Research Ideas and Outcomes 5: e38698
doi: 10.3897/rio.5.e38698

Review Article

A literature review of scholarly communications metadata

Will James Gregg[‡], Christopher Erdmann[§], Laura A D Paglione[|], Juliane Schneider[¶], Clare Dean[#]

‡ Simmons University, Boston, United States of America
§ California Digital Library, Oakland, United States of America
| Spherical Cow Group, on behalf of Metadata 2020, New York, United States of America
¶ Harvard Catalyst | Clinical and Translational Science Center, Boston, United States of America
# Public Library of Science, San Francisco, United States of America

Corresponding author: Will James Gregg (willjgregg@gmail.com)

Reviewed v1

Received: 31 Jul 2019 | Published: 05 Aug 2019

Citation: Gregg WJ, Erdmann C, Paglione LAD, Schneider J, Dean C (2019) A literature review of scholarly communications metadata. Research Ideas and Outcomes 5: e38698. https://doi.org/10.3897/rio.5.e38698

Abstract

The purpose of this literature review is to identify the challenges, opportunities, and gaps in knowledge with regard to the use of metadata in scholarly communications. This paper compiles and interprets literature in sections based on the professional groups, or stakeholders, within scholarly communications metadata: researchers, funders, publishers, librarians, service providers, and data curators. It then ends with a 'bird's eye view' of the metadata supply chain which presents the network of relationships and interdependencies between stakeholders. This paper seeks to lay the groundwork for new approaches to present problems in scholarly communications metadata.

Keywords

Scholarly communication, metadata, metadata supply chain, research data management, Metadata 2020

Introduction

The network of those who prepare and publish scholarly research is made up of individuals who work in many different professions, disciplines, and institutions worldwide. They share a common interest, however, in that “All of those who are involved with the publication of scholarly works have the same end goal: to conduct, facilitate, and/or communicate research” (Kemp et al. 2018). Metadata 2020*1 is a group that advocates for rich, reusable, and open metadata for scholarly research in all outputs, and seeks to clearly establish the cost of inadequate metadata. This literature review, authored on behalf of Metadata 2020, compiles, summarizes, and interprets literature on the topic of metadata in scholarly communications in order assist those who work with it to better understand the challenges and opportunities at hand.

After laying out definitions for central terms, this review organizes the literature into different categories based on professional group: researchers, funders, publishers, librarians, systems and service providers, and data curators. It then presents a subset of the literature, which attempts to capture the bigger picture: the network of relationships and interdependencies among all of these groups. The final section discusses new standards, guidelines, and initiatives for improving metadata and offers preliminary suggestions for solving current problems with the scholarly communications metadata supply chain.

Methods

The idea for this literature review originates with Metadata 2020’s Researcher Communications project. A number of the members of this project identified a need for a comprehensive review of the challenges, opportunities, and gaps with metadata in scholarly communications with the aim that it would foster further conversations among the stakeholders involved. The review itself was researched and written from August of 2018 to April of 2019. Initial structure and guidance was provided by a sub-group of Metadata 2020 participants, after which the author performed literature searches, summarized articles, and categorized them based on their research topic, output type, and relevance to particular stakeholder groups. The author sought to limit the review to resources published within the last 10 years and attempted to include literature which was written by or about each stakeholder group. In meeting this goal, the review sought resources from outside traditional venues for scholarly publishing; it includes a number of white papers, blog posts, newsletters, videos, and other grey literature (literature that does not pass through traditional academic publishing channels). Throughout the process, the author regularly collaborated with participants of Metadata 2020 in seeking out literature and structuring the content of the review. In March of 2019, a draft of the review was edited by a small number of Metadata 2020 participants. In April, the review was published and opened to comment via RIO journal.

Definition of core terms

This review assumes at least a basic knowledge of how metadata records are created, shared, and used. Nevertheless, two core terms are defined below to frame the discussion.

