Documenting the Supply Chain for Open Journals and Data

19 August 2024

Laura Ricci, Senior Consultant

Michael Clarke, Managing Partner

Clarke & Esposito

CLARKE & ESPOSITO

A C&E Report

For more information, contact:

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

Clarke & Esposito | www.ce-strategy.com

About this Report

This research report was made possible by NSF Grant # 2335827.

About Clarke & Esposito

C&E is a consulting firm that helps forward-thinking organizations grow and diversify revenue, better engage customers and communities, and operate more efficiently. C&E focuses on the professional and scholarly information market.

Learn more at www.ce-strategy.com.

Contents

1 Introduction

In November 2022, the Association of Research Libraries surveyed the collaboration priorities for US-based research data organizations (Kennedy and Hudson Vitale 2022)*. Strategic priorities for action included: public-private partnerships, creating services to support compliance, coordinated research infrastructure, and the sharing of sensitive data.*

Two years prior, in May 2020, the authors of this report documented the distribution and usage supply chains related to OA monographs, illuminating the complex business relationships and data flows that support the evaluation of such scholarly communications (Ricci & Clarke, 2021). This research has since informed the Open Access eBook Usage Data Trust's development of a minimum viable data governance "data space" for controlled multi-party usage data exchange across public and private scholarly communications stakeholders.[1] *Organizations that generate and/or innovate with book usage data are at present exploring sovereign data federation within a trusted "Data trust" membership network of organizations to expedite data exchange across public and commercial data stores while reducing duplication, improving data quality, and increasing economies of scale and trust.*

In 2023, the US National Science Foundation funded a workshop, "Exploring National Infrastructure for Public Access and Impact Reporting." (FAIN 2315721) Thought-leaders from publishers, repositories, scholarly infrastructures, and funding agencies collaboratively identified gaps and opportunities given the current state of impact reporting and analytics across the diverse research outputs, such as books, journal articles, and datasets.[2] *Stakeholders identified ways to improve the efficacy, findability, access, interoperability, and reusability of research impact metrics while considering opportunities for economies of scale in institutional reporting. Discussions surfaced a need to better understand the state of the usage and impact data supply chains. What roles were played by different types of actors in different "market channels"? How dissimilar were book, data, and article level usage data flows? The promise of object-agnostic infrastructure surfaced as participants identified research avenues to advance the FAIRness of usage and impact metrics about publicly accessed digital content. (Kemp, Watkinson, Drummond, 2024)*

[1] The OAEBUDT is developing a data commons network for controlled multi-party public and private scholarly communications data exchange that adheres to the emerging common European data spaces protocol. More information about this protocol is available from the International Data Spaces Association.

[2] (Drummond, C. (2023). Proceedings of the Workshop Exploring National Infrastructure for Public Access Usage and Impact Reporting. Zenodo. https://doi.org/10.5281/zenodo.8335916)

Questions abound as scholarly monograph stakeholders learn from piloting a minimum viable data space. Could, or should, a data-commons for scholarship usage and impact reporting be created for each type of scholastic output? Should such data sharing governance cyberinfrastructure be global in nature, or federated and national in focus given the diversity of stakeholders and data regulations involved?

Made possible through a grant from the US National Science Foundation (FAIN 2335827), this report is a first step towards answering such questions. Laura Ricci and Michael Clarke complement their prior monograph distribution supply chains by highlighting herein the ecosystems supporting publicly accessible article and data distribution and evaluation. This report addresses questions such as:

- *How are journal distribution practices changing, particularly with the growth of new open access business models (i.e., Gold OA, transformative agreements, Subscribe to Open, and Diamond OA)?*
- *What are the metrics that matter in the evaluation of these business models?*
- *How are open datasets being shared, distributed, and evaluated within the research community?*
- *How do the journals and open data supply chains differ from each other, and from that of open books?*
- *Do these differences support a similar or a different approach to managing usage and impact data? What infrastructure is needed to support this approach?*

With their comparative analysis of stakeholder relationships, opportunities and gaps across research distribution supply chains, Ricci and Clarke identify more than just information flows. They surface pain points and potential value propositions while identifying ways to strengthen the foundations of research evaluation and research information management systems.

After describing their methodology and providing an overview of the usage and impact reporting landscape, Ricci and Clarke outline key stakeholders, data intermediaries, standards and infrastructures related to journal article and dataset usage and impact analysis. Gaps and opportunities for distribution supply chains are identified alongside visualizations of the usage and impact data flows for open data and journals. In sum, we hope this report provides valuable information to inform future scholarly communications infrastructure development and investment.

Christina Drummond, Principal Investigator
University of North Texas

2 Methodology

Building on our previous analysis of the supply chains for OA monographs, C&E developed a research plan in consultation with the project Principal Investigator Christina Drummond (University of North Texas) to identify the following:

- Key stakeholders involved in funding, creation, distribution, and evaluation of journal articles and open data,
- Connections and relationships between stakeholders to facilitate the flow of information about each type of research output,
- Key metadata elements used to describe articles / datasets and in provision of usage and impact information, and
- Important standards and tools developed to facilitate this information flow.

Research interviews solicited a range of stakeholder perspectives from subject-matter experts embedded throughout journal and data ecosystems. Interviews aimed to surface relationships among different stakeholder groups and the perceived gaps and opportunities for each supply chain. The interview format was semi-structured in that C&E touched on specific topics from a list of questions but did not read from a script, allowing the conversation to progress naturally and be adapted to the specific perspective of the interviewee. This approach provided consistency in coverage of the most important issues, but also enabled a degree of flexibility for deeper exploration of productive avenues of discussion.

Twenty-five senior leaders were interviewed from 21 organizations across 23 interviews. Positions held by the interviewees spanned leadership and operational units. Within the interview pool were: executive directors (4), program and center directors (2), product managers or vice presidents of product (4), executives of various levels with oversight over content or publisher affairs (4), a startup founder, a managing director, a general manager, a chief strategy officer, a data intelligence unit lead and an editorial director.

Interviewees were employed at the following institutions at the time of their interview. Notably, some of the institutions represented focus on a single research output type, while others work with multiple types of outputs, as annotated below.

Organizations Participating in Research Interviews	
Organization	**Organization Focus* (Journals, Books, Data, Infrastructure)**
Atypon	Journals, Books, Infrastructure
Berghahn Journals	Journals, Books
ChronosHub	Journals
DataSeer	Data
DeGruyter / Ubiquity Press	Journals, Books
Digital Science	Journals, Data, Books
Dryad	Data
EBSCO	Journals, Books
JISC	Journals, Data
JSTOR	Journals, Books
DataCite / Make Data Count	Data, Infrastructure
OA Switchboard	Journals, Books
OpenAIRE	Infrastructure
COUNTER Metrics	Infrastructure
Project MUSE	Journals, Books
Protocols.io	Data
ResearchGate	Journals
Silverchair	Journals, Books, Infrastructure
Springer Nature	Journals, Books
Taylor & Francis	Journals, Books
University of California Curation Center (UC3)	Data, Infrastructure
*Individuals interviewed for this project were those identified as most knowledgeable about open journals and open data content and metadata distribution and evaluation workflows; these individuals were not required to have expertise in all organizational activities or in supply chains outside of open journals and open data.	

Interview questions (**Appendix C**) addressed the value derived from various stakeholder interactions, metadata and other information flows, supply chain gaps, pain points related to incomplete or missing information, and recommendations for improvements and remediation of problems.

Interview findings informed the creation of schematic maps of the distribution and evaluation supply chain for open journal articles and open data (**Appendix A**). This report provides explanatory text for each map, describing in detail the key gaps, challenges, and opportunities expressed by interviewees. The maps are also accompanied in this report by an index of stakeholders within the open journals and open data supply chains (**Appendix B**).

2.1 DEFINITION AND TREATMENT OF TERMS

Supply chains are the interconnected networks of organizations, people, activities, information, and resources involved in the production and distribution of a product or service from supplier to end customers and users.[3] The business model for that product or service defines the relationship between supplier and customer, and so directly influences the development and evolution of that supply chain. In the case of journals and data, stakeholders within these supply chains have developed specific terminology for the prevailing business model(s), which we start by defining below.

Open Access (OA). Open access is defined as when a research output is made freely accessible online (at no cost to the reader) with no restrictions on access and no (or limited) restrictions on *reuse*, other than requiring acknowledgment of the original author.

Public Access. Public access is defined as when a research output is freely accessible to the public, but may carry copyright restrictions or a restrictive reuse license. For example, a work may be freely accessed by the public, often through an open repository such as PubMedCentral, while still maintaining a publisher copyright..

Open and Public Access Business Models. In journal publishing, open or public access can be achieved via several different business models, designated by a color system.

- ***Gold Open Access.*** In this model, an article is made *open access* (i.e., free to the reader under an open license) upon publication in the journal itself, typically after payment of an article processing charge (APC).
    - Commonly, the APC is paid for by the author directly (whether from money allocated from a grant, an institutional allowance, or personal funds).
    - In an increasing number of cases, APCs may be paid by an institution for authors affiliated with that institution. One important example of this type of institutional deal is ***Transformative Agreements*****,** also known as "read and publish" or "publish and read" deals. Transformative Agreements bundle subscriptions to journals with article processing charges (APCs) for researchers affiliated with the institution or consortia to publish on an OA basis (in those titles).

[3] Note that, consistent with C&E's previous analysis of the open monographs supply chain, we are focused on the publication processes for these three research outputs. In the terminology of the National Institute of Standards and Technology (NIST) Research Data Assessment Framework (RDaF), this corresponds with the "Share / Use / ReUse" stage of the research lifecycle. For more see: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/1500-18/NIST.SP.1500-18r2.html

- ***Diamond Open Access.*** These are journals in which all articles are free to read and reuse, with no author (or institution) fee for publication. Expenses are often covered by grants or host institutions.
    - A further subset of the Diamond Open Access model is the ***Subscribe to Open (S2O*)** model. Libraries participate in the model by paying the journal (or journals) an annual fee. Each year the journal achieves its target level of library participation, all articles published in the journal that year are made OA in perpetuity. For any year the journal fails to hit the participation target, that year's articles are published under the subscription model, with only paid subscribers able to access the content.
- ***Green Open Access.*** Green OA refers to making a version of an article available in a public repository. This model typically results in a freely available version of the research output that is not the ***version of record (VoR)*** – *i.e.*, the final formatted, typeset article published in the journal. Rather, a freely available, Green OA work is often an earlier, unformatted version of the work such as the ***Author Accepted Manuscript (AAM)*** produced once a publisher accepts a work after peer review.
    - America's federal research funding agencies mandate public access to research outputs and the unformatted AAM is acceptable to fulfill the requirement; an open access final publication version is not required). In this report, we note how the supply chain enables ***public access*** via the Green OA route.[4]
- ***Bronze Open Access.*** Bronze OA refers to articles voluntarily made free to read by the publisher in the journal, without an explicit open license. This is often temporary and intended as a marketing activity, to bring attention to the journal or a specific high-profile order.[5]

Note that the definitions above are specific to journals publishing; the same terminology is not typically used for open monographs or open data.[6]

[4] Despite being generally classified as an "Open Access" model, Green OA does not necessarily result in the article being made OA in the full sense of the word. For example, an article can be made Green OA without a permissive license or without copyright restrictions. We use this specific terminology because it is industry standard and comports with how the terms are used by stakeholders in the supply chain.

[5] As with Green OA, Bronze OA typically does not include a permissive license or and the paper may still fall under copyright restrictions.

[6] For a more detailed treatment of open monograph business models, see Ricci, L. "Every Book, Everywhere, All at Once: Exploring the Multiverse of eBook Open Access Models". Against the Grain. Vol 34, July 2023 Special Report. https://issuu.com/against-the-grain/docs/atg_july22_special_report/s/16467256

In this project we took an inclusive view of "open" journal articles, with the understanding that many stakeholder interactions and metadata standards for research outputs are the same regardless of access model. This approach was justified as many of the information flows described in this report explicitly distinguish between these different forms and flavors of access, highlighting how each must co-exist within the same supply chain.

Open data. In the case of open data, we have defined open data as that which meets the formal definition of open access given above (i.e., available to access at no cost, under a permissive license which allows re-use and redistribution). For open data, the requirement to give attribution to the original author may be optional (for example, under a CC0 "no rights reserved" license[7]).

2.2 EXCLUSIONS FROM THE ANALYSIS

Many journals, particularly those in medical disciplines, rely on digital advertising as a revenue stream. This digital advertising relies on a cost per impression (web-page load) or per click (reader engagement). While article page view and engagement related web analytics are important to these journals, advertising click-thru streams, cookie trackers, and other digital signatures of advertising engagement involve distinct data supply chains adjacent to research distribution and evaluation supply chains. Given our focus on research use and impact, advertising supply chains are outside of this report's scope.

3 Landscape Overview

Previous research on the supply chain for open access monographs found that the interactions between stakeholders in that supply chain is heavily influenced by the monograph's legacy sales model. As an open monograph does not generate "sales", there are opportunities and challenges for stakeholders to adopt *usage* as a new metric for success. However, the dispersed nature of distribution in the books supply chain means usage often happens across many platforms, with currently limited central visibility.

To fully compare open journals and open data supply chains with that of open books, **it is likewise necessary to understand how the business and distribution models have informed the development of each supply chain**, where there are common problems, and where each is unique.

[7] https://creativecommons.org/public-domain/cc0/

3.1 JOURNAL ARTICLES

Journal articles are the primary research output for Science, Technology, and Engineering (STEM) fields and certain social sciences, which differs from the scholarly monograph/book being the primary research output in certain social sciences and humanities. Researchers in these disciplines publish articles to describe their research results for their peers, promotion and tenure committees, and the financial sponsors funding their efforts. Career progression and employment is heavily influenced by the quantity, reach, prestige and impact of the articles they publish.

Unlike book publication, where books are discrete research outputs published one at a time and often sold individually, journal publication relies on a persistent and overarching journal brand.[8] Journal articles are published together within issues and volumes, adding to the journal's publication history which can span many years. (The oldest journal is over 350 years old.) Historically, readers or their institutions would gain access to the newest articles by paying an annual subscription to that journal – securing in advance the ability to access and read all articles published in the forthcoming year. Even when the primary mode of access to journals shifted from print to digital in the 1990s and 2000s, the subscription model persisted. Institutions and individuals pay for access on a recurring basis to the journal, while authors seek to publish their articles under the umbrella brand of the journals that best fit their desired audience and perceived level of significance of their work.

The growth of open access models, specifically the Gold OA author-pays model, represents a "break" from the journal subscription model. Two major differences Gold OA introduced into the supply chain are:

- **Authors become the customer.** In the Gold OA author-pays model, authors (not readers) are the payer or "customer", and must secure the funds to pay publishers (the payee) to make their article publications freely available without requiring a subscription.[9]
- **Revenue is primarily transactional.** The Gold OA model is transactional (author-pays revenue is generated just once for each article published, at the time of publication), whereas subscriptions are a recurring revenue model (the same subscribers continue to pay every year).

[8] The one exception among books products are books series, which are in many ways a special case and behave much in the way that journals are described here.

[9] Even in the case of institution-funded Gold OA deals, such as Transformative Agreements, the author's choice of publication venue determines where each payment is directed.

Yet despite the differences in stakeholder relationships between the subscription model and author-pays open access (different stakeholders act as customers, different value proposition to that customer), many journals and journal publishers support both business models simultaneously. This means the supply chain must support multiple relationships between stakeholders.

Elsewhere on the spectrum between author-pays OA and the subscription model, other forms of open and public access are increasingly supported as well. The most common forms of public access (the Green OA model, in which a version of the article is made available to read at no cost in a publicly accessible repository) does not represent a new business model per se, but instead can be understood as subsidized by another business model – Green OA is sustainable for a subscription journal as long as that journal is able to continue to sell subscriptions. The volatility of purely transactional author-pays models is also why many Gold OA publishers have introduced recurring institutional models (such as PLOS's Community Action Publishing) and large publishers of hybrid journals are increasingly striking Transformative Agreements – each mimics a subscription model to make spending more predictable to the payer and payee.

There are also aspects of the supply chain which have *not* changed in response to new business models.

- A journal is designed to both generate revenue and require investment with no pre-determined end date. It does not have a "frontlist" or "backlist" period (like books do).
- The journal brand and performance history persists in both the distribution and the evaluation of research, including backward-looking evaluation (for author-pays: "am I satisfied with the publication I paid for?"; for subscriptions: "did we use this subscription enough to justify the cost?") and in forward-looking evaluation (for author-pays: "should I pay for publication in that journal again?"; for subscriptions: "does this fit into my institution's budget and research priorities?").

The result of these dynamics is a **cyclical** supply chain for journals, including a well-developed infrastructure designed to support recurring decisions about the journal. This is a stark contrast with books and open data, neither of which rely on recurring models to the same degree.

