repository or Platform**	Content delivery platform entirely dedicated to the delivery of free or OA content; there is no legacy B2B or B2C transactional functionality required to access titles.	OAPEN Library Internet Archive Project Gutenberg Hathi Trust Ubiquity Manifold Institutional repositories	ONIX	OAI-PMH MARC KBART Other formats (non-standard)
Sales Channels	**Library Sales Channels**	Booksellers who conduct sales and offer post-sales support services for the institutional / library market. Support for OA titles depends on the provider.	GOBI Library Solutions ProQuest OASIS Askews and Holts Casalini Libri	ONIX Other formats (non-standard)	MARC
	Consumer Ebookstore	Booksellers serving the retail markets, from whom individuals purchase access to titles on a consumer ebook platform.	Amazon.com Barnes & Noble Kobo.com Google Play	ONIX	none
Catalogs and Indices	**Library Management System**	Umbrella term for systems and technology products which allow librarians to manage local holdings and catalogs. Covers both the Integrated Library System (ILS) and any associated modules, as well as next-gen Library Services Platforms (LSP) which include a broader feature set.	ExLibris Alma OCLC WorldShare Management Service FOLIO Sierra Koha	MARC Usage reporting	Usage reporting

Open Access Books Supply Chain Map - Stakeholder Index					
Category	**Stakeholder**	**Role in OA Supply Chain**	**Example(s)**	**Inputs**	**Outputs**
Catalogs and Indices (cont'd)	**KnowledgeBase**	Extensive catalog of content metadata shared across libraries and institutions (we include under this umbrella any associated union catalogs). KnowledgeBases combine metadata from multiple sources and can provide in structured format to library management systems, as well as underpin a library's Discovery Service	OCLC WorldCat Alma Central KnowledgeBase (CKB) EBSCO Integrated Knowledge Base	MARC KBART ONIX OAI-PMH Other formats (non-standard)	MARC KBART
	Third-Party Content Index	Specialized product for metadata curation and discovery; can include topic-specific A&I databases or other curated directories. Developed to provide structured and organizational metadata and discovery.	Directory of Open Access Books (DOAB) Clarivate Analytics Book Citation Index MLA International Bibliography Unpaywall Digital Science Dimensions	Other formats (non-standard) OAI-PMH	Other format (non-standard)
	DOI Registries	Issuing body for unique Digital Object Identifiers (DOI), which can be applied at the book or chapter level. The primary DOI registry for OA books, Crossref, tracks DOI resolutions and reports back to publishers	Crossref	Other format (non-standard)	Usage Reporting (DOI Resolutions)

Open Access Books Supply Chain Map - Stakeholder Index					
Category	**Stakeholder**	**Role in OA Supply Chain**	**Example(s)**	**Inputs**	**Outputs**
Catalogs and Indices (cont'd)	**Web-Scale Search Engines**	Search engines developed to index and retrieve content (including OA books) from across the web; important discovery mechanism for users especially outside of a library context	Google Scholar Google	OAI-PMH Other format (non-standard)	Other format (non-standard)
End User Platforms	**Online Ebook Viewer**	Software embedded in a webpage or device which allows the user to access and read an ebook	Embedded web viewer on ebook aggregator websites	Content files	Usage data
	Ebook Viewing Apps / Devices	Dedicated device or app operated by a platform provider; allows information about reading behavior to be trackable by the content platform.	Kindle Device Kindle App iBooks Kobo Device	Content Files	Usage data
	Offline Ebook Viewer	Third-party software which allows a user to access and read an ebook, but which does not provide any information back to the hosting platform	Adobe Acrobat Bluefire Reader	Content files	

**Note: Example organizations or entities are provided for illustrative purposes only; examples given are not meant to be exhaustive. While we endeavored to speak with as many of the categorized stakeholders as possible not every stakeholder listed here was included in research interviews.*

APPENDIX C: Stakeholder Semi-Structured Interview Guide

Overview

Goal duration: 45 minutes

Interview format: Semi-structured; interviewer will touch on specific questions but not read from a script. This format provides consistency in coverage of important topics but also enables a degree of flexibility for deeper exploration of productive avenues of discussion.

Introduction

Introductory remarks to be shared with interviewee at the start of the interview.

- We are with Clarke & Esposito, a consulting firm that provides business strategy services to publishers, mostly not-for-profit societies and associations and university presses.
- We are currently working on a project with the Book Industry Study Group, Curtin University, Educopia Institute, University of Michigan, University of North Texas, and others to map the supply chain for Open Access monographs. This is part of a Mellon-funded initiative to develop an OA Data Trust to allow different stakeholders to share and analyze data on OA monograph works.
- We've scheduled 45 minutes for this call – will that work for you today?
- We have a list of questions, but first we wanted to give you the opportunity to ask us any questions, or to offer anything that is top of mind for you right off the bat.