Scholarly communication is the process by which research is conducted, transformed into content, and distributed to a wider audience. The majority of the resources featured in this literature review are concerned with the physical and life sciences. The implications of this review, however, can be extended in many cases to the social sciences and humanities which have the same need to describe and share scholarly works. The majority of resources also treat the published journal article as the primary form of scholarly output, though emerging literature increasingly focuses on research data and the research process as opposed to finalized articles.

Metadata in this context is the information that accompanies the various stages and outputs of research. Common to most scholarly research are metadata elements such as author, date, title, subject, language, and standard identifier. In the case of research data, metadata describes specialized aspects such as the geographic location where the data was collected, the name and identifier of the research funder, the institutional affiliation of the researchers, contributors such as editors and data curators, or the number of the grant award funding the research (DataCite Metadata Working Group 2016). Metadata for journal article includes identifiers such as the International Standard Serial Number (ISSN, the numeric identifier assigned to journals), volume, issue, page numbers, or epub identifier. In the case of monographs (books), specialized metadata includes the International Standard Book Number (ISBN), title, page number, and author of individual chapters. These and other metadata elements are essential for disseminating research outputs. Metadata gives proper credit to research funders, ensures that research can be reproduced, facilitates others in finding research, allows authors to build a portfolio, and measures the scholarly impact of an article.

Stakeholders

Publishers, service providers, researchers, funders, librarians, and data curators are the stakeholders of metadata in scholarly communication. The relationships between these stakeholder groups have serious implications for scholarly communication metadata. For example, funders have requirements for researchers, stipulating that the source of the funding must be stated in their research outputs. Publishers ask researchers to supply metadata about themselves and the subject of their projects during the submission, editorial and publication process. Vendors, who package content from publishers and re-sell it to libraries and others, use publisher metadata to keep track of their products including ebook and journal articles. Librarians then use and enhance metadata supplied by publishers and vendors to make their articles discoverable.

The “supply chain” of metadata is best elaborated in the literature by studies that examine stakeholders’ relationships with one another. Bull and Quimby (2016) and Bascones and Staniforth (2018), for example, focus on librarians and end users at libraries. They profile problems that arise when libraries take in poor metadata from publishers and service providers and also elaborate challenges that libraries face in managing metadata for thousands of electronic resources. They conclude that, while these stakeholders are co-dependent, the existing channels for resolving problem metadata are limited. Flynn (2013) analyzes the relationships among librarians, publishers, and service providers in terms of how each would stand to benefit from open access metadata, defined as "bibliographic information describing library content that is openly licensed and freely accessible" (p. 29). Meanwhile, Mercer and Dyas-Correia (2011) suggest ways that newly-formed journals could benefit from working with librarians. Librarians "can choose to act as distributors or publishers via an institutional repository," provide information on acquiring ISSNs or digital object identifiers (DOI), help to modify or enhance machine-readable cataloging (MARC) records, and help to explain funder or open access policies (p. 239). Meanwhile, de Moor and van Zanden (2008) find opportunities among data archives, journals, and groups of researchers. For instance, researchers who agree to share data (known as researcher "rcollaboratories") could partner with data archives from the outset to ensure long-term preservation of high-quality data. Journals with a data availability policy (DAP) could similarly benefit from partnering with data archives.

Examinations of the researcher-funder relationship are also present in the literature, most frequently revolving around discussions of research data management (RDM) and sharing. Some propose best practices for RDM and data sharing for publishers, funders, and researchers (Michener 2015; Dietrich et al. 2012, p. 3) while others evaluate the problems associated with current practices: Dietrich et al. (2012) surveyed 22 data management policies from 10 funding organizations and found that data policies overall were lacking a “significant number of elements” from the quality rubric established by the researchers. 11 of the policies included fewer than half of the elements on the rubric (2012, p. 4). Poole (2016) writes that researchers must comply with funder requirements which often ignore important "inter-domain differences in research practices" (p. 971). In addition he argues that funders often emphasize quantity of data without considering the likelihood it will be reused or building infrastructure for digital curation. Couture et al. (2018) find that funder data publication requirements rarely result in actual sharing, bringing home the practical consequences of incomplete policies and researcher confusion.