3.2 OPEN DATA

Publishing research data often is tied to, if not required by, the publication of another primary research output. Making research data available for peer inspection and verification is important to research integrity and reproducibility, as is evidenced by longstanding disciplinary expectations transforming into the FAIR (Findable, Accessible, Interoperable, and Reusable) data movement to increase dataset access and use. Often, open data is published in conjunction with

a publication, not as a standalone research output for its own sake. The process of making data FAIR is itself dependent on ensuring the linkages between data outputs and their publications.

Online data sharing greatly enhances scholars' ability to make data openly available to the public, as discipline-focused online databases continue to arise across fields for researchers to contribute their observations and findings. After being developed with significant federal support, the staff caretakers and technical infrastructures to support these data repositories are often hosted by research institutions as part of their contribution to the field (sometimes collaboratively via groups of research funders) or have spun out as nonprofits. Many such databases serve a particular discipline, research area, or use case (for example, the Protein Data Bank [PDB], in existence since 1971, is dedicated solely to the three-dimensional structures of proteins, nucleic acids, and complex assemblies; the NASA Science Data Portal provides access to different repositories for astrophysics, planetary science, heliophysics, earth science, etc.). The federal agencies developing such data centers often play a role in overseeing and approving data repositories for their related disciplines.

Strong cultural motivations to share data are increasingly becoming formal requirements to make data available as part of the publication process and reinforcing stronger data sharing practice. Many funders around the world (including most of the major US funding agencies) require researchers to provide a data management plan (DMP) as a condition of grant funding to articulate their sustainable plans for access and data protection or embargo (as needed). Other funders (particularly in the European Union) also require open data publication, although there is variability in the degree to which compliance is monitored. To comply with the 2022 OSTP Nelson Memo, federal US agencies are enacting new data sharing requirements of their funded authors in 2026.[10]

Similarly, journal publishers are strengthening their research data sharing requirements and incorporating more forms of data sharing into the publication process.[11] Previously, many publishers would simply request or require an author's acknowledgment that they would make full data available upon request. However journals are increasingly primary drivers of data deposits into their field's disciplinary repositories. For a growing number of journals, publication is now conditional on making data available via an appropriate data repository (with measures

[10] The implementation details of this aspect of the mandate are still in the process of being defined as of July 2024

[11] The National Institute of Standards and Technology (NIST) Research Data Assessment Framework (RDaF) lists seven modes of dissemination, including publication of a traditional journal article and supplementary files, which have always been part of a publisher's remit. Increasingly, publishers are encouraging additional forms of dissemination, going beyond allowing authors to make data available on request toward ensuring materials are available via standalone web pages. See: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/1500-18/NIST.SP.1500-18r2.html

applied if data is sensitive or confidential). As an example: for several decades, most biomedical journals have required deposit of any DNA sequence included in a research publication in GenBank, and the article must show the database's accompanying accession number for the sequence. For the journal publisher, enforcing a data publication requirement strengthens their ability to detect and deter fraudulent research, and thereby creates more trust in their publication brand. Infrastructure to appropriately link to and cite data used in research is increasingly necessary (although, as we will discuss, not always used robustly).

Importantly, although access to a journal article may be enabled through an institutional subscription, access to the supporting research data is almost universally expected to be available for free.[12] However, the repository landscape is not devoid of commerce – there are repositories that charge fees for deposit; others that license "white label" versions of their repository to institutions for in-house branding; and still others that charge for access to data visualization tools or data mining of the objects within the repository.

As the research data itself is not typically monetized, the distribution supply chain differs from books and journals. The primary "buyers" (in supply chain terms) are often those choosing to invest for the purpose of keeping the supply chain going (as in, making sure scholars continue to have a place to share and re-use open data). Rather than being sustained through *purchases*, like books and journals, the open data supply chain is more often sustained through *funding*. This directly influences the evaluation decisions the supply chain must also support.

3.3 LANDSCAPE OVERVIEW SUMMARY AND COMPARISON

Table 1 provides a summary overview of the comparison between journals, books, and data and these factors which have influenced the development of each supply chain.

[12] See also the NIST RDaF list of seven modes of access:
https://nvlpubs.nist.gov/nistpubs/SpecialPublications/1500-18/NIST.SP.1500-18r2.html

TABLE 1. Comparison of Factors Influencing Books, Journal, Data Supply Chains*			
	Books	**Journals**	**Data**
Business Model	• Books evolved from transactional business models (recurring revenue is limited)	• Journals evolved from recurring business models (subscriptions), though author-pays Gold OA is growing and is a transactional revenue stream)	• Data sharing is typically not monetized, but rather is supported by funding by governmental organizations and research institutes
Brand	• Books typically published as discrete titles, typically no over-arching brand	• Articles are published under the brand of their parent journal	• In many cases, data is made available in association with another research output such as a journal article. However data may also be shared independently.
Distribution	• Books often distributed across multiple platforms simultaneously in order to maximize distribution	• Under subscription model, journals typically hosted on one primary platform, usually managed by the publisher, with access through secondary platforms (like aggregators) following an embargo. • Open access articles are often published under a permissive license (like open access content), in which case further and simultaneous distribution of articles is allowed . • To broaden readership, publishers are also increasingly enabling formal syndication across platforms for both OA and subscription articles	• Data is not formally distributed, though an author may make data available in multiple repositories

TABLE 1. Comparison of Factors Influencing Books, Journal, Data Supply Chains*			
	Books	**Journals**	**Data**
Unit of Distribution	• Most books are distributed as whole titles, with limited distribution at the chapter level • Chapters are often (but not always) the same business model as the parent text	• Journal distribution supply chains support both journal-level and article-level flow of information • Many journals are "hybrid" meaning articles within the same journal can be published under different business models	• Metadata capture is at the dataset level, though each dataset may contain more discrete data items • Data items typically do not have their own unique metadata elements / identifiers
Individual vs. Institutional Sales	• Many scholarly monographs are available through B2C channels (such as Amazon) in addition to institutional channels (such as JSTOR, EBSCO)	• Journal subscriptions may be sold to individuals (or in the case of society journals, available to individual society members as a member benefit) but the majority of journals are reliant on the institutional subscription market as a primary source of revenue	• Data is typically free to access for the end user. Institutional support is in the form of maintaining the data repository where data is stored and/or funding other data infrastructure.
Versioning	• Books are typically published under a single Version of Record (though sometimes with multiple editions) • Distribution of a non-peer reviewed or preliminary version of a book manuscript is not common	• The final, formatted version after peer review is designated by the publisher as the Version of Record (VoR) • Previous versions of the article may also exist within the supply chain, such as the preprint (initial version of the article before peer review) and the Author Accepted Manuscript (AAM)	• Data outputs typically do not designate a single formal "Version of Record" • Depending on the license applied, data outputs may be re-used or adapted / forked by other researchers

TABLE 1. Comparison of Factors Influencing Books, Journal, Data Supply Chains*			
	Books	**Journals**	**Data**
Evaluation	• Book sales are a primary metric of success for publishers, authors, distributors within the books supply chain • For open books, sales is not a pertinent concept so many stakeholders are reorienting towards usage as a metric of success • Evaluation metrics such as citations have a weaker cultural hold in monograph publishing	• Usage information is an important metric for an institutional subscription purchase decisions (specifically, usage within that institution) • Citations are a commonly used quality metric for evaluation of journals and articles (and authors) • Initiatives such as DORA encourage additional metrics be used for evaluating quality, such as global usage and contribution to policy	• Usage of data repositories and platforms can inform further investment in that repository and justify continued funding into that platform • Initiatives to capture use and re-use of individual datasets are nascent, as are initiatives to encourage more consistent citation of datasets in scholarly research
*As the supply chain for open books and open journals are overlapping and in many cases not distinct from the supply chain for paid access books and journals, this table includes comparisons of legacy sales infrastructure to illustrate how this leads to differences in how each supply chain operates.			

4 Key Stakeholders and Intermediaries

This section provides an overview of the key stakeholders and intermediaries that play a role in the journals and data distribution supply chains. In this report, we separate out various stakeholder types not because they are distinct groups of organizations, but because we hope to illustrate the roles and motivations in place for the production and evaluation of research. A map of each supply chain, illustrating the interactions between stakeholders for different research outputs, are included in this report's **Appendix A**.

It is important to point out that organizations can play multiple stakeholder roles; for example, an academic institution may act as both a funder (by supporting OA publication for researchers in their employ) and a publisher (via support for a university press). Commercial organizations like Elsevier, Wolters Kluwer, and Springer Nature also span multiple stakeholder roles, simultaneously acting as publishers, content platforms, creators of discovery indexes, data repositories, and analytics providers. There are also multiple subcategories and subrelationships, such as the professional society that contracts with a larger publisher to provide publishing services on their behalf.

4.1 CONTENT FUNDERS

Content funders are organizations that provide financial (and other) support for conducting research and publishing the outputs of that research. These stakeholders are instrumental in setting publication policy and evaluating researcher performance or "return on (research) investment" (ROI) based on the resultant research outputs. Subcategories of content funders include:

- **Research funding agencies and sponsors.** Research is ultimately enabled by the government bodies and private organizations that provide funding and support. Such "funders" can influence the publication model of the finished work, by making funding conditional on open and/or public access to resultant research outputs within a certain period of time, or under specific copyright provisions. Information about the impact and importance of sponsored projects affiliated with a given author or institution is sought by funders to inform their future funding decisions.
    - This type of stakeholder can at times be further distinguished between organizations that fund the *research* and those that fund the *publication* of the research. Stakeholder interviews surfaced a desire to clearly indicate the difference between research and publication funders within article metadata.
- **Institutions.** Institutions employ researchers and provide research facilities; their employees are both creators and consumers of research outputs. Defined broadly this

group includes: academic institutions, research institutes, government labs, commercial organizations, nonprofits, etc. Institutions, like funders, may also have regulatory or policy mandates that influence the open/public availability of the research outputs they support (including both journal articles and datasets). Following publication, institutions rely on information about research impact and importance for scholar performance evaluation as discussed above. When researchers collaborate across different institutions, many institutions may become connected to the resulting article(s) or dataset(s), making it important that the supply chain support this many-to-one relationship between institutional provenance and research output.

- **Libraries**. While libraries are historically the purchasers of research publications, they are playing a growing role in promoting public and open access publication and increasing the discoverability of research outputs. Libraries may support the funding of open publication through multiple methods. They may: 1) provide article processing charge (APC) funds to researchers, 2) enter into Transformative Agreements with publishers to to dedicate a portion of their subscription spending to support OA publication by their institution's researchers, and 3) pool their funds to support collaborative open access business models such as Subscribe to Open (described in Section 2.1, "Definition and Treatment of Terms").

Content funders are important in the supply chain of any research output. However, in building comparisons between the supply chains of open journals, open data, and monographs, funders play a stronger, more active role in the journals and data supply chains. The journal article is the primary output for many resource-intensive fields that command significant funding (such as most STEM fields and to a lesser extent the social sciences). In addition, articles and data publication are more often addressed in funder publication mandates; outputs like monographs are often absent from such mandates (or else the mandated policies are as yet under development). [13] In the journals and data supply chain, in other words, funders have a strong influence on the business decisions and capabilities that must be built by stakeholders downstream.

[13] This includes the draft public access plan released by the National Endowment for the Humanities, released early 2024 and due to be finalized in December 2024, for implementation by December 2026. The draft plan explicitly states it does not apply to monographs or chapters in edited volumes, but it does cover journal articles and any datasets created as the result of NEH-funded research.

CONTENT FUNDERS *Organizations that provide financial (and other) support for conducting research and the publication of the outputs of that research*	
Supply Chain Contribution	• Establish publication policies for authors • Source payment for open and public access business models • Provide institutional support for platforms and repositories
Information Needs	• Usage performance of funded research • Citation performance of funded research • Policy compliance • Publication activity of affiliated researchers / grants (for example, to confirm coverage of an APC payment under a Transformative Agreement)

4.2 BUSINESS INTERMEDIARIES

Within the journals distribution supply chain, intermediaries exist to facilitate the business of publishing. As noted in Section 3, "Landscape Overview", the traditional model for journals is the subscription, in which access is pre-purchased for all articles published within the coming year alongside access for older archived content held behind a paywall. With the rise of open access, those in the supply chain must support additional models as well (as described in Section 2.1, "Methodology and Treatment of Terms").

Different types of intermediaries have accordingly emerged to support content funders and publishers by facilitating OA related spending, manage institutional subscription purchases and support per article open access cost recovery through transactional Article Processing Charge (APC) payments.

- **Subscription agents.** For journal publishers following a subscription or hybrid model, direct institutional sales can be facilitated by subscription agents. The largest agents, like EBSCO, provide libraries with a single point of contact for handling the subscriptions and renewals of multiple journal titles including the management of multi-publisher payments and invoicing. Agents can provide "boots on the ground" support for publishers in regions where a local presence is the preferred means of doing business and the publisher does not have dedicated staff in that region themselves; examples of such regional agents include Kinokuniya in Japan or Charlesworth Group in China. Over time, many subscription agents have developed decision support services for institutional buyers, such as analytics dashboards to compare cost per use across titles. As more libraries have begun to incorporate APC payments into their spending via transformative agreements, subscription agents have needed to expand their services to

facilitate decision-making around anticipated publication volume by affiliated researchers, and appropriate spend to cover the associated publication costs.

- **APC management services.** Just as subscription agents support subscription purchases across many publishers and libraries, APC management services provide a common intermediary for content funders and authors. In addition to handling payments, invoicing, and reporting for both, APC management services can use logic to apply the appropriate funder-supported publication model and associated pricing (for example, identifying authors covered by a transformative agreement and applying appropriate funds, in addition to ensuring the appropriate license is applied). APC management services may be automated, such as CCC's RightsLink service, or be facilitated manually during the article production process by a production outsourcing vendor. Some services, such as ChronosHub, are evolving to serve multiple functions such as payment handling, policy compliance management, repository deposit, and reporting. Another subtype of APC management service, exemplified by the OA Switchboard, allows for the structured exchange of publication-related financial information between institutions, publishers, and funders to allow parties to track funding eligibility and policy compliance.

The journal publishing supply chain differs from that of book publishing in that business transactions supported by this class of intermediary occurs before article publication, with some portion of transactions supporting recurring spend. (Books, by contrast, earn most of their money after publication, and are far more reliant on transactional revenue streams. See Table 1.) Journal subscription agents and APC management services also often address different parts of an institution's content budget, and so there is often not overlap in the services of journal business intermediaries and book distributors or sales channels.

As access to open research data is typically not monetized, subscription agents and APC management services do not play a role in the open data supply chain.

BUSINESS INTERMEDIARIES *Organizations that facilitate the business of publishing*	
Supply Chain Contribution	• Facilitate article business model and license selection • Facilitate author eligibility check and payment processing • Facilitate author compliance • Facilitate exchange of business model metadata and performance metrics
Information Needs	• Author, institution, and funder information • Funder policy and compliance requirements

4.3 AUTHORS / RESEARCHERS

Authors and researchers are the primary stakeholders responsible for the creation and publication of research articles and outputs; their needs and decisions have a significant impact on the overall workings of both the journal and data distribution supply chains.

- **Authors / researchers.** Articles and data are ultimately produced by those who conduct or collaborate on research. Authorship on a particular research paper can take a number of forms: for example, corresponding authors (or the corresponding author's institutions) are typically seen as responsible for payment of any necessary publication fees. In certain fields (such as biomedicine), author order is determined by the type and amount of contribution to the work. NISO's Contributor Roles Taxonomy, or CRediT, provides structure for the range of possible author and contributor roles.[14]

  Authors are also responsible for choosing the publication option that complies with any appropriate funding or institutional mandate, whether it be publishing their article open access, making a version available via an appropriate repository (either immediately or after some embargo period, depending on the mandate), and/or providing access to associated research data.

AUTHORS / RESEARCHERS *Conduct research and create outputs such as journal articles and datasets*	
Supply Chain Contribution	• Content creation (often multiple versions) • Selection of business model and license
Information Needs	• Research funder and publication funder information • Publisher or repository policies and requirements

4.4 CONTENT PUBLISHERS

Content publishers are responsible for the formal publication of research outputs. Their author services span: editorial development and curation (including facilitating the critical step of peer review), publication formatting and production, and distribution of the final publication (i.e. a formal version of record [VoR] for journal articles.) Publishers are typically the stakeholders responsible for gathering, structuring, and distributing (or causing to be distributed) the metadata and content files for the VoR. In turn, by providing access to the content, they accrue the relevant metrics to measure a publication's impact. Journal publishers determine the business model of each journal and determine what licensing and public access funding choices are made available to a journal's authors. For example, a publisher may require authors to apply

[14] For more on NISO's Contributor Roles Taxonomy (CRediT), see https://credit.niso.org.

a CC BY, CC BY-NC or CC BY-ND licenses to their article when publishing OA. Publishers also dictate whether an author is allowed to deposit a Green OA version immediately upon publication or only after a prescribed embargo period.