NOTE: Interviewees will be asked to complete a brief questionnaire before the interview (see Appendix D)

Interview Template

Italics indicates instruction for interviewer.

NOTE: Questions were adapted for each interviewee based on primary role and organization.

Name	
Affiliation	
Role	
Add'l context	
Date of interview	
GENERAL BOOKS SUPPLY CHAIN BACKGROUND	
1. Describe your role in the general supply chain for books. What providers and roles sit immediately upstream from you? What kind of stakeholders or 'customers' do you serve?	
•	
(NOTE: Interviewees will have been provided a pre-interview questionnaire to capture more info on metadata standards support. Review prior to the interview.)	
2. *[[Explore use of particular standards as needed, in particular proprietary formats.]]*	
•	
3. When you send and receive information from others in the supply chain, do you augment / add metadata to these feeds in any way? Are there any metadata elements that tend to get dropped or lost?	
•	
4. How do you keep track of versions of books in multiple formats (e.g., print with e versions, EPUB with PDF)? Do you make efforts to reconcile these multiple formats in any way?	
•	
5. What data points are most important to you to measure performance or success of a particular book? Do you share that information in either direction in the supply chain?	
•	
6. How are you capturing discovery and usage of the books you *[[produce / sell / catalog / etc]]*? Does how you capture usage differ by format (PDF vs. EPUB, chapter vs. whole book)?	

7. Where do you get this information from? What information are you not able to get — what's missing?

-

8. Are there any standards and formats that you anticipate will become a greater priority in the coming years (either for you or for others in the supply chain)?

-

OPEN ACCESS MONOGRAPHS [[adapt as needed based on if interviewee is fully OA or hybrid]]

9. How does your role in the supply chain change for Open Access monographs? Do you handle Open Access monographs today?

-

    a. How do you distinguish Open Access monographs in your system? [[Note – this question should ideally have been addressed in the pre-interview questionnaire, but is included for discussion if needed]]

-

    b. Are you able to distinguish your Open Access books in the information you provide to other stakeholders? Are there any standards and formats where this is particularly difficult?

-

10. Do you measure performance differently for Open Access books than traditionally published books? How important is usage as a success metric for you?

-

11. Are there examples where a view into the usage of Open Access monographs might provide a benefit to other parts of your business / to other activities?

-

12. What is your anticipated engagement with Open Access monographs generally in the next five years?

-

13. Is there anything you had hoped we had asked you about that we haven't touched on yet?

-

APPENDIX D: Pre-Interview Questionnaire

Clarke & Esposito is working on a project to map the supply chain for open access monographs. This is part of a Mellon-funded initiative to develop an OA Data Trust to allow different stakeholders to share and analyze data on OA monograph (you can learn more about the project here: educopia.org/data_trust).

As part of this project, we'd like to understand what traditional and emerging metadata formats and best practices are currently being used and supported in the marketplace. To help us do this, please take a moment to answer the brief (>5-minute) questionnaire below. If you are unable to answer these questions yourself, please feel free to forward the questionnaire to the appropriate colleague.

1.How would you best describe your organization's role in the book and ebook supply chain? (You may choose more than one)
Publisher
Distributor / Wholesaler
Aggregator
Platform Provider
Retailer
Purchaser / Consumer
Service Provider

2.Which of the following metadata standards does your organization use to SHARE information? In other words, which of these formats are your systems able to EXPORT? (You may choose more than one)
MARC
ONIX
KBART
COUNTER 5 Usage Reporting
BIBFRAME
In-House Format (Unique to my Organization)
OPDS

3.Which of the following metadata standards do you use to IMPORT / INGEST information into your own systems? (You may choose more than one)
MARC
ONIX
KBART
COUNTER 5 Usage Reporting
BIBFRAME
In-House Format (Unique to my Organization)

4.What is the primary unique identifier you use to distinguish ebook titles in your systems?
ISBN
DOI
OCLC Control Number
Library of Congress Control Number
In-House Book Identifier (Unique to My Organization)

5.What other unique ebook identifiers does your system collect or support? (You may choose more than one)
ISBN
DOI
OCLC Control Number
Library of Congress Control Number
In-House Book Identifier (Unique to My Organization)

6.Which book formats do you support? (You may choose more than one)
EPUB / EPUB3
PDF
MOBI
Print
HTML5
XML / BITS

7.(Optional) How does your organization flag or differentiate Open Access books in your systems?

8.Name / Organization