Studies of the relationship between librarians and researchers are also present but somewhat less prevalent. Two authors who examine this relationship in depth, Marsh (2014) and Jennings (2017), do so with a focus on institutional repositories (IRs). An IR is a service, most often operated by librarians in an academic institutions, which seeks to collect, maintain, and distribute the intellectual products of that institution's researchers. Marsh reviews literature and current practice with regard to IRs and finds that, while IRs are proliferating, they are slow to populate. One barrier to access is low engagement by faculty at the institution, possibly because of lack or hard or soft requirements, confusion about its purpose, or a more fundamental division between researchers and information

professionals: "academics . . . are in some cases even reluctant to dismantle a structure which has enabled them to be successful and in which they continue to thrive as authors, advisors and editors" (p. 166). Jennings profiles the creation of the Research Data Archive at the University of Bath using EPrints, an institutional repository software and describes the process by which researcher-provided metadata is imported into the IR. In implementing a rich metadata schema for the description of research data, Jennings found that "our guidance for using the fields has many caveats and options, which actually makes it harder for researchers to use the system" (p. 6). These articles illustrate the problems that can result when there is a disconnect between one stakeholder group and another with regard their metadata needs and expertise.

Yet other research has focused on metadata problems and opportunities for single groups, such as the call of McNutt et al. (2018) for publishers to adopt the CRediT (Contributor Roles Taxonomy) standard for describing scholarly contributions or the study by Kratz and Strasser (2015), “Researcher perspectives on publication and peer review of data.”

The articles referenced above and many others share a method of grouping stakeholders by professional or institutional identity. Perrier et al. (2017), for instance, divide stakeholders into professional groups (researchers, administrators, participants, and funders) in their survey of research data management (p. 9). Another formulation of stakeholders is possible, in that those who work in scholarly communications could define themselves by the metadata roles they have in common. For instance, rather than the terms “publisher” or “author,” these individuals could think of themselves as metadata creators, and thus highlight their co-dependencies with metadata stewards (custodians) and consumers.*6 This review, however, follows the example of the majority of literature in grouping stakeholders by profession rather than metadata role. By doing so, it hopes to present the literature in a way that will be approachable to a wide audience while raising alternative ways of viewing stakeholder relationships. Below, this review presents challenges, opportunities, and gaps for publishers, service providers (vendors), researchers, funders, librarians, and data curators. In doing so, it seeks to outline states of practice, demonstrate incentives for improvement, and suggest areas for future research within each group.

Publishers

Academic publishers distribute scholarly research by making journal articles, books, theses, and data available online or in print. Though publisher metadata is sometimes criticized, metadata’s return on investment is increasingly recognized and publishers are projected to invest more in quality metadata. Publishers are also on the forefront of adopting emerging technologies for automatic generation of metadata including full-text semantic analysis. Other stakeholders stand to benefit from hearing from publishers directly about their day-to-day practices.

Challenges

Of all the stakeholder groups, calls to improve metadata quality fall perhaps most frequently on publishers. Publishers could “clean up the information that they provide to vendors about their items, which would help vendors create higher quality records” (Flynn 2013, p. 30), and thus also benefit those who use vendors’ services. Fault is also found with vendors’ and publishers’ semiautomated metadata collection techniques, which can leave out core information. One case study profiles the difficulties encountered by librarians due to the lack of ISSN as a standard piece of metadata in “unsolicited submissions or in collected aggregator/publisher notifications”; without it, it is much more difficult to determine the status of access restrictions (Bull and Schultz 2017, p. 14). Another finds that publishers “could hire professionally-trained catalogers on staff to create high-quality cataloging records for their items,” though there is perhaps a more direct incentive to do so for vendors than publishers (Flynn 2013, p. 30).

Opportunities

On the other hand, there is an observable trend of publishers giving metadata more serious consideration. According to a 2017 survey of industry leaders, 90% of all publishers are planning to invest in metadata over the next three years (Imbue Partners 2017, p. 11). Publishers are looking to take this action despite budget concerns and the fact, acknowledged by Kemp et al. (2018) that “... enriching existing metadata records can be difficult and time consuming” (p. 208).