Publishers (as classically defined) play an important role in the supply chains for journal articles, monographs, and other peer-reviewed outputs. Publishers may be societies, associations, or research institutes that own and publish journals, in addition to commercial publishers and university presses, and in some cases libraries or groups of scholars who lead their own publishing activities (aka library publishers and scholar-led presses). While there has been some experimentation blurring the lines within the publication process (such as inviting peer review *after* publication, as in F1000's post-publication peer review model[15]), these are typically variations on a theme with "publication" still the primary step in establishing the article VoR. A publisher's role in curation and editorial oversight are critical for maintaining scientific integrity of research outputs and underpins the value of the supply chain overall.

CONTENT PUBLISHERS *Curate, produce, and distribute formal research outputs and associated metadata, and define the version of record (VoR)*	
Supply Chain Contribution	• Content production • Content publication (primarily the version of record) • Metadata creation, provision, and curation • Provision of content to platforms and distributors • Implementation of business model and license terms
Information Needs	• Research funder and publication funder information • Funder and institutional publication requirements • Usage performance of published research articles • Citation performance of published research articles

4.5 CONTENT PLATFORMS

Content platforms are the content management infrastructure and web delivery interfaces that make articles or datasets accessible. They are may be provided by a publisher, institution, or operated by an independent service provider:

- **Publisher platform.** Typically referred to as "platforms", these websites provide access to content – specifically the VoR. In some cases, a publisher may have more than one

[15] For more detail on F1000's model, see: https://www.f1000.com/resources-for-researchers/how-to-publish-your-research/peer-review

platform serving slightly different audiences, which provides multiple points of access to the VoR. For example, a journal publisher may provide a white-label website for a research society's members to access VoRs in a society branded environment while also providing a separate website for library patrons accessing via an institutional subscription. A publisher may also provide different platforms under related brands (for example, Springer Nature hosts many journals on both its BMC and Springer Nature-branded platforms). What's important is that the publisher itself is providing the platform for its own content. These platforms are a primary distribution method in the journals supply chain; as publishers typically are not the primary distributor of data sets this type of platform is not a part of the open data supply chain.

- **Content aggregators.** Publishers may license their content for discovery and access via aggregator platforms. Aggregators sell collections of content (including articles, books, proceedings, magazines, primary source documents, etc., though typically not datasets) from multiple publishers and sources, to library buyers and institutions. These packages can be more cost-effective for library and institutional buyers due to their heavily discounted bulk pricing model. Aggregators may also add additional value by organizing content into topical collections or packages and add additional discovery and search capabilities to serve a specific audience, discipline, or stakeholder. Aggregator royalties have historically provided a supplemental revenue stream to subscription journal publishers, only hosting articles (typically the VoR) after a rolling embargo period to avoid the cannibalization of subscriptions on the primary journal platform. Increasingly aggregators are also exploring ways to incorporate open content into their content databases, to increase the usage and utility of their platforms for users and subscribers. Examples of common aggregators within the journals supply chain are EBSCO, ProQuest, JSTOR, and Project MUSE.
- **Scholarly Collaboration Networks (SCNs).** These are platforms which support content discovery alongside social features designed to build connections across a core researcher user base. Examples include ResearchGate or Academia.edu. Access to the platform and to content is often free for researchers, and researchers can share their articles with others through the platform (after confirming they have the appropriate license to do so[16]). ResearchGate has also integrated with GetFTR, a service which allows users to access content on the platform to which they are entitled through their library's institutional subscription.[17] Publishers are also increasingly striking deals directly with

[16] https://www.researchgate.net/press-newsroom/acs-elsevier-and-researchgate-resolve-litigation-with-solution-to-support-researchers

[17] https://www.getfulltextresearch.com/researchgate-integrates-with-getftr

ResearchGate to provide their content to the platform, in exchange for (anonymized) audience usage and engagement data.

- **Content repository.** These platforms provide a space for authors (or institutions, or publishers on the author's behalf) to deposit a version of their research outputs. The managers and host organizations operating content repositories are as diverse as content funders (*e.g.* government-managed repositories like PubMed Central hosted within the National Library of Medicine, or institutional repositories like the University of Michigan's Deep Blue).

  For journal articles, whether an article is eligible for repository deposit will depend on the publication status and rights retained by the author for the corresponding VoR. For example, many publishers do not allow VoR deposit within another content repository but do allow the posting of an earlier version (like the AAM), sometimes after an embargo period. For a Gold OA article published under a CC BY license, the permissive license means there are no restrictions on the author's (or anyone's) ability to upload the final VoR immediately wherever they choose.

  Repositories play a more central role in the sharing of datasets, which typically do not receive intervention from a publisher. For datasets we can provide an additional layer of definition:

  - ***Generalist repositories*** which host data outputs across subject areas and disciplines, data types, and formats. Examples include Zenodo, Dryad, and Figshare. Given the broad range of content these repositories must accommodate, deposit and metadata requirements for deposit are basic, set at the lowest common denominator across their wide range of outputs.
  - ***Subject-specific repositories*** host data outputs for a particular discipline, often in a specific format. Given the narrower focus, metadata captured about each dataset can be far more granular and specific to a particular research use case or format. Examples come from medical research (i.e. WormBase and FlyBase, used in genetics research and the Protein Data Bank (PDB), chemical and materials science (e.g. the Cambridge Structural Database (CSD), and Inorganic Crystal Structure Database (ICSD), and behavioral and social sciences (e.g. ICPSR).
  - ***Institutional repositories*** are associated with a particular institution to host research outputs from work conducted at that institution. Like generalist repositories, institutional repositories are often not based around a specific data format or disciplinary focus, but are instead designed to promote greater accessibility and visibility of institutional outputs.

Content platforms are the hub of the usage supply chain. They provide the interfaces where content is accessed and thereby usage is generated. However, the type of usage information available will depend on the underlying technical capabilities of the platform.

CONTENT PLATFORMS *Provide the content management infrastructure and hosting services to make research outputs discoverable and available to readers*	
Supply Chain Contribution	• Content management and hosting, including access authentication, preservation, and cybersecurity • Content indexing and discovery channels • Usage capture and structuring usage reporting
Information Needs	• Standard metadata describing content • Key license terms governing content use (from content creator)

4.6 CATALOGS AND INDEXES

Catalogs and indexes aggregate and augment research output metadata across journals, articles, and datasets, to improve discoverability and completeness of the scholarly record. Indexes can also play an important role in creating links between versions of articles, for example connecting a preprint version of an article to its AAM stored in a repository and its final VoR.

Different subcategories include:

- **Knowledgebases.** These global content indexes help libraries discover and manage digital content. Knowledgebases underpin the library **discovery services** which enable search across a library's entire set of entitlements. KnowledgeBases combine metadata from multiple sources and can provide in structured format to **library management systems**. Examples include EBSCO Knowledge Base (KB), Ex Libris Central Knowledge Base (CKB), and the OCLC WorldCat Knowledge Base.
- **Content indexes.** These directory services index and organize content across publishers or platforms, often applying further organizing principles or content curation to promote more specialized search and discovery. For example, the Directory of Open Access Journals (DOAJ) provides an index of fully OA journals that meet prescribed criteria including availability of key journal information, adherence to minimum peer review standards, and licensing and copyright parameters[18]. Another example is MEDLINE,

[18] https://doaj.org/apply/guide/#basic-criteria-for-inclusion

which indexes content in select biomedical journals based on National Library of Medicine (NLM) Medical Subject Headings (MeSH). For data, indexes such as Data.gov allow for search across all of the US Government's open data.

- **Search engines.** As one of, if not the most important method, of discovery, search engines help researchers discover articles on the open web relevant to their interests. For example, Google Scholar was cited by many interviewees as the largest source of referrals to articles, and a critical piece of journals supply chain infrastructure. Google also offers a separate search engine for open datasets.
- **DOI registries**. Managed by organizations as open infrastructure, these registries generate a unique Digital Object Identifier (DOI) for individual research outputs to support persistent linking across systems. In addition, DOI registries can also track the resolutions to each DOI and provide additional helpful data points (such as funding information, license, version, and references.). The de facto standard DOI registry in use for journals publishing is Crossref; for datasets it is DataCite.

Some indexes, like Dimensions, compile metadata feeds across multiple sources (including the publisher, Crossref, and its own machine reading tools) to add key metadata elements and allow users to conduct analysis across different pivots and parameters. Specific to open journal articles, an important resource is Unpaywall, which indexes content across the web to capture each article's OA status; this data on OA status is then passed on to other platforms and services including the Web of Science, Dimensions, and several discovery services.

CATALOGS AND INDEXES *Aggregate and augment information about research outputs to improve discoverability and completeness of the scholarly record*	
Supply Chain Contribution	• Add or enhance metadata (e.g., to support specialized use cases) • Combine metadata across content sources • Structure and restructure information to support better discovery • Build connections between multiple versions
Information Needs	• Standard metadata describing content (from content publisher or platform) • Full content access (from content publisher or platform) • Key license terms governing content use (from content platform)

4.7 ANALYTICS AND METRICS SERVICES

These services compile and consolidate data about the research outputs themselves (e.g. DOIs) as well as tertiary information about the impacts of such outputs, such as usage, citations, and other metrics. Analytics providers are closely related to catalogs and indexes, but provide an extra layer of insight that is important for stakeholders interested in monitoring research performance and impact. This focus on evaluation means they play a very different role in the supply chain, and are a primary channel for information traveling "upstream" (i.e., *decision support*) rather than "downstream" (discovery).

- **Citation metrics providers**. Clarivate's Web of Science includes the most influential citation databases, including several indexes (Science Citation Index, Social Sciences Citation Index, Emerging Sources Citation Index) that underlie the annual Journal Citations Reports (JCR) and the Journal Impact Factor (JIF) metric. JIF is a metric of journal citation frequency commonly used in research evaluation, but not the only index of article citations – Elsevier's Scopus index, the Dimensions database, and Google Scholar all calculate their own citation metrics. These metrics are inherently proprietary in that they rely on the specific index (one cannot reproduce a JIF, for example, without Clarivate's dataset). For datasets, DataCite plays an important role in capturing citations and has founded the Make Data Count initiative to support open data metrics across the board.
- **Usage analytics dashboards**. For intermediaries facilitating and managing ongoing subscription access to journals, information about annual usage is critical to making and managing decisions about subscription renewals (or cancellations). As one example, EBSCO (a sales agent) has produced usage analytics tools like the Panorama dashboard to help library customers with their decision-making. For open content, usage analytics may be used by stakeholders to understand readership patterns and gauge the degree to which they can appeal to a high-value audience; an example of this is ResearchGate's Journal Home reporting dashboards which allow publisher clients to understand usage patterns among researchers at various stages of the research lifecycle (including those close to publishing their own research, who might consider the journal in question as a potential publication venue).
- **Compliance dashboards.** These tools allow stakeholders to monitor compliance with funder mandates and open science principles. Examples include the Open Science Indicators for PLOS Journals developed in partnership with DataSeer, the Open Science Observatory which is part of the European Commission's OpenAIRE catalog of services, and CHORUS dashboards which enable tracking OA compliance metrics across CHORUS member publishers.

- **Combination metrics dashboards.** In response to efforts to diversify research metrics, additional services have emerged to help stakeholders illustrate and evaluate the broader impacts of research on society. Some of these dashboards aim to illuminate engagement beyond scholarly citations and institutional usage (referred to as "altmetrics," as in "alternative metrics"). Metrics combined for a specific research output or scholar could include social media mentions, press coverage, and patent or grant application references. For example, Sensus Impact, launched by Silverchair, Hum, and Oxford University Press, features a dashboard that organizes impact reporting on articles by funding body, capturing and presenting many different metrics including usage (reads and downloads) across multiple platforms, altmetrics, and citations to allow funders to see the impact of the research they fund.

As noted previously, the same organizations that offer analytics and metrics services are often playing another role within the supply chain (for example, subscription agents offering usage analytics dashboards to support their customers with subscription decisions.) The importance of data for making renewal or other recurring funding decisions makes this type of decision support service particularly important in the journals supply chain.

In the data supply chain there is less demand for robust analytics of usage or impact to inform purchase decisions (as the underlying business model is not predicated on purchasing at all). However, some metrics may still be important to gauge return on investment and effectiveness at the platform level, as well as of specific research works.

ANALYTICS AND METRICS PROVIDERS *Tools and services that consolidate data about articles as well as tertiary information about those items (such as usage, citations, and other metrics)*	
Supply Chain Contribution	• Consolidate article/journal/dataset performance information across the supply chain • Provide article/journal/dataset performance data to stakeholders upstream (content creators, content funders) • Add additional pivots or views to data to support stakeholder decisions
Information Needs	• Standard and enhanced metadata about content and related versions • Captured citation frequency and source (from catalogs and indexes) • Captured usage information (from content platforms) • Captured engagement from other third parties (for altmetrics providers)

5 Key Standards and Infrastructure

Table 2 illustrates the key standards in use for journals and research data publication distribution as information is shared "downstream" throughout the scholarly publication supply chain. This section examines the application and interoperability of metadata and data transfer standards in both the open journal article and open data supply chains, and opportunities for new infrastructure to fill gaps in distribution and evaluation.

TABLE 2. Distribution and Discovery Metadata Standards Used in the Journals and Datasets Supply Chains			
Standard	**Parent Organization**	**Research Output Supported**	**Description**
KBART	NISO	Journals, Books	KBART is a NISO standard typically generated by each publisher platform, used to provide journal-level metadata to a knowledgeable or index.
MARC Records	Varies; there are many versions of the MARC standard including MARC21 (US Library of Congress) and UNIMARC (International Federation of Library Associations and Institutions)	Journals, Books	MAchine-Readable Cataloging records are used to exchange bibliographic information. They are used most commonly at the journal level to support librarian cataloging of journal holdings.
JATS (Journal Archiving and Interchange Tag Suite) XML	NISO	Journals, Journal Articles	A common format enabling publishers and archives to exchange journal content; JATS XML is created by publishers during the production process and is often required for indexing services, Google Scholar, downstream discovery, etc.
DataCite Metadata Schema	DataCite	Datasets	A schema governing the core metadata properties for datasets necessary to provide accurate and consistent identification of a resource for purposes of discovery and citation.
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)	Open Archives Initiative	Resource-Agnostic Standard	A low-barrier mechanism for repository interoperability. Data providers are repositories that expose structured metadata via OAI-PMH. Service providers then make OAI-PMH service requests to harvest that metadata.
Google Scholar Indexing	Google Scholar implementation guidance, maintained on the Google Scholar website[19]	Journal Articles, Books	Google Scholar is a search engine dedicated to indexing and searching academic literature (articles, theses, books, abstracts, and court opinions) from academic publishers, professional societies, online repositories, universities and other websites.

[19] https://scholar.google.com/intl/en/scholar/inclusion.html

TABLE 2. Distribution and Discovery Metadata Standards Used in the Journals and Datasets Supply Chains			
Standard	**Parent Organization**	**Research Output Supported**	**Description**
Datasets Type – Schema.org[20]	Schema.org	Multi-output; data providers can use Dataset standard for discoverability in Google Dataset Search	Open standard used to describe datasets to enable discoverability via open web searches like Google Dataset Search.
Other	N/A	Variable	Organizations may use their own proprietary process for gathering information; for example stakeholder interviews noted several journal hosting platforms and indexes which adapt the JATS XML standard into their own proprietary metadata schema; aggregators also frequently collect key metadata from article PDFs using machine reading rather than metadata feeds.

[20] https://schema.org/Dataset

5.1 JOURNALS DISTRIBUTION

5.1.1.1 Journal and article-level distribution

Journals distribution can be understood in two different tiers: **journal-level**, and **article-level**.

Historically, journal subscriptions have operated on the journal level – meaning business model, distribution agreements, and library entitlements were captured at the journal level. A set of standards served this journal-level flow of information, such as KBART and MARC records (which provide information about entitlements to library knowledgebases and library management systems). As journal titles represent the enduring brand that unites articles within issues and volumes, and act as the unit of purchase for library subscribers, journal-level information can be a useful organizing level for evaluation and metrics.

However, an additional layer of information is captured at the article level. As journals began to support a mix of business models with hybrid OA, standards and infrastructure now must expressly distinguish the business model of *each article* published, as a journal may include articles with different terms of access. Accordingly, support for article-level metrics and evaluation have grown supported by digital distribution channels that allow for more granular metrics and data capture.

5.1.1.2 Role of unique identifiers

For both journal title and article level analysis, the use of persistent unique identifiers (PIDs) is critical. These research output identifiers enable the durable, unique reference to a specific document, entity, object, or other item. They are critical for disambiguation, automation, and machine-readability of metadata.

Widely considered to be key metadata elements for research outputs, the following PIDs are seen as critical to understanding the impacts of open and publicly accessible journal articles:

- **Digital object identifier (DOI)** – At the heart of article-level infrastructure is the Digital Object Identifier (DOI), which is a unique and persistent identifier for digital objects (such as articles) issued by a central DOI registry. For journals, this is typically Crossref.
- **Scholar identifiers** – Required to uniquely identify authors and disambiguate author names, these PIDs are increasingly required as part of the publication process and in some cases by national agencies under national security directives[21]. Examples include the commonly used Open Researcher and Contributor ID (*ORCID)* PID.

[21] https://www.whitehouse.gov/wp-content/uploads/2022/01/010422-NSPM-33-Implementation-Guidance.pdf

- **Institution identifiers** – Similarly, unique institutional identifiers are necessary when authors may be affiliated with multiple institutions and therefore have a one-to-many relationship. Examples include the open registry of research organization Research Organization Registry (ROR) and Ringgold IDs.
- **Research funding agencies or sponsors** – An organization(s) that directly funds phases of a research project lifecycle. This institutional identifier emerged to identify the agencies or sponsors that fund scholarly activity that results in a research output. This may need to be further distinguished, as the funder supporting the research (of which there could be more than one) may be distinct from the entity that funded the publication. PIDs used to identify funders may include the same institutional identifiers noted above (in particular ROR, which is in the process of merging with the Open Funder Registry [OFR] maintained by Crossref) but are separate here as stakeholders typically capture research funder and institutional affiliation as separate metadata elements.
- **License type** – The license governing the terms under which the research output can be further distributed, remixed, adapted, and reused; for open access articles these are typically Creative Commons (CC) licenses.

Stakeholders often require additional pieces of information to describe articles for more refined analysis. Examples include:

- **Article type**, which defines whether an article is a research article, review article, letter to the editor, etc., which could be valuable information for research impact related evaluation.
- **Contributor type**, for example to distinguish between different author roles via the CreDiT taxonomy.
- Many OA agreements also hinge on the identification of the **corresponding author**, distinct from other authors. For example, in many Transformative Agreements, eligibility for APC coverage is based on whether the corresponding author comes from that institution (i.e., payment from an institutional deal could not be applied if the author affiliated with that institution were a non-corresponding author).

5.1.1.3 The case for moving metadata authority and assignment upstream

In interviews, stakeholders typically expressed a preference that the authority on metadata elements come from as far upstream in the supply chain as possible so that any modifications flow down to others.

Similarly, to improve metadata quality and reduce effort duplication, the labeling of a specific research artifact with metadata elements should ideally be the responsibility of those "upstream" in the distribution supply chain. For example:

- Funders could couple author (e.g. ORCID), funder (e.g. ROR), and grant identifiers at the time of research grant award.
- Corresponding authors could supply comprehensive funding and institutional affiliation information for themselves and their fellow co-authors at the point at which they submit the article for peer review.
- Publishers could add additional identifiers at the start of the publication event (e.g., DOIs, license type identifier).

If such processes and practices were standard throughout the supply chain, it would improve the quality of metadata flowing throughout the system, ensuring completeness while also eliminating the need for multiple (sometimes conflicting) metadata inputs from downstream stakeholders.

5.1.1.4 Highly variable metadata quality

In practice, publisher provided metadata quality and completeness is highly variable. Certain distribution standards may not provide adequate space to include each metadata element (for example, the KBART standard which populates global knowledgebases does not provide fields for article-level business model data). Indexes and APIs such as the Unpaywall data feed play an important role in filling metadata gaps and providing cross-publisher flows of article-level information.

New distribution tools are appearing that promote and enable standardization around article-level metadata exchange. The OA Switchboard leverages an exchange protocol to support publisher-library-funder information flows during the article publication process, to support use cases such as checking prior to publication for funder mandate compliance, or Transformative Agreement eligibility. The OA Switchboard schema is focused on the article, not the journal, and so the quality of the interactions it enables relies on high-quality article-level metadata being provided from publishers.

However, one of the key challenges faced by publishers is that current processes rely on the author to be the primary and ultimate source of publication metadata at the point of submission and publication. As authors are the primary "customer" of publishers, publishers are wary of introducing roadblocks or requirements in the publication process that increase the time to publication, require more time from busy researchers, or otherwise cause friction with their authors. Such actions would provide a competitive disadvantage for a brand compared to other journals with simpler submission processes in the eyes of the author.

Some providers have emerged (such as the previously-mentioned ChronosHub, or Wiley's Research Exchange [ReX]) to help publishers achieve metadata standardization through author submission interfaces that provide a more intuitive submission process that simplifies the author experience related to funder mandates and transformative agreements.

Better metadata capture from authors promotes better reporting at the article level, and better decision-making. This is especially true in the context of open access business models – which by necessity rely on very different decision points than decisions about subscriptions.

5.1.1.5 Comparing journals and books supply chains

The dual-level distribution model (journal and articles) currently provides an interesting contrast with that of books. As described in C&E's analysis of the open monograph supply chain, books are primarily distributed as complete titles, and while chapter-level metadata may be available, there is not yet a robust flow of information about individual chapters. In the future, mandates requiring open publication of book chapters may increase the frequency that paid access titles will include one or more OA chapters and therefore necessitate more granular information capture. Distinguishing title- and chapter-level metadata was one of the primary challenges identified in C&E's report on the open access books supply chain.

Journal publishing, in contrast to books publishing, has been much more successful in establishing infrastructure for both journal and article level metadata collection and distribution, given the importance of the individual journal article as well as the overarching journal brand in citation, scholarship, and academic reward. Much of this infrastructure has evolved due to the rise of Gold OA and the broader implications of this business model (described in Section 3, "Landscape Overview"), including the increasing shift toward an article-level economy.

5.2 OPEN DATA SHARING

The open research data landscape is diverse, with many repositories serving specialized, niche communities. The degree of independence between repositories means many do not participate in a broader "supply chain" as classically defined (as in, research datasets are hosted and made available but are not then included within broader distribution networks or discovery indices).

Many repositories are designed around the specifications of a single community, and have accordingly adapted their own unique metadata structures.[22] However several initiatives are

[22] The Research Data Alliance (RDA) Metadata Standards Catalog Working Group maintains a list of metadata standards in use by various repositories, available at https://rdamsc.bath.ac.uk

active at the time of writing that seek to increase interoperability and discoverability across repositories, and to formalize the evaluation and management of research data as its own output, rather than the byproduct of other outputs like journal articles. While the metadata schema in development by these initiatives are focused on the most general use cases, they may nevertheless be instructive in identifying the common data points that might be shared across repositories with diverse disciplinary focus.

5.2.1.1 DataCite

In its role as primary DOI registry, DataCite manages an extended metadata schema that supports 28 types of research outputs including not just datasets but samples, software, and also journals and books at the title and chapter level. It also manages other key identifiers such as instrument IDs, data management plan IDs, and project IDs.

The DataCite schema, which can apply to any of the three repository types listed in Section 4.5 "Content Platforms" provides a list of core metadata elements necessary to describe a dataset at the most basic level, with six required properties and four additional recommended properties:

- **Required**:
  - Unique identifier (i.e., a DOI or accession number)
  - Creator / author
  - Dataset title
  - Platform
  - Year of publication
  - Type of resource (i.e., file format and contents)

- **Recommended**:
  - Subject (particularly helpful for comparing impact assessment within a particular discipline)
  - Content license terms (e.g., CC BY)
  - Related identifiers
  - Description or abstract

These properties are repeatable, so that there may be versions of each metadata field in multiple languages.

5.2.1.2 Generalist Repository Ecosystem Initiative (GREI)

A second key initiative defining the metadata elements necessary to support a more comprehensive open data supply chain is the **Generalist Repository Ecosystem Initiative** (GREI), funded by the US National Institutes of Health (NIH) to understand how generalist repositories can better support sharing NIH-funded research data. NIH is a major funder of many specialty databases and options for disseminating research, but there remain many outputs that do not have a natural home. The work of GREI is intended to work directly with a broader group of repositories to establish best practices and make sure they are suitable for NIH requirements, where a subject-specific option does not exist. Seven generalist repositories (Dataverse, Dryad, Figshare, Open Science Foundation, Mendeley Data, Vivli, and Zenodo) are participating in the initiative. A secondary outcome of the project is to raise general awareness of the importance and utility of research data in sharing among researchers themselves.

One of GREI's initial goals has been to establish a Generalist Repository Metadata Schema[23] which specifies metadata fields that should be collected, included in public metadata, and provided to DataCite for DOI registration, including recommended use of controlled vocabularies and persistent identifiers. Because of the importance of DOI registration for overall interoperability, much of the GREI schema overlaps with the DataCite schema (specifically the DataCite Metadata Schema 4.4). The GREI initiative is envisioned to apply these recommendations to support the specific needs and interests of both generalist repositories and institutional repositories (which, as noted above, by nature must support discovery of diverse research outputs).

5.2.1.3 Diversity of metadata standards in use

Though GREI and DataCite are two examples of initiatives to better unify metadata standards and enable discovery, application of such initiatives are still not universal. Some databases might issue their own accession number, which is an identifier that is unique within that data repository but that does not ensure global uniqueness nor follow any generally recognizable format, rather than apply a DataCite DOI. The use of nonstandard persistent identifiers limits the effectiveness of the data supply chain, as datasets cataloged with proprietary IDs are more difficult to connect to the broader ecosystem. For example, AI tools trained to identify citations to each dataset or to affirm compliance with data sharing policy may not recognize all types of accession numbers and therefore miss important references.

Subject-specific repositories or those serving a niche audience also typically have their own highly-developed classification schema developed over time within a field or discipline. The

[23] The initial recommendations, published in June of 2023, are available at this link: https://zenodo.org/records/8101957

specificity of these schema means they may not map neatly to external standards (which indeed is part of their value). In other cases, a domain-specific schema may be an extension of a more generalized schema: for example, the bioschemas.org[24] schema is an extension of the more general schema.org dataset markup which enables indexing in the Google Dataset Search engine.

5.2.1.4 Comparing open data and books supply chains

Comparisons between open data supply chains and the open articles and books supply chain further illustrate the contrast between business and funding models. Because many journals and books supply chain requirements are defined by the distribution model – be it recurring journal subscriptions, transactional article-level payments, or distribution of individual titles – suppliers, distributors, and agents are instrumental in defining the level of information and metadata standards required to participate in the broader publishing industry. (For example, the ONIX book distribution standard grew in popularity as it became a requirement for distribution via Amazon.)

In the data supply chain, the stakeholders involved are not driven by a business model but rather by the need to maximize and evaluate funding decisions. This makes it less likely for outside actors to influence standards uptake for better interoperability. Instead, the open data supply chain relies on stakeholder communities (and their funders) working collaboratively on initiatives such as DataCite, GREI, IRUS, etc. A particular platform's ability to achieve that standard depends on the strength of its need and ability to fund the technical development toward that standard.

5.3 MEASURING USAGE

Table 3 illustrates the different methods of capturing usage for journals and datasets.

[24] https://bioschemas.org

TABLE 3. Usage Reporting Standards and Services in Use for Journals and Datasets			
Standard	**Defined By**	**Research Output Supported**	**Description**
COUNTER 5.1	COUNTER Metrics	Journals, Books, Other	COUNTER reports are the industry standard for developing consistent, comparable, and credible usage reports for scholarly publishing platforms. The COUNTER 5.1 standard, released in May 2023, will take effect in January 2025, and defines reporting at the platform, database, title, and item level.
Standardized Usage Statistics Harvesting Initiative (SUSHI)	NISO	Journals, Books, Other	A protocol defining the automatic harvesting of COUNTER usage data
Code of Practice for Research Data	COUNTER Metrics	Open Data	COUNTER's Code of Practice for Research Data was published in 2018 and is designed for use by research data repositories. In May 2024, COUNTER initiated a consultation on a proposal to merge the Code for Research Data with COUNTER 5.1.
Web analytics (i.e., Google Analytics, Adobe Analytics)	Google / Adobe / other third parties	N/A	Other forms of web analytics are not specific to research outputs, but can provide insight into user engagement, search engine optimization, popular search keywords, and so on.
Proprietary analytics	N/A	Variable	Platforms may have developed their own analytics to capture user behavior and illustrate patterns not captured in generalist web analytics. For example, tools which identify potential authors and analyze behavior to measure engagement and potential interest in submission to a particular journal(s).

Usage is captured at the platform level (see Section 4.5, "Content Platforms"), and the platform determines the reporting standard in use and its implementation. Often a platform may provide different types of reports to different audiences to address different types of use cases and decisions (e.g. purchase decisions, funding decisions, audience engagement, etc.) In this section we cover the most common usage reporting standards and their application.

5.3.1 COUNTER Reporting

COUNTER is a key reporting standard initially designed to help libraries evaluate usage of purchased content by readers at their institution. For journals, COUNTER reports have historically been critical in the evaluation of subscription purchases and to inform decisions about whether to renew a subscription. In keeping with distribution and sales infrastructure built around the journal subscription, the most important information was captured at the journal or title level. At the simplest level, COUNTER's code of practice defines how web analytics metrics should be "counted" and curated for institutional reporting and analytics. It focuses reporting practice on counting, "genuine user-driven usage, successful valid page requests, and relevant content items" while requiring the exclusion of, "robot usage, pages that fail to load and bad page requests, and the counting of non-content records (e.g. style sheets)."[25] It also defines a set of machine-readable report formats, enabling interoperability and downstream harvesting (via the SUSHI protocol) of COUNTER reports via APIs.

In 2017, Project COUNTER (recently renamed "COUNTER Metrics") released the new COUNTER 5 reporting standard, which was updated to help libraries and publishers address questions about usage of OA content. A newly defined "item-level" report format supports granular journal article (or book chapter) level reporting, supporting consistent analysis beyond the journal or book title level. This new item-level reporting format distinguishes usage for subscription access, open access, and free to read content. This directly informs library purchase decisions, as staff can now determine cost per use for only the articles accessed via subscription as they can filter out usage related to open or free content.

In 2024, COUNTER is adopting a Code of Practice for Research Data developed by the Make Data Count initiative, providing an iteration of the COUNTER standard customized for dataset usage tabulation and reporting. In May 2024, COUNTER sought community consultation on the proposal to merge the MDC Code for Research Data Practices with the COUNTER 5.1 standard released in May 2023, thereby providing the new prevailing code of practice for COUNTER compliance from January 2025 on.[26]

[25] https://medialibrary.projectcounter.org/file/compliance-guide-1

[26] https://www.countermetrics.org/crd-consultation/

5.3.1.1 Varying levels of COUNTER adoption

To reflect their adoption and compliance with the COUNTER code of practice, content platforms may market to libraries using terms like "COUNTER compliant", "COUNTER conformant", or "COUNTER like", indicating varying levels of adoption of COUNTER's code of practice. To be listed on COUNTER's registry of "COUNTER compliant platforms"[27], a platform must pass an independent third-party audit (from a certified CPA or one of three listed auditor agencies -- a process that takes 3-12 months[28]), and agree to provide data in specified report formats via JSON and spreadsheets accessible via the SUSHI protocol, aggregated at a monthly level going back two years.

Such compliance requirements may be financially difficult to achieve for content platforms with little resourcing (e.g. scholar-led initiatives, startups). Platforms that do not serve libraries via subscription sales may have little incentive to participate formally ("COUNTER-like" reporting may be all that is needed). Stakeholders interviewed for this project highlighted the need for a more accessible means of participating in the full standard. In the meantime, some data repositories might design their reporting standard around the COUNTER Code of Practice, but not complete the formal steps to demonstrate compliance and be listed on COUNTER's registry.

5.3.1.2 Navigating Reader Privacy and IP Addresses Blocks for Reporting

The community developed COUNTER Code of Practice aims to provide usage tracking and reporting that protects reader privacy. User-specific "item-level" (i.e. research output) reports summarizing what a unique individual accessed over time are not supported, to protect individual user privacy. Rather, institutional access (meaning, usage across all individuals located at an institution) is the most granular level of reporting currently allowed in the COUNTER Code of Practice. As a result, many platforms traditionally built to support library subscriptions are not equipped to report on usage more granularly than at the institutional level.

This leads paradoxically to one of the main challenges of the COUNTER standard when it comes to tabulating usage of publicly accessible content: As usage occurs increasingly outside an institutional paywall (because users no longer have to authenticate into their institutional account to gain access), it is more difficult to attribute usage to a specific institution.