The reason for this shift in focus is in part due to the incentives that have been associated with good metadata. Two white papers by Nielsen, a publisher with a presence in the US and UK, found increased sales to be associated with quality metadata. Specifically, a quantitative analysis on the sales of Nielsen’s top 100,000 book titles in the period between July 2015 and June 2016 showed higher sales for titles which included basic metadata, a set of 9 elements from the Book Industry Communication standard: ISBN, title, format/binding, publication date, Book Industry Standards and Communications (BISAC) subject code, retail price, sales rates, cover image, and contributor. Titles with these elements had sales 75% higher than those with incomplete metadata. The study also found that supplemental descriptive metadata such as author biography, reviews, and title description, give a boost to sales. On the whole, each element of descriptive metadata boosts sales, with an increase of 72% for titles having 3 descriptive elements over those having none. (Walter 2016, p. 9). The fact that “metadata was considered of highest importance in digital transformation – a 4.6 out of 5” by industry leaders, combined with the fact that metadata “had the second lowest rating for current organizational ability – a 2.0 out of 5,” suggests that publishers will invest more in metadata in coming years (Imbue Partners 2017, p. 6).

Moreover, investments in developing technologies for metadata promise substantial rewards. Some publishers are exploring AI-generated keywords and abstracts as well as chapter-level metadata which may lead to more granular search functionality (Tan 2018, p. 20). Ruth Pickering of Yewno sees possibilities in the developing area of full-text semantic

analysis. Line-by-line subject analysis of publications, also known as semantic analysis, will allow publishers to generate categorization and keyword metadata for backlogs of poorly cataloged publications (Silverchair Information Systems 2018, 18:00). While semantic analysis will likely not replace traditional metadata generation and search techniques, it might also help to overcome inconsistencies among publishers (Silverchair Information Systems 2018, 20:00). In a test of this technology, Westchester Publishing Services applied semantic analysis to a backlog of 200 academic book titles, hoping to "'... create discoverability-focused metadata tags and other supporting metadata and content to add value for these titles when they are placed online'" (Tan 2018, p. 30). The results of this effort have not been reported. Similarly, AIP Publishing is using a semantic tool from Access Innovations to generate keyword metadata for articles focused on the physical sciences (Society for Scholarly Publishing 2018, 16:17). Semantic analysis may have wider implications for the publishing industry and scholarly communications as a whole. In a panel discussion at Davos 2016, Stewart Russell of University of California at Berkeley predicted that "If the search engine industry is worth a trillion dollars roughly right now, this new level of technology [semantic analysis] could be worth 10 trillion because it will have so many more applications" (World Economic Forum 2016, 7:52).

Gaps

New technologies pose interesting opportunities for metadata, but their effectiveness is not yet explored in the literature possibly because of their proprietary nature. The extent of the use of semantic analysis within publishing as a whole has not yet been surveyed. Equally lacking is a deeper look at the roles of traditional catalogers and metadata experts in publishing companies, a study of which could provide more clarity regarding the differences between various publishers in the way that metadata is created and used. Again, proprietary concerns might be a cause, in addition to publishers who think of themselves more as self-contained units rather than part of a larger metadata ecosystem. The same concerns may prevent adoption of share controlled vocabularies which would aid in making discoverability easier across publishers.

Service Providers, Platforms, and Tools

Scholarly communications service providers (vendors) create tools and platforms to disseminate and facilitate the use of scholarly research. They include library system vendors (such as ExLibris, OCLC, EBSCO), E-retailers (Amazon), publishing services companies (Cenveo, Overleaf), and metrics organizations (Altmetrics, BibExcel), among others. Depending on the nature of the service offered, one service provider's use of metadata will vary significantly from another's. Such inconsistencies cause problems for resource discovery and for accessing full text content. By contrast, initiatives for open metadata and usage data, transparent pricing and contracts, allowance for community input, and use of well-established international standards to promote interoperability would bring positive change of the sort modeled in other industries.