COUNTER's Code of Practice includes a "global report" standard, to reflect the total usage of an item on a given platform. Some publishers and service providers leverage IP address level web analytics to attribute usage to organizations or at the very least to various geographic areas,

[27] https://registry.countermetrics.org/

[28] See https://medialibrary.projectcounter.org/file/compliance-guide-1 pg 5.

often relying upon service providers such as PSI / IPregistry. This is an imperfect science – as users may not be accessing content from an IP range associated with an organization (e.g. from home) or may be accessing by Virtual Private Network (VPN) to shroud their actual location. That said, IP address affiliated usage could nonetheless provide useful and directional information about regional usage and the ability for OA content to reach new audiences.

5.3.1.3 Report consistency

A second challenge for the COUNTER standard is ensuring compatibility of implementation across platforms. For example, the COUNTER Code of Practice specifies how web traffic generated by bots and spiders should be processed and omitted. It also guides how duplicate webpage visits to an item in a given period of time by a single IP address should be counted. Yet, it is difficult for content platforms to reliably distinguish machine-generated traffic from actual reader engagement, especially when there is no paywall or authentication layer turning a large fraction of bots away. While COUNTER maintains a list of known bots and sites whose traffic should be removed from usage statistics, stakeholders nevertheless reported a lack of consistency in adherence to such lists across platforms, making cross-platform comparisons of usage difficult if not invalid. The same challenge then occurs in comparing usage patterns across subscription and open content, i.e. how much of an increase in usage for OA content is due to increased engagement versus increased bot accessibility?). Issues resulting from inconsistent bot management will only be exacerbated with the growth of AI and the increased interest in leveraging open content to train AI models.

Currently supply chain value is derived as platforms consistently track and report usage data for research output "items" over time. However, the above challenges become more significant as usage reporting relies on interoperability across platforms, i.e. when reports are combined or compared *across* platforms for evidence based decision-making. This was noted in C&E's previous report as an important challenge within the supply chain for open access books, given its reliance on distribution of the same title across multiple platforms. Journal articles and open data will share this same problem to the extent they are distributed across multiple platforms, and in multiple versions (see Section 6, "Gaps and Opportunities", below).

5.3.2 Make Data Count

The Make Data Count initiative is a project hosted by DataCite to increase support for measuring research data usage and citation. Its goal is to ensure that these categories of metrics are built in an open, transparent, and meaningful way that encourages better evaluation of open research data. This includes being able to provide a normal and rationalized view of usage, so that the same definition is applied across platforms. This shared definition then allows comparison and evaluation of usage in more meaningful ways (addressing the platform discrepancy issues noted above impacting journals and books).

Make Data Count aims to increase normalized dataset view and download reporting by advocating for specific solutions for repositories, including becoming COUNTER compliant (which, as noted above, may be a challenge for repositories that lack technical support and funding), or by adopting the embedded DataCite usage tracker into their content platform's discovery pages where "items" are accessed. In either approach, repositories can make usage data available via DataCite via the DataCite API or the DataCite Commons index. Resulting usage metrics may be summarized at the repository or item DOI level.

It is notable that Make Data Count and DataCite go beyond advocating for COUNTER standards adoption by directly supporting repositories with lightweight implementation tools designed to simplify standards implementation. While the uptake of the Usage Tracker tool is still in the early stages beyond generalist repositories, the initiative nevertheless provides an important example of how efforts might support stakeholders with COUNTER adoption.

5.3.3 Institutional Repository Usage Statistics (IRUS)

Institutional Repository Usage Statistics (IRUS) is an aggregation service which works across institutional repositories to help collect and standardize usage data for institutional repositories and institutional data repositories. The service was initially launched by JISC in the UK, but has since expanded to other regions; it is also in use by repositories in Australia and New Zealand through a partnership with JISC and CAVAL and, in 2020, became available in to libraries United States through an exclusive distribution partnership with LYRASIS[29]. It is also in use by the CORE (COnnecting REpositories) database of open access content.

IRUS works by providing a means to capture, normalize, and report on usage in a standardized way, following the COUNTER Code of Practice (Release 5). Implementation is via the IRUS Tracker, which is compatible with many common repository platforms (such as DSpace, Esploro, Equella, Figshare, and others). Usage is then processed by IRUS into a COUNTER-compliant format and made available via an online web portal, an API, and embedded widgets. Examples of information IRUS is able to capture include:

- Views
- Downloads
- Usage of items authored by a single individual (identified via their ORCID iD)
- Usage statistics for a specific DOI
- Platform-wide aggregated usage stats
- Usage benchmarking against other repositories

[29] https://lyrasisnow.org/press-release-jisc-and-lyrasis-help-us-universities-and-research-organizations-gather-new-usage-insights/

- Combined usage statistics (i.e., usage for content included in multiple repositories and via the CORE database).

As the IRUS service is primarily designed for use by institutional repositories, it must accommodate a wide variety of content types, including journal articles, books, book chapters, datasets, chemical structures, and many others[30]. As noted in our discussion of institutional repositories in Section 4.5 "Content Platforms", the version of the article stored in those repositories may be the AAM, rather than the VoR. Other outputs stored in institutional repositories can be diverse, but implementation of the IRUS service can allow for a standardized and verifiable means to capture usage of these items, which can contribute to further evaluation of data sharing best practices and investment (including identifying which data types are assets to future research, and which do not tend to lend themselves to future reuse or utility).

5.3.4 Proprietary Analytics

Platforms and repositories often incorporate additional web analytics to better understand their audience, item engagement and website user experience. Examples include the following:

- Google Analytics is commonly used across journal platforms, although stakeholders noted that Google Analytics statistics are not typically used to compare usage across platforms or over time because of the potential for differences in measurements. Such web analytic metrics are not repeatable, replicable, nor auditable by external parties, making them an inappropriate tool for research evaluation. Web analytics are, however, very good at helping a platform understand how the user interface attracts and services different audiences.
- ResearchGate is a social network for researchers which also acts as an important discovery hub for research. It provides proprietary analytics to publisher members of its Journal Home service to inform how they can best use tools to increase content discoverability on the ResearchGate platform. ResearchGate provides a dashboard illustrating usage patterns across institutions and by different audience segments, including levels of engagement with a specific journal. Such metrics are the natural next step in the journal supply chain's evolution toward an article-level economy, in that the metrics of success are built around attracting potential authors. Publishers can use this information to evaluate their performance along these axes: for example, how many researchers engaged with their journal several times this month? How many of those

[30] https://irus.jisc.ac.uk/r5/about/policies/item_types/

researchers were early career researchers? How many researchers went from accessing the journal infrequently to becoming a more engaged reader?

While analytics providers are not compelled to follow the same privacy guidelines as a library-focused standard like COUNTER, publishers and platforms are nevertheless restricted in the information they can gather by privacy law. The most common benchmark is the GDPR guidelines enacted in the European Union, which place restrictions on the sharing of personal information (and indeed have led many platforms to avoid storing personal information altogether). Even ResearchGate, which as a social network develops researcher profiles, does not provide publishers with reporting beyond the institutional level. While services providing granular metrics to platforms and publishers are increasingly valuable (for both editorial and commercial reasons), there were few expectations expressed among stakeholder interviews that these metrics will be part of public-facing reporting and research evaluation in future.

5.4 CITATIONS AND OTHER MEASURES OF IMPACT

It must be noted that there is already strong infrastructure in place within the journals supply chain to measure research impact, with citations acting as the leading metric. This space has a high involvement of commercial players, with the Clarivate Journal Citation Reports indexes and Journal Impact Factor (JIF) the most prominent. While there are other citation databases and metrics (Elsevier's Scopus, Google Scholar Metrics, and potentially in the future a not-for-profit addition from OurResearch's OpenAlex), the JIF has played a unique role in career evaluation and advancement for researchers, who are often incentivized to publish in journals with a sufficiently high JIF.

That said, citations are an imperfect measure of impact in some fields (like the arts and humanities) and for certain research outputs (such as clinical practice papers, which are widely read but rarely cited). Many funders, institutions, and publishers are signatories to the 2012 San Francisco Declaration on Research Assessment (DORA), which promotes moving away from journal-based metrics such as the JIF and instead using a range of article-level metrics for evaluation. Twelve years after DORA's release, we found continued support expressed among stakeholders for a deemphasis of JIF, but research culture and incentives have shifted only slowly. In many parts of the world, publishing in journals that have attained a certain JIF is still necessary for career advancement, and C&E's market research has found many researchers still report feeling pressured to make publication decisions with a journal's impact factor in mind.

For datasets, the use of citations and impact are far more nascent. Here, as with usage data, initiatives such as Make Data Count are actively working to establish a greater understanding of the value of data citation and evaluation to stakeholders. The Data Citation Corpus developed by DataCite in partnership with the Chan Zuckerberg Initiative (CZI) is a public dashboard and

data file which aggregates citations related to various repositories to support the analysis of data reuse. It currently includes datasets from 40 repositories in the life sciences disciplines as a proof of concept, with the goal to expand to further disciplinary repositories.

The more nascent status of citation capture for open data means that these metrics may not be widely used in research evaluation and career progression as yet, but stakeholders are anticipating this development. For both data and journals, any consideration of measuring impact via usage should not ignore the parallel role that citations will play in the same supply chain.

6 Gaps and Opportunities

In this section we assess the overlap and distinctions between the distribution and evaluation of open journals, data, and books, to understand where there is potential (or not) to build shared infrastructure for usage and impact assessment.

6.1 JOURNALS SUPPLY CHAIN

The best way to understand the open journals supply chain is to understand its evolution. Journals have historically been sold via subscription, on a recurring basis, and are published continually over the course of years with no fixed lifespan (in contrast to books). This has resulted in the traditional journals supply chain's two primary features:

- **Cyclical flow of information.** Because subscription purchase decisions must be renewed every year, there needs to be a strong and simultaneous flow of information back "upstream" to content purchasers, and many services have emerged to provide support for these decisions. The criteria defining whether a journal is valuable to its readership and community are complex – usage (especially as it relates to cost per usage) is a well-understood metric, as are metrics of research quality such as citation frequency. The Gold OA open access model changes the central purchaser from the library to the author, which has in turn changed the decision-making calculus (libraries prioritize patron usage; authors prioritize citation impact), but the supply chain nevertheless benefits from a legacy of strong infrastructure around metrics capture and reporting.
- **Centralized and controlled distribution.** Journals tend to provide their own primary "distribution" in the form of a dedicated web platform to which an entitled user is given access. This is true whether the buyer is an institution (as with a library subscription) or an individual (as with a member subscription). While there are sales agents who can (and do) facilitate subscription purchases in many cases, and there may sometimes be distinct sites for institutional vs. member users, the important thing is that the publisher is the ***primary*** channel for access. In the traditional journal model, secondary channels –

such as aggregators and repositories – may host the same content, but typically after a short (6-month, 12-month, or longer) embargo period.

These features are in contrast to those for outputs like books, which evolved as outputs published (and often purchased) one at a time as discrete titles, with a limited time to recover their initial investment. Book publishers are accordingly more heavily reliant on third-party channels for sales and distribution. The influence of third-party channels and platforms is one of the primary challenges behind book usage aggregation and measurement, and underpins the need for infrastructure to better support metrics exchange across platforms, such as the Open Access eBook Usage Data Trust (OAEBUDT).

In an increasingly open environment, the journals supply chain is seeing a shift in the dominance of the primary distribution model and channels. There are several changes underway that are creating new infrastructure needs and opportunities, as outlined below.

6.1.1 Emerging Trends

6.1.1.1 Shift from journal- to article-level assessment

As more journals follow a hybrid OA model (in which some articles are published OA and others are published on a subscription basis), there is the shift from a **journal-level** to an **article-level** economy. Key metadata about business models must increasingly be applied to the article. In reporting, the COUNTER standard has adapted to introduce item-level reports and to segment out OA and free articles from subscription articles, so that libraries can evaluate the effect of spending with only paid-access content in mind.

In addition to the business model shift, research culture has itself begun shifting (albeit slowly) toward an article-level understanding of impact. The move to digital delivery has led to new capabilities for capturing metrics (including altmetrics) for specific articles and a push to move away from journal-level metrics like JIF, exemplified by funders, institutions, and publishers pledging support for DORA.

These new capabilities are powerful insofar as there is article-level metadata flowing through the supply chain. A few standards, particularly those that have historically focused on library entitlements (i.e., KBART and MARC records), remain focused at the journal level. Other workflows, such as transmission of content and metadata from publishers to certain aggregator platforms, lack consistent standards to describe business models at an article level – publishers may vary in the format of their data feeds, or may simply provide an article PDF from which the aggregator extracts what information it can. Third-party services such as Crossref and Unpaywall can provide an important source of article-level data to close the gap, but the more different metadata streams are combined, the more opportunities there are for metadata errors to be

introduced. Many aggregators, as a result, do not "ungate" OA articles in hybrid journals, but instead set the business model universally for a journal based on their contractual agreement with that publisher.

Interviewees for this project agreed that content creators (authors and publishers) are the ideal source of truth for information about an article's business model, funding sources, and other critical metadata. Not all of the systems and infrastructure in place allow for capture of these metadata fields, let alone in a standardized or machine-readable way. Initiatives to encourage further improvements in metadata quality (such as the OA Switchboard's "Year of Metadata" and NISO's Working Group to Develop Recommended Practice for Operationalizing Open Access Business Processes) are important to ensure reporting is available and, importantly, accurate.

6.1.1.2 Increasing syndication and version proliferation

The increase in articles published via Green and Gold OA is creating opportunities for article discovery beyond the publisher's own platforms. Articles published with a permissive CC BY license can be reposted or reused without the publisher's (or author's) permission, and so copies may proliferate across the web without the publisher's consent or even knowledge. Some publishers are also increasingly open to content syndication, in which content is posted to a partner site simultaneously with publication on the journal's primary platform(s). (This is in contrast to secondary distribution via aggregators, in which articles are typically embargoed for a period of months after initial publication.) Content syndication deals are typically intended to expand a journal's audience and gain additional exposure to a journal; some examples include syndication partnerships with ScienceDirect (Elsevier) and ResearchGate.

This proliferation across platforms increases the need to report aggregated usage across platforms – to the extent that this usage information is available. Usage data sharing is typically a condition of syndication partnerships – for example ResearchGate, an increasingly important syndication channel for publishers, provides institutional COUNTER statistics to its publisher partners, which the publisher can in turn incorporate into its own reports to libraries. It should be noted that because ResearchGate limits access to subscription content to only those from subscribing libraries, these COUNTER reports are intended to support subscriptions to paid access content, not the evaluation of open content. ResearchGate also provides additional analytics, such as usage and engagement patterns, to publishers via its Journal Home services. Publishers can use these tools and analytics to gauge and refine their B2C marketing capabilities and quantify Journal Home's ability to attract submissions from promising authors. This latter type of usage sharing is an increasingly valuable part of the "upstream" supply chain for journal articles as models like author-pays open access reorients the author as the primary "customer" publishers must serve.

Outside of such formal syndication partnerships, visibility into usage of alternative article versions is low. This is especially true of preprints and AAMs, which often have much less key metadata (or slightly different metadata) than the published article and are therefore difficult to link with certainty to the appropriate VoR. It is an open question to what extent usage of these alternative versions matter to stakeholders. Several publishers and librarians interviewed for this project were generally not invested in the performance of preprint or AAM versions. However, the interest and need to consider the non-VoR may become increasingly important as funder policies shift: in the US, the Nelson Memo requires immediate public access to research outputs, which would include Green OA deposit of the AAM (if the publisher allows it). The Gates Foundation's new policy taking effect in 2025 introduces a new mandate requiring preprint deposit, whether the preprint will move forward to formal publication or not. Efforts to monitor the effects of these mandates (including any unintended consequences) undoubtedly will create new reporting needs among content creators and funders.

Broadly speaking, existing performance metrics have focused on the VoR for both practical reasons (the publisher's platform containing the VoR contains strong connections to the appropriate indexes and dashboards) and cultural reasons (the journal's brand itself can convey a level of prestige to authors publishing in that journal).[31] While there have been improvements in the mechanism to connect different versions, these connections appear to be largely developing further "downstream," by catalogs and indexes, rather than by publishers upstream (who have an incentive to prioritize the VoR as the culminating peer-reviewed output). Some publishers are bucking this model by supporting post-publication peer review – for example, eLife and F1000 – but these are largely exceptions that prove the rule.[32]

As the US OSTP "Nelson Memo" mandate comes into effect, resulting in an increase in simultaneous access of multiple article versions – both the AAM and the VoR, services like Unpaywall and Google Scholar that index and connect these different article versions may prove increasingly disruptive. Better discovery of alternative versions will have as-yet-unknown effects on usage of the VoR. Navigating this tension will be at the heart of how publishers will have to navigate the potential risks of immediate Green OA to journals that operate on a hybrid

[31] Lisa Janicke Hinchliffe wrote an excellent *Scholarly Kitchen* post in 2022 providing a landscape scan of the "version of record" and illustrating exactly why it continues to take primacy in research evaluation and career advancement for authors; available at https://scholarlykitchen.sspnet.org/2022/02/14/the-state-of-the-version-of-record/

[32] The Bill and Melinda Gates Foundation, a previous member of the cOAlition S funders, changed its policies in March 2024 to require authors to deposit a preprint of their article, and to cease support for APC payments. This may result in publishers increasing service provision for preprints (for example, automating support for preprint deposit for authors) but it remains to be seen whether this will influence the version considered in research performance metrics.

subscription basis, putting downward pressure on subscription pricing and potential cancellations. There is potential for a negative paradox to emerge: linking article versions is a prerequisite for monitoring the effects of alternate versions on usage and citation, but may also expose for a publisher that more usage accrues for a version other than the VoR.

Even for the VoR, not all sources of usage are equally transparent. Aggregators, for example, are typically not good at providing granular usage reporting to publisher partners. Within the aggregator / publisher relationship, the key metrics of performance are typically at the journal or even the publisher level (as in, how much usage was generated across all content from that publisher), so limited information is available about specific articles. Interviewees noted that even if article-level data were available, it would be difficult to combine with other data sources in a meaningful way. It is also difficult to know exactly how much the CC BY license has allowed informal proliferation of articles across platforms, repositories, or other websites, let alone to gather usage data from those sites. A clearer value of that usage information and *why it matters* will be necessary to increase the incentive to track down and consolidate, or even to provide usage information for compilation.

6.1.1.3 Success criteria in research assessment

A final distinction of the journals supply chain is a more mature range of services focused on monitoring research performance, including citations, compliance, and increasingly other performance indicators such as usage. Journal articles are a primary research output for many fields, including those heavily reliant on research grant funding (as opposed to books, which are more typically a secondary output in STEM fields, or are a primary output for HSS fields with lower levels of funder influence). Due to the journal article's importance in assessing research performance on many levels (individual tenure and promotion, institutional performance tables, and awards for future grant funding), a significant amount of infrastructure has developed around capturing and measuring article and journal performance. Most analytics providers also consider performance metrics along other pivots as well – to measure, for example, the research performance of a particular institution (or to compare across institutions), across funders, across individuals, and across fields.

Among success criteria already in place for research assessment, it must be noted that different stakeholders will have different interests. Librarians, as noted previously, have been historically interested in usage because it provides evidence that purchased content is valuable to patrons. As libraries increasingly fund the publication of open access content, usage generated by a particular institution becomes less important. Instead, success is better measured in global usage and reach – how is this research output being used around the world? This global perspective overrides the importance of cost per use, the traditional quantitative assessment benchmark used to support subscription renewal decisions. What is less clear is the benchmark

for the value of open or public access content – for a library funding a Subscribe to Open program, for example, what usage threshold or metric defines whether participation has been worth the investment? And to what degree will these metrics inform future participation in that program?

For both a funder and an author, usage is important to illustrating the impact of their work. However, in certain fields, the leading success criteria for a research publication may not be usage at all – these stakeholders may focus more heavily on citations as the key indicator for valuing a "purchase" (i.e., publication) decision, in that it can illustrate more concretely that the published research is influencing further research in the field. Other measures of value might be captured in a research output's contribution to policy documents, patents, patient education, or in engagement of a lay audience (through social media or news mentions, for example). Ultimately, an author's career advancement prospects will be based on a hierarchy of metrics: highly cited works may create more value to that researcher than works that are highly used. No one metric captures all measures of value.

This is still a time of experimentation around supporting funder and author decision-making, and new services will continue to emerge to serve this goal. Ultimately, the evolution of further infrastructure around usage depends on there being a suitable appetite for this information, and evidence that it is informing funding and publication decisions.

6.1.1.4 Summary

Table 4 summarizes the emerging gaps and opportunities within the journals supply chain.

TABLE 4. Journals Supply Chain Gaps and Opportunities			
Trend	**Description**	**Gaps**	**Opportunity**
Shift from journal-level to article-level assessment	Access model increasingly determined at the article level rather than at the journal level (for example, an article made be made Gold OA in a hybrid OA journal; articles may be made OA only for a particular year via Subscribe to Open)	Improvements needed in article-level metadata flows and accuracy Inclusion and support of key business model information needed at the article level for all stakeholders	Assessment of ROI focused on specific articles rather than relying on journal-level metrics
Increasing syndication and version proliferation	Articles may appear on multiple platforms, and in different versions, due to an increase in Green OA, permissive licenses, and syndication deals	Article versions not always connected within the supply chain Usage of different article versions not always visible Platforms not equally transparent in sharing article-level usage	Syndication deals include sharing of usage information as a key component of the partnership Indexes increasingly linking different article versions create opportunities for consolidated reporting across versions
Success criteria in research assessment	Different business models create different metrics of success for evaluating return on investment	Traditional cost per use metrics of ROI for library buyers no longer applicable when paying for OA publication Researchers and funders have numerous metrics they may consider when determining the impact of an article	Services continue to emerge analyzing impact along new benchmarks (for example, across funders, individuals, and fields)

6.2 DATA SUPPLY CHAIN

Access to published research data is customarily free of charge, with data provided as a courtesy by the authors of research publications to confirm the reproducibility of their work, or to advance available knowledge within a field. The elements of each underlying dataset are also extremely diverse, with an almost unlimited range of formats. Accordingly, the supply chain for data has evolved less from a need to *curate* and *publish* (in the formal sense of the word) but rather to *share* and *distribute*, with the intended audience being primarily other specialists within the data output's relevant field[33]. It is only relatively recently that attention has been paid to adding more formal infrastructure to support the impact and value of data sharing more broadly, and for data sharing itself to become a mandated step in the research lifecycle.

The primary characteristics of the open data supply chain can be summarized accordingly:

- **Outputs can be made available in more raw form.** Many stakeholders that play a critical role in the publication of formal outputs (articles, monographs) are not required for the publication of data. This means many of the entities that play a critical role in the creation and curation of metadata (publishers, aggregators, analytics providers) are not providing similar services for datasets. This is not to say that there is limited information available about each dataset, but rather that the authors themselves (at this point at least) are responsible for the provision of accurate and structured information to enable reuse and discovery. Some platforms conduct basic checks or validation protocols to ensure the minimum viable level of information is provided to describe or identify the data provided. Given the diversity of data outputs that may be provided, much of this is left up to the author.
- **Infrastructure and services to monetize access to data are more limited.** While some repositories charge for access to analysis tools or visualizations, the data ecosystem does not focus on monetization of the content itself, unlike journals and books. The incentive structure for stakeholders in the open data supply chain is not built around a specific business model but rather driven by the incentives of research integrity – providing information that enables replication of a given experiment or finding – and acknowledgment – where data is used as the basis of future research. Therefore, stakeholders who have historically supported the monetization of research access – such as publishers, aggregators, commercial databases, and so on – are largely absent in the data distribution supply chain.

[33] See above reference to the NIST RDaF.

Rather than finding “shifts” in the open data supply chain, we found instead that this supply chain is at an earlier stage of evolution. This leaves significant room for emerging needs and opportunities, as outlined below.

6.2.1 Emerging Trends

6.2.1.1 Raising awareness of the role of open data in research evaluation

Although many funders have data management requirements and indeed mandate the open sharing of data resulting from grant-funded research, compliance with these requirements remains patchy and difficult to track. Some early solutions are emerging: DataSeer.ai is an example of a service provider helping funders, publishers, and institutions to identify and track data sharing compliance and gaps; it recently partnered with PLOS to develop a series of Open Science Indicators including an indicator illustrating the frequency of data sharing. However, the lack of consistent metadata standards and persistent identifiers used across repositories makes the identification and tracking of datasets challenging.

Organizations like DataCite are taking important first steps in building adequate tracking mechanisms for open data. However, there must be a great deal more standardization throughout a wide range of repositories, especially discipline-specific or specialty repositories, to scale an automated solution that detects and verifies connections between research funding, published articles, and datasets. This foundational work must be in place as a prerequisite before stakeholders can fully understand the potential value of bringing usage of that data into research evaluation metrics.

6.2.1.2 Capturing data sharing use and reuse

One of the primary pain points the Make Data Count initiative aims to address is inconsistent data citation by journal publishers, who do not often act as data repositories themselves but who do incentivize data sharing through author requirements. Repositories using a DataCite DOI benefit from having a consistent and persistent identifier recognizable to both humans and machines. Other repositories using a proprietary accession number may not have the same level of visibility – an algorithm may not identify the persistent identifier as such and therefore not count the citation. (This challenge is why the DataCite Commons citation corpus is focused only on a limited number of repositories in a certain discipline, for now.) Journal publishers as well may fail to appropriately capture the reference, for example allowing authors to reference an article describing the dataset without linking directly to the dataset in a repository.

Addressing this gap requires improving the practices of those outside the formal data supply chain. The benefit of increasing research integrity and complying with increasing mandates around data sharing provides publishers with plenty of incentive to improve such connections in the future.

6.2.1.3 Appetite for expansion of data sharing with limited visibility into costs

The lack of a formal "publication" process for open data does not mean the costs for maintaining data services are insignificant. Data as an output is often larger in size and can be more unwieldy than books and articles, or is data as a whole a standardized format. Storage costs and infrastructure maintenance can be significant, requiring substantial infrastructure to sustain.

For institutions and funders footing the bill, usage attributed to a repository can be a valuable metric to evaluate whether that repository is fulfilling its mission in increasing visibility of research data and outputs. Institutions and funders of repositories have the potential to capture a wide range of different types of data, but need to make decisions about how to direct investment towards data outputs that will generate utility and re-use in the long term. Usage patterns can help to inform decisions about allocating resources toward data with clear and enduring community benefit (such as longitudinal datasets).

More broadly, repositories of all kinds are expected to become more widespread and necessary in future. The 2022 Nelson Memo introduced a new requirement for data sharing by authors receiving US federal funding (the details of the policy for each agency will be finalized by the end of 2024). One of the most notable aspects of this federal requirement was a lack of additional funding to support it as the mandate lacked a full cost accounting of what it would take to expand data sharing across all federally funded research. In early 2024 the Association of Research Libraries published a study[34] noting that individual research institutions are already spending between $800,000 and $6,000,000 per year for the current data management and sharing practices in place. If the number of researchers sharing data increases further after the new policies come into effect in 2026, it would be reasonable to expect these costs to go up. Closer evaluation and scrutiny of data sharing practices might reasonably be expected to follow, which would increase the importance of metrics like usage, citation and reuse for funding decisions.

[34] https://www.arl.org/resources/making-research-data-publicly-accessible-estimates-of-institutional-researcher-expenses/

6.2.1.4 Summary

Table 5 summarizes the emerging needs and opportunities within the data supply chain.

TABLE 5. Data Supply Chain Gaps and Opportunities

Trend	Description	Gaps	Opportunity
Increasing awareness of the role of open data in research evaluation	Increasing frequency of funder mandates requiring open sharing of data will increase attention on the data supply chain and infrastructure needs	Standardization lacking across data repositories which creates challenges for developing new cross-stakeholder solutions	Industry groups now working to create best practices and build connections between outputs Potential to increase connections with infrastructure from other research outputs (i.e., the journals supply chain)
Capturing data sharing and reuse	New standards are emerging to better capture and report on citation and reuse of datasets	Standardization gaps among repositories (i.e., using a unique accession number instead of a DOI) makes scaling new solutions more difficult Implementation of standards requires support of many stakeholders, including publishers in the journals supply chain	Increasing frequency of data sharing mandates and resulting evaluation of open data impact increases the incentive for repositories and other stakeholders to participate in emerging standards
Appetite for expansion of data sharing with little visibility into costs	Distribution of open data is largely subsidized by institutions or research funders	Repositories often lack funding to implement new standards Costs of data storage and sharing can be high overall	Capturing usage can increase the incentives to fund necessary development for data repositories

7 Conclusions

C&E's analysis of the supply chain for open access books found that the primary challenge in developing infrastructure to improve the sharing and consolidation of usage data was the need to shift the metrics of success from "sales" (in the traditional supply chain) to "usage" (in the open access supply chain). Insofar as the open access books supply chain relies upon intermediaries whose business models are built upon the purchase and paid access to books, these suppliers need to be provided an incentive to support infrastructure that enables the flow of usage information upstream.

This is not an issue shared with the open journals or open data supply chain. In these supply chains, usage is already understood as a measure of value and the industry did not evolve from a legacy of transactional sales and distribution. Furthermore, usage is not the only measure of impact: other metrics, such as citations in other research outputs, are often valued just as much as if not more than usage by certain stakeholders.

Notwithstanding these differences, each supply chain would receive different benefits from infrastructure that connects usage and impact across platforms. We summarize our findings below with an eye towards informing future research and development for supporting infrastructures:

7.1.1 Journals Supply Chain

- Within the journals supply chain, usage infrastructure has historically been in place for supporting subscription decisions at the journal level, on a specific platform, focused on the final published version of record.
- Cross-platform distribution is less prevalent for journal articles than for books, but will grow with the rise of syndication agreements and the increase in permissive CC BY licenses applied to articles in hybrid and Gold OA journals. Changes to funder mandates also will increase the simultaneous availability of alternative article versions (such as author accepted manuscripts).
- Increases in cross-platform distribution will simultaneously increase demand to understand article usage across platforms. There are some new services beginning to offer this, such as Sensus Impact (which is in its early stages). That said there is a strong ecosystem of platform-specific analytics and reporting focused on the journal and increasingly at the article level.
- The role of usage as a measure of value for open content will depend on the importance of other legacy metrics, such as citations, which are often used as the basis for research assessment and benchmarking for researchers, institutions, and funders.

- Gaps that need to be overcome to develop a better understanding cross-platform usage and impacts include:
    - Stronger and more consistent information throughout the supply chain at the article level, using authoritative metadata (ideally generated by content creators themselves) and article-level reporting (from all platforms and distribution channels, including aggregators and repositories)
    - Consistent implementation of standards like COUNTER, to ensure usage can be accurately and fairly compared from platform to platform
    - Better connections and visibility between article versions, including between the author accepted manuscript (AAM) and version of record (VoR), and greater visibility of article reuse even in cases where a permissive license is used
    - Clarity on the value of usage information to the primary decision-makers in the supply chain, particularly in an author-pays model, where other metrics (such as citations) are already firmly established

7.1.2 Open Data Supply Chain

- The open data supply chain evolved primarily around the desire to make data available as a secondary research output, focused not on the monetization of data content but rather the goals of research transparency and compliance with funding mandates.
- Infrastructure for data sharing varies widely by field and supports a mix of discipline-specific and generalist repositories. Because content is not monetized, the data distribution supply chain lacks many of the stakeholders that support the business of books and journals publishing and distribution (such as publishers, distributors, and sales channels).
- Existing initiatives, such as DataCite / Make Data Count, the Generalist Repository Ecosystem Initiative, and IRUS seek to identify the commonalities and build infrastructure that can span across data repositories. These include efforts to support better capture and consolidation of usage information about data repositories and datasets.
- The challenges such initiatives will have to overcome include:
    - Many repositories are operated as public resources with limited funding models beyond institutional support. These repositories, although they contain robust metadata about each dataset, would require greater resourcing to adopt more industry standards such as DOIs, COUNTER compliance, and so on.
    - Appropriate use of standards is also necessary from stakeholders outside the supply chain, such as journal publishers, to ensure research outputs appropriately capture citations of related data.