Challenges

Metadata problems can arise when indexing practices differ between publishers and service providers. Publishers may index content at a different level than service providers. For instance, a database vendor might index several small articles under one title when a publisher indexes them discreetly. Service providers may not index some but not all of the articles in a journal issue, or they may index each article in an issue but not the issue itself (ExLibris 2017).

Problems also arise in the area of linking to content. Service providers create platforms, such as Integrated Library Systems (ILS), that allow users to discover bibliographic resources. These platforms feature links to full-text versions of content housed on external websites. The act of clicking on a link and arriving at the correct resource, known as link resolution, can be adversely affected by inconsistencies in metadata between content vendors, publishers, and providers of platfoms. ExLibris (2017) provides the following example: "You find an Article from Provider A, and go through a link resolver to link to Full Text from Provider B, it can happen that the metadata for the same Article from Provider A and Provider B are not the same, so Provider B cannot match the article, and the link fails." In the absence of a direct stable URL, the OpenURL standard (NISO Z39.88-2004) is used to facilitate linking from central indices like Primo or Summon to a target (e.g. article or ebook) on a publisher or aggregator platform. But when an OpenURL contains typographical errors or is poorly formatted, links fail to resolve.

Inconsistent or incomplete use of metadata fields also makes it a challenge to disambiguate similar titles, journals, and authors across distribution systems. If the metadata utilized in these systems is not normalized to authority sources or linked to unique identifiers, it is difficult to tell if two similarly-described resources are in fact the same. This serves as an inconvenience for individual researchers but also impacts our knowledge of the scholarly ecosystem as a whole. One research study detected significant problems in researchers' use of bibliometrics to analyze author networks because of lack of disambiguation. Analyses which falsely equate authors with the same or similar names distort our understanding of author networks and "may result in ill-informed decisions about research policy and resource allocation" (Kim 2017). Disambiguation could remain an obstacle, especially for text and data mining and the continued use of bibliometrics, though the adoption of ORCID is providing a solution.

Opportunities

Organizations such as the Scholarly Publishing and Academic Resources Coalition (SPARC) and the Confederation of Open Access Repositories (COAR) have developed best practice principles for data repositories which can also be applied to vendor services. Among the principles are calls for open metadata and usage data, transparent pricing and contracts, allowance for community input, and use of well-established international standards to promote interoperability (Confederation for Open Access Repositories (COAR) and Scholarly Publishing and Academic Resources Coalition (SPARC) 2019).

Several organizations have fulfilled aspects of these guidelines through policy changes and initiatives. In 2014, EBSCO opened metadata in its 139 databases for use by third parties (Enis 2014). In 2015, a partnership of health information vendors, health information management professionals, and other medical professional associations formulated standards for ensuring interoperability between different health information systems (Information Technology Infrastructure (ITI) Planning Committee 2015). The standards include statements as to the importance of quality metadata in maintaining vital information about data provenance (ibid., Table 8, Table 10.C). These two major initiatives demonstrate the rewards of improving metadata and access to metadata among services providers in scholarly communications.

Gaps

One challenge for other stakeholders in understanding the metadata practices of service providers is the large number of providers and tools offered. The literature would benefit from a comparison of different kinds of platforms and tools offered by service providers. Platforms and tools could be classified according to the impact of metadata for collection, curation, or consumption. Librarians in particular benefit when service providers are willing to adopt open metadata and usage data, transparent pricing and contracts, allowance for community input, and use of well-established international standards to promote interoperability.

Researchers

Researchers generate the ideas and data that precede all publications. As such, researchers are also a starting point for metadata: they use their own metadata while organizing their writing and research data, and are usually responsible for formally submitting metadata to a publisher or data collector when a project is ready to be released to the public.