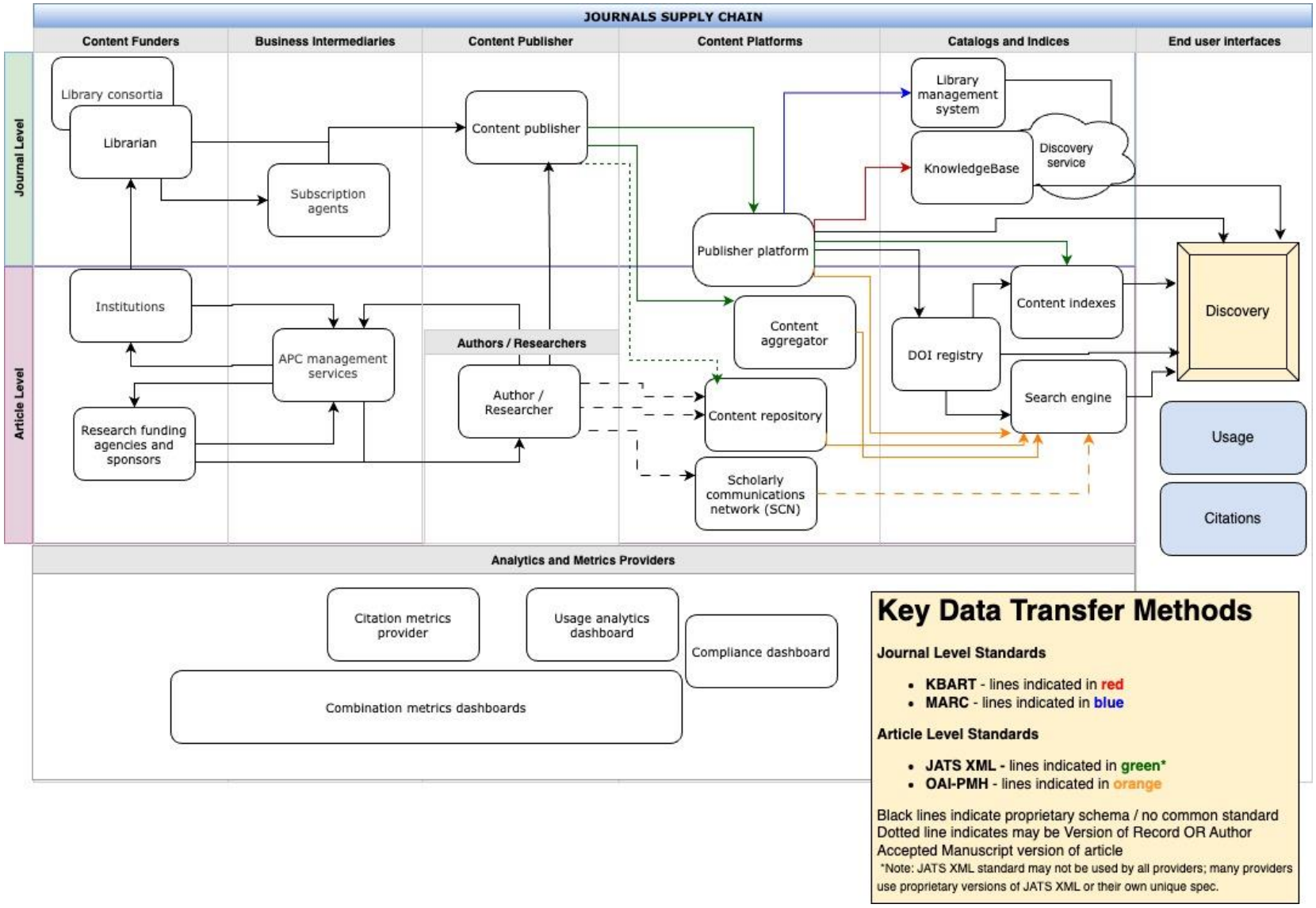
Creating infrastructure to support better usage and impact sharing for open journals and open data will require a collaborative approach focused on these supply-chain specific challenges. This means that just as solutions focused on the books ecosystem need to be fit-for-purpose and designed with books in mind, a solution that addresses the journals or data ecosystem should likewise be expected to address different challenges, and accommodate different business needs. To fully extend a data space like the Open Access eBooks Usage Data Trust to serve journals and open data, additional work would be necessary with each set of respective stakeholder groups to better understand their unique problems and possible solutions (both technical, and strategic).

Yet as with books, both journals and open data share the challenge of determining which "metrics of success" will matter in a new open environment. For open access books, where "sales" no longer apply, usage is a natural place to look to gauge the value of research. For journals and data, usage and citations are both valued as measures, and the challenge is capturing this information for open resources in the appropriate way and for the right stakeholders. The need for additional infrastructure is ultimately defined by the appetite for those "upstream," especially those who fund and create the research, to better understand usage of journal articles (including across versions) and datasets. Cultivating this interest and defining the value of usage as a metric among these stakeholders is just as important a criteria of success as building the tools to gather and analyze the data.

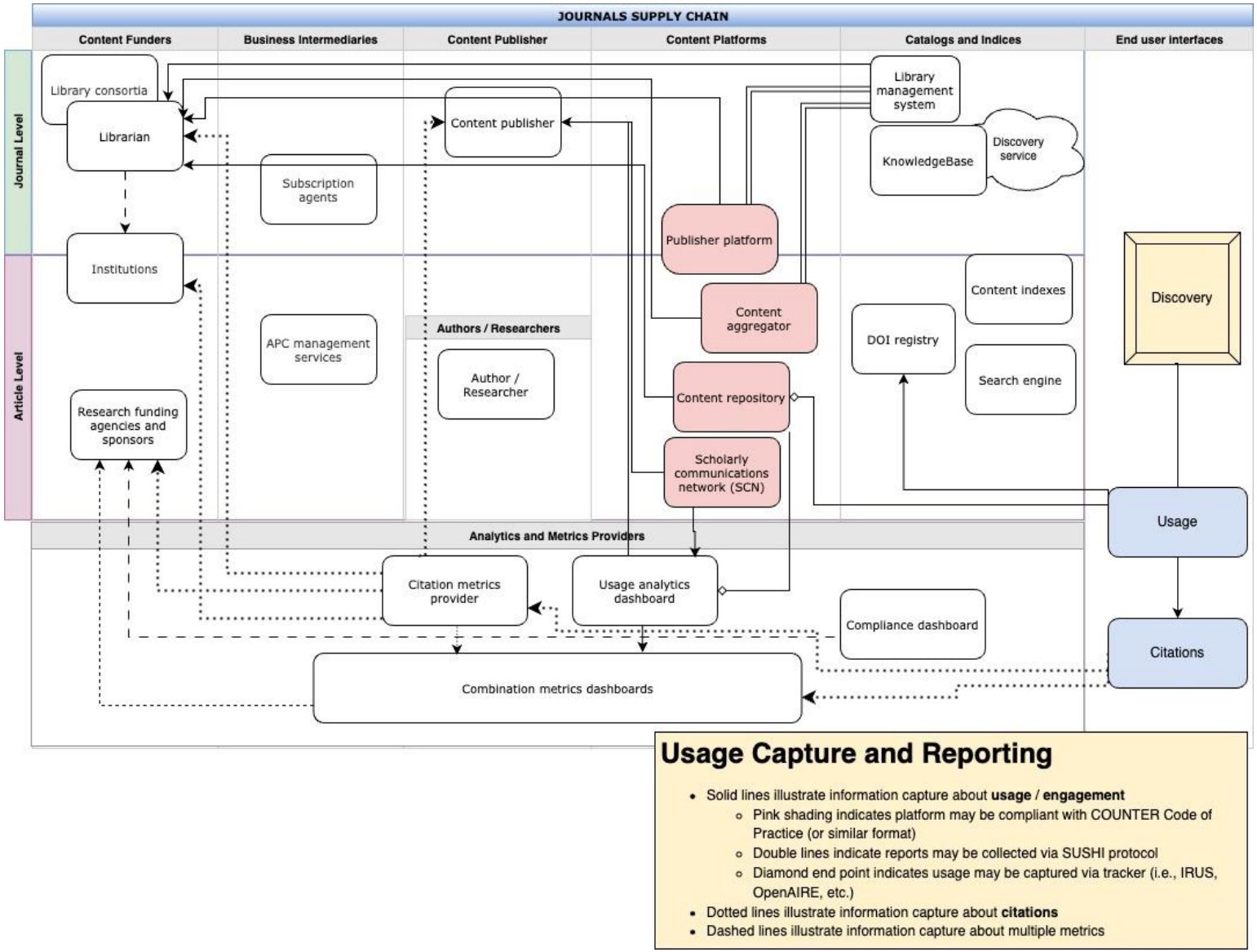
Appendix A: Open Journals and Data Supply Chain Maps

APPENDIX A.1: SUPPLY CHAIN DIAGRAMS

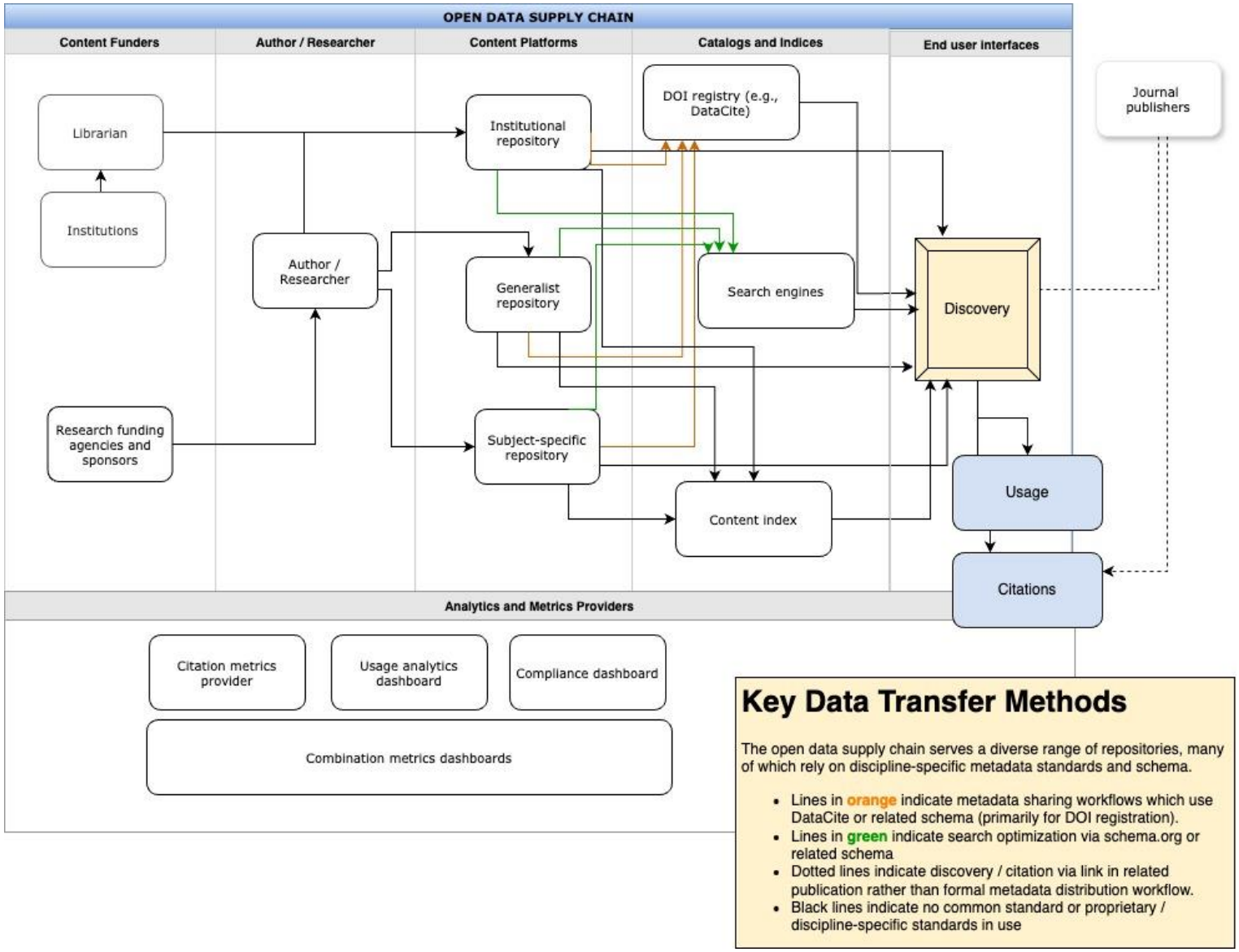
Journals Supply Chain – Downstream: Distribution to Discovery



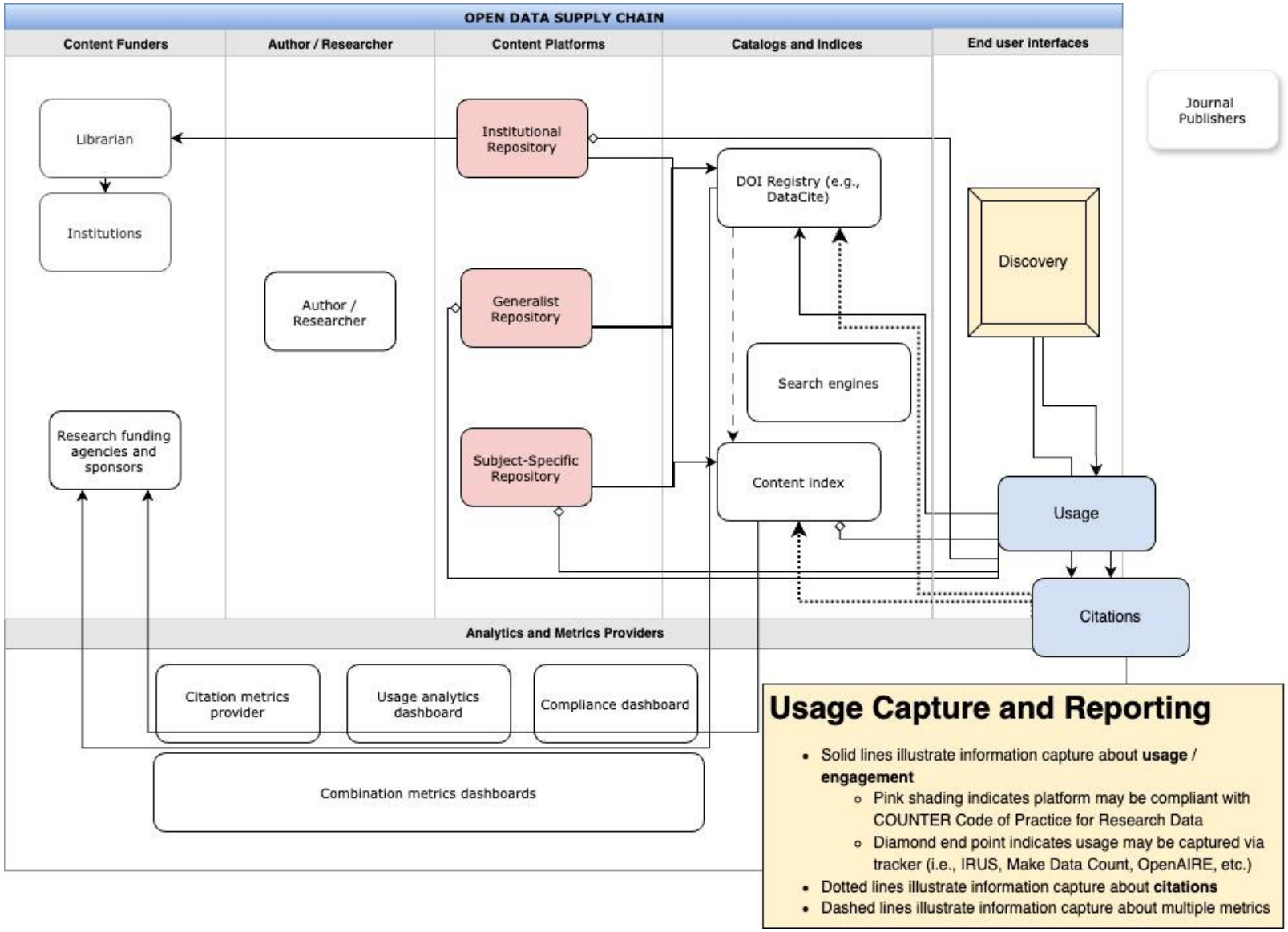
Journals Supply Chain – Upstream: Discovery to Assessment



Data Supply Chain – Downstream: Distribution to Discovery



Data Supply Chain – Upstream: Discovery to Assessment



APPENDIX A.2: SUPPLY CHAIN COMPARISON

The distribution and evaluation of supply chains for open books, journals, and data each have their own unique characteristics and terminologies, which have been reflected in their respective supply chain maps. Below we provide a brief description of the key similarities and differences between supply chains to allow easier cross-comparison. See also **Section 3.3, "**Landscape Overview Summary and Comparison**"** for a higher-level comparison of business dynamics underlying each supply chain.

A.2.1. Stakeholder Types

Shared Stakeholder Types	Unique to Books	Unique to Journals	Unique to Data
• Research funders / sponsors; librarians; institutions • Authors/researchers • Content publishers (journals and books only) • Publisher platform (journals and books only) • Content aggregator (journals and books only) • Content repositories • DOI registries (Crossref for journals and books; DataCite for data) • Content indices; search engines • Knowledgebases; library management systems (journals and books only) • Citation metrics providers • Compliance dashboards	• Book distributors • Consumer ebook platform • Consumer sales channels • Library sales channels • Ebook viewing apps / devices and interfaces	• Subscription agents • APC management services • Scholarly communications networks (SCNs)	• Content publishers are typically not involved in the process of storing open data in repositories • In addition to more general content repositories, open data may be shared in discipline-specific repositories focused on sharing data

A.2.2. Metadata Distribution Standards

Shared Metadata Distribution Standards	Unique to Books	Unique to Journals	Unique to Data
• KBART (for distribution to knowledgebases; journals and books only) • MARC Records (for cataloging in library management systems; journals and books only) • OAI-PMH enables discoverability for all resource types	• ONIX • BITS XML (similar to JATS XML; less commonly used)	• JATS XML and related schema	• Many discipline-specific metadata schema are not applied to other publication output types • DataCite DOI registration schema • Data repositories use a specific schema.org metadata schema for Datasets

A.2.3. Evaluation and Metrics

Shared Evaluation standards and metrics	Unique to Books	Unique to Journals	Unique to Data

• Usage, following the COUNTER Code of Practice / Code of Practice for Datasets
• Protocols/services like SUSHI and IRUS may be used to collect usage reports for any content type hosted on appropriate platform
• Citations (primarily for journals and data)
	• Sales (for paid access / print versions), including
	  ○ Library channel sales
	  ○ Consumer channel sales
	• Usage generated on consumer ebook platforms which are not COUNTER-compliant (i.e., Amazon Kindle)
		• Cross-disciplinary normalized measures of citation performance (i.e., Journal Impact Factor)
		• Compliance with funder mandates (shared with books but influence is stronger for journals due to higher exposure to funder mandates)
			• Re-uses and adaptations

A.2.4. Other Key Attributes Documented in Supply Chain

Shared Key Attributes	Unique to Books	Unique to Journals	Unique to Data
• n/a	• *Unique identifiers*: While books may use DOIs issued by Crossref at the title or chapter level, books primarily use the ISBN as a unique product identifier used throughout the supply chain. • *Formats:* Distribution channels may sell print editions alongside open monographs. • Titles may have multiple ISBNs (i.e., a unique ISBN for each format).	• *Versioning:* Open journals supply chain supports the Version of Record and Author Accepted Manuscript • Distribution and evaluation may be either at journal- or article-level	• No common format for open data; format is determined by the experiment conducted and instruments / software used

Appendix B. Stakeholder Index

Open Journals and Data Supply Chain Mapping - Stakeholder Index				
Category	**Stakeholder**	**Role in Open Journals / Data Supply Chain**	**Open Journals / Data Examples***	**Supply Chain**
Content Funders	**Research funding agencies and sponsors**	Organization providing funding support for research activities and outputs. Can be further distinguished between the funder of the research activity and the funder providing payment for publication (i.e., in the author-pays Gold OA model).	• National Science Foundation (NSF) • National Institute of Health (NIH) • European Research Council • Bill & Melinda Gates Foundation	• Journals • Data • Books
	Institution	Organization that employs researchers and provides facilities to conduct research. Sub-types include academic institutions, research institutes, government labs, commercial organizations, nonprofits, and others.	• Scripps Research Institute • Lawrence Berkeley National Laboratory • CERN • Howard Hughes Medical Institute	• Journals • Data • Books
	Libraries	Historically the purchaser of research publications but increasingly allocating budgets to support open access publication methods (i.e., providing funds for Gold OA publication, supporting Subscribe to Open models). May also be responsible for institutional repository funding and management.	• Individual libraries (i.e., Massachusetts Institute of Technology, University of Michigan) • Library consortia (i.e., JISC, , OHIOLINK, PALCI)	• Journals • Data • Books
Business Intermediaries	**Subscription Agents**	Businesses that facilitate sale and renewal of subscriptions to institutions and libraries. Increasingly providing additional decision support and handling for open access agreements (i.e., Subscribe to Open, Transformative Agreements)	• EBSCO • Harrassowitz • Charlesworth Group • iGroup • Kinokuniya	• Journals

Open Journals and Data Supply Chain Mapping - Stakeholder Index				
Category	**Stakeholder**	**Role in Open Journals / Data Supply Chain**	**Open Journals / Data Examples***	**Supply Chain**
Business Intermediaries (continued)	**APC Management Services**	Intermediaries that facilitate the collection of article publication charges (APCs) for journals offering the author-pays option (i.e., hybrid titles, Gold OA titles).	• Copyright Clearance Center Rightslink • ChronosHub	• Journals
Authors / Researchers	**Authors / Researchers**	Original creator of the article or dataset. Can be further distinguished into various contributor roles (i.e., via the CREdiT taxonomy). Also may be distinguished by corresponding authorship (often the author responsible for APC payment in Gold OA model and which determines eligiblility for publication waiver / coverage under Transformative Agreements).	Any author or researcher falls into this category	• Journals • Data • Books
Content Publishers	**Publisher**	Entities who facilitate peer review and publication of research articles (specifically the Version of Record) and initiate distribution via the supply chain. Under this category there are business ranging from large to small, commercial, not-for-profit, society, library publishers, and scholar-led presses	• Elsevier • Springer Nature • Oxford University Press • Rockefeller University Press • Cold Spring Harbor Lab Press • American Chemical Society	• Journals • Books
Content Platforms	**Publisher Platform**	Platforms that are dedicated to hosting content of a specific publisher. Typically the home of the article Version of Record. Publishers may have multiple platforms for different audiences (i.e., a member site and institutional subscription site, or different sites for different publisher brands).	• ScienceDirect • SpringerLink • BMC • Oxford Academic • American Society for Microbiology Journals Platform	• Journals • Books

Open Journals and Data Supply Chain Mapping - Stakeholder Index				
Category	**Stakeholder**	**Role in Open Journals / Data Supply Chain**	**Open Journals / Data Examples***	**Supply Chain**
Content Platforms (continued)	**Content Aggregator**	Content delivery platform that aggregates articles and other content across multiple publishers. Typically serves B2B markets (institutions and libraries) although some aggregators host OA-specific collections on their platform.	• EBSCOhost • ProQuest • JSTOR • Project MUSE • BioOne • GeoScienceWorld	• Journals • Books
	Scholarly Collaboration Network (SCN)	Social network for academic audiences. Individuals and journals / brands can create profiles and generate engagement as well as share content (either Version of Record or Author Accepted Manuscript). Increasingly a platform for journal publishers to syndicate access to the Version of Record.	• ResearchGate • Academia.edu	• Journals
	Content Repository	Platform where content can be deposited by author or publisher, often freely available to access. For journal articles, version may be the Version of Record or Author Accepted Manuscript. Repositories can be further distinguished as *general, subject-specific,* or *institutional*.	*General*: • Dryad • Figshare *Subject-specific*: • PubMed Central • National Climatic Data Center (NCDC) *Institutional:* • Deep Blue (University of Michigan)	• Journals • Data • Books

Open Journals and Data Supply Chain Mapping - Stakeholder Index				
Category	**Stakeholder**	**Role in Open Journals / Data Supply Chain**	**Open Journals / Data Examples***	**Supply Chain**
Catalogs and Indices (cont'd)	**KnowledgeBase**	Extensive catalog of content metadata shared across libraries and institutions. 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 Knowledge Base • EBSCO Knowledge Base (KB) • Alma Central KnowledgeBase (CKB)	• Journals • Books
	Content Indexes	Specialized product for metadata curation and discovery; can include topic-specific A&I databases or other curated directories. Developed to augment structured metadata for more refined search capabilities and thereby promote discoverability for relevant users.	• Scopus • Web of Science • Directory of Open Access Journals (DOAJ) • Unpaywall • Digital Science Dimensions • Data.gov • Kaggle	• Journals • Books • Data
	Search Engine	Search engines developed to index and retrieve content from across the web; important discovery mechanism for users especially outside of a library context.	• Google Scholar • Google Dataset Search	• Journals • Books • Data
	DOI Registries	Issuing body for unique Digital Object Identifiers (DOI). The primary DOI registry for journals is Crossref, while the primary DOI registry for open data is DataCite.	• Crossref • DataCite	• Journals • Books • Data
Analytics and Metrics	**Citation Analytics Providers**	Provider which creates analytics regarding citation behavior for research outputs. Often affiliated or overlapping with content indexes but may be distinguished as providing *decision support* rather than enhancing discovery.	• Journal Citation Reports (JCR) • Scopus • Google Scholar Metrics	• Journals • Data

Open Journals and Data Supply Chain Mapping - Stakeholder Index				
Category	**Stakeholder**	**Role in Open Journals / Data Supply Chain**	**Open Journals / Data Examples***	**Supply Chain**
	Usage Analytics Providers	Provider which offers visualization and analytics focused on usage behavior for research outputs.	• OpenAIRE PROVIDE • ResearchGate Journal Home	• Journals • Data
	Compliance Reporting	Provider which provides visualization and analytics focused on compliance with open publication mandates or principles, to serve funders, authors,	• CHORUS • Dataseer.ai	• Journals • Data
	Combination Metrics Dashboards	Provider which combines multiple analytics and visualizations covering citations, usage, compliance, and other factors.	• Sensus Impact • EBSCO Panorama • Altmetric • Make Data Count	• Journals • Data

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. Examples given in the table apply to **journals and **data** supply chains only.*

Appendix C: Interview Question Template

Interview Structure

- We use a *semi-structured* format for interviews, where we touch on specific questions but do 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.
- The questions you see below are phrased as questions, but again, will not be read as a script – therefore, don't worry about wordsmithing them. We will adapt them on the fly in interviews, to fit the role and specific circumstances of the interviewee, as well as the direction the interview naturally progresses.
- Questions were adapted for each interviewee based on primary role and organization.
- Interviews are planned for 60 minutes; however, if someone can't set aside the full time, we'll take what time we can get and focus on priority questions.

Introduction

Introductory remarks to be shared with each 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.
- C&E are conducting research to document the open journals and data supply chain, from metadata description to distribution to the capturing of usage data. This research follows the work we conducted in 2020-2021 to map the supply chain for open access books, and is funded by a grant from the National Science Foundation (NSF).
- We've scheduled 60 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.

Journals Interview Template

INTRODUCTION

1. Tell us about yourself and your role at [organization].

2. Tell us about your OA journals program/activities.
    a. What "type" of OA do you offer / support? (Author-pays gold, green, subscribe to open, transformative agreements)?
    b. Who are your customers or "buyers"?
    c. What other stakeholders do you serve?

3. What is the flow of the "supply chain" when it comes to OA journal articles, and where does your part fit? Who do you receive information/data from, and who do you supply information/data to?

4. How do you measure the success of your OA journals program? What are the important metrics to measure that success? Do you share any data on these metrics with other stakeholders?

DISTRIBUTION AND DISCOVERY

5. What is your approach to the distribution and discovery of OA articles (specifically the Version of Record)? Is this different than for paid-access articles and journals?
   Prompt for:
    a. What is the role of aggregators in distributing OA articles?
    b. Are there benefits to the VoR appearing on multiple platforms?
    c. Are there risks to the VoR appearing on multiple platforms?
    d. Are there intermediaries (indices, A&I databases) you work with to increase discoverability on your own platform?
    e. Do you connect versions of an article? For example, the author's accepted manuscript to the final published version of record? Or a preprint to the published article? What article versions do you connect? How is the connection made?

6. Do you connect open data and supplemental files associated with an article via metadata elements? How?

7. How are open access articles distinguished from other types of access (paid access, free-to-read)? What metadata elements are used? Is everyone following the same standard? Are there any gaps in OA status being made available in machine readable way to platforms / intermediaries?
    a. Do you capture/provide information about license type in article metadata?

8. What are some of the challenges to increasing the reach and discovery of OA articles? What improvements would help to overcome those challenges?

USAGE REPORTING

9. Do you ***capture*** usage of OA articles on your platform? Do you distinguish usage of OA articles vs. paid access articles? What about OA articles vs. "free to read" articles?
    a. If so - what kind of usage information do you capture? [[e.g., COUNTER, Google Analytics data, other]]
    b. What is the value of that OA usage data to you? How do you use it to make decisions or to support other stakeholders?

10. Do you ***provide*** OA article usage data to other organizations or stakeholders? In what format?
    a. What is the value of OA usage data to those stakeholders?
    b. What is the value to you in sharing that information?

11. Do you ***receive*** OA article usage information from other stakeholders? In what format?
    a. How do you use that data? What is the value to you?

12. Is confidentiality or privacy important as this information is passed down the supply chain?

13. Are there any platforms where OA articles may be used where you do ***not*** have access or visibility into usage? Would having this usage be valuable to you?

14. What role does OA usage information play overall in evaluating the success of your program (both within your business and for your customers and stakeholders)?

    a. Do you consider usage of the Version of Record only in your evaluation, or do you consider usage of Author Accepted Manuscript or Preprint versions of articles?

15. What are the challenges in capturing usage information about OA articles?

16. What are the opportunities in capturing better usage data for OA articles?

FINAL THOUGHTS

17. What have we not asked about that is an important component of the supply chain for OA articles?

18. As part of the project, we are speaking with these categories of stakeholders: publisher, aggregator, index/database, analytics provider, infrastructure provider

    a. Are we missing any categories of stakeholder?
    b. Are there any specific people or organizations you would recommend we speak with as part of our research?

Open Data Interview Template

INTRODUCTION

1. Tell us about yourself and your role at [organization].

2. Tell us about your open data program/activities.
    a. What is your primary business model? (author pays, diamond OA, other)
    b. Who are your customers or "buyers"?
    c. What other stakeholders do you serve?

3. What is the flow of the "supply chain" when it comes to OA data, and where does your part fit? Who do you receive information from, and who do you supply information to?

4. How do you measure the success of your data sharing program? What are the important metrics to measure that success? Do you share any data on these metrics with other stakeholders?

DISTRIBUTION AND DISCOVERY

5. What is your approach to the distribution and discovery of open data? *Prompt for:*
    a. Are datasets typically available on more than one platform?
    b. Are there intermediaries (indices, A&I databases) you work with to increase discoverability?
    c. Do you connect open data to associated research articles? How is the connection made? What article versions do you connect?

6. What pieces of information (aka metadata elements) are most important to describe each dataset? What metadata elements are used? Is everyone following the same standard? Are there any gaps in OA status being made available in machine readable way to platforms / intermediaries?

7. What are some of the challenges to increasing the reach and discovery of datasets? What improvements would help to overcome those challenges?

8. Is confidentiality or privacy important as this information is passed down the supply chain?

USAGE REPORTING

9. Do you ***capture*** usage of open datasets on your platform?
    a. If so - what kind of usage information do you capture? [[e.g., COUNTER, Google Analytics data, other]]
    b. What is the value of that usage data to you? How do you use it to make decisions or to support other stakeholders?

10. Do you ***provide*** open data usage data to other organizations or stakeholders? In what format?
    a. What is the value of usage data to those stakeholders?
    b. What is the value to you in sharing that information?

11. Do you ***receive*** open data usage information from other stakeholders? In what format?
    a. How do you use that data? What is the value to you?

12. Are there any platforms where datasets may be used where you do ***not*** have access or visibility into usage? Would having this usage be valuable to you?

13. What role does usage information play overall in evaluating the success of your program (both within your business and for your customers and stakeholders)?

14. What are the challenges in capturing usage information about open datasets?

15. What are the opportunities in capturing better usage data for open datasets?

16. Is confidentiality or privacy important as this information is passed down the supply chain?

FINAL THOUGHTS

17. What have we not asked about that is an important component of the supply chain for OA data?

18. As part of the project, we are speaking with these categories of stakeholders: publisher, aggregator, index/database, analytics provider, infrastructure provider
    a. Are we missing any categories of stakeholder?
    b. Are there any specific people or organizations you would recommend we speak with as part of our research?

Journals / Data Supply Chain Infrastructure Interview Template

INTRODUCTION

1. Tell us about yourself and your role at [organization].

2. Tell us about your program/activities to support open journals and data.
    a. Who are your primary stakeholders for journals?
    b. Who are your primary stakeholders for open datasets?
    c. What other stakeholders do you serve?

3. What is the primary problem that you help open journals and data repositories solve?

DISTRIBUTION AND DISCOVERY

4. What pieces of information (aka metadata elements) are most important to describe each article or dataset? What metadata elements are used? Is everyone following the same standard? Are there any gaps in OA status being made available in machine readable way to platforms / intermediaries?

    a. [if appropriate] How do stakeholders distinguish usage of OA articles vs. paid access articles? What about OA articles vs. “free to read” articles?

5. What are some of the challenges in the distribution and discovery of OA articles and datasets? What improvements would help to overcome those challenges? Which stakeholders are best positioned to make those improvements?

6. To what degree does the infrastructure currently support the linking of different items (such as different article versions, a dataset to a research article, the AAM to the VoR)? How is the connection made?

USAGE REPORTING

7. What is the value of understanding article or data usage among stakeholders? How does usage information influence their decision-making and business decisions?

8. How does the importance of usage information differ between the Version of Record and previous versions (Author Accepted Manuscripts)? Does usage of the AAM matter?

9. What are the challenges in capturing OA article / data usage information across stakeholders?

10. What are the opportunities in capturing better OA article / data usage information across stakeholders?

11. Is confidentiality or privacy important as this information is passed down the supply chain? How does the infrastructure support confidentiality and privacy?

FINAL THOUGHTS

12. What have we not asked about that is an important component of the supply chain for OA journals and data?

13. As part of the project, we are speaking with these categories of stakeholders: publisher, aggregator, index/database, analytics provider, infrastructure provider

    a. Are we missing any categories of stakeholder?
    b. Are there any specific people or organizations you would recommend we speak with as part of our research?