Challenges

While researchers understand the value of making their work available to others, literature on the subject draws attention to the fact that the quality of their metadata does not always follow suit. Authors may be required to submit metadata along with traditional articles or monographs. Author-supplied metadata may benefit from the author's expertise, but is also, writes Tan (2018), prone to typos and outdated or incorrect information. One common problem, for instance, is the rate at which email addresses and other contact information becomes dated (p. 28). An earlier study gives weight to this claim, concluding that authors and article contributors are in the ideal position to provide valuable metadata for their articles given their deep knowledge of the subject (Wilson 2007, p. 16). Problems arise however, in the area of formatting metadata consistently. For 104 articles in the discipline of music, researchers identified 201 structural metadata issues, or mistakes in formatting (ibid., p. 21). While contributors expedite the metadata submissions process and lend

expertise, “efficient, usable systems and interfaces . . . guiding correct structural entry of metadata” are needed (ibid., p. 26).

On the whole, however, recent literature has focused more on researcher-generated metadata for data sets than on metadata submitted for published articles. On the subject of metadata for research data, a consensus emerges that researchers, while invested in the idea of sharing their data and findings, often lag behind in practice. Two studies of researcher perspectives on publication of data find that “researchers frequently fail to make data available, even when they support the idea or are obliged to do so” (Kratz and Strasser 2015, p. 2) and that “despite its potential and prominent support, sharing data is not yet common practice in academia” (Fecher, Friesike, & Hebing 2015, p. 18). One reason for this phenomenon could be the time and effort required to do so. Researchers most frequently cited time investment as a barrier for sharing research in a number of surveys (Kratz and Strasser 2015, p. 2).

The ability to manage large sets of data is increasingly important for researchers across disciplines. In the sciences, however, educational and training programs for researchers are still catching up. In one case, researchers published their experiences adopting new practices to manage data for the Ocean Health Index project. They found that “the need to improve practices is common if not ubiquitous” among environmental scientists who work with large data sets (Lowndes et al. 2017, p. 1). Echoing this concern, Barone et al. (2017) found, among a set of principal investigators in the environmental sciences who received National Science Foundation grants, 87% work with big data sets. When surveyed, the PIs reported that their research institutions were not meeting their training needs. 78% stated that their needs went unmet in the area of “data management and metadata” (p. 4).

The learning curve which researchers must negotiate to manage research data is not the only barrier for researchers who want to publish their data with appropriate metadata. Researchers must also manage the requirements and standards, sometimes conflicting, of different repositories, journals, and funders. In “Data management assessment and planning tools," Andrew Sallans and Sherry Lake find that “researchers’ current practices appear fragmented largely because funding agencies propagate broad requirements and provide few resources” (Sallans and Lake 2014). Similarly, McQuilton et al. (2016) find that “the proliferation of content standards and databases creates a barrier for researchers and database maintainers, creating confusion over which standard they should use to format their data, or which database to submit their data to” (p. 2).

Interestingly, researchers’ reliance on informal sharing might create a disincentive for quality metadata. In exploring how researchers share and obtain data, direct contact rates higher than all other methods, including sharing and retrieval via data repository (Kratz and Strasser 2015, p. 7). On the one hand, these numbers could be somewhat inflated because researchers are more likely to remember incidents of sharing which took place as a result of direct contact. On the other, the ubiquity of personal contact could explain a finding in the same study that researchers have relatively low expectations that data sets be accompanied by formal metadata.

Again, specifically in regard to data sets, a lack of shareable metadata has contributed to relatively low citation rates for data sets in published articles. While 49% of researchers thought that citation would be the most appropriate way of giving credit for data consulted (Kratz and Strasser 2015, p. 10), citation rates for data sets continue to remain low, with two previous studies finding that authors cited the data sets they consulted as little as 19% and 17% of the time (p. 16).

Opportunities

Researchers have a strong incentive to create good metadata. The quality of metadata associated with their scholarly outputs will affect the number of citations they receive, while utilization of quality metadata early in the research process will facilitate organization and save time down the road. A more selfless incentive, researchers also want to see knowledge advanced in their fields through better findability and wider access. These interests can be seen at work in the wide adoption of the Open Researcher and Contributor ID (ORCID), an initiative to assign a unique identifier to every author. These unique identifiers prevent confusion when authors change names or share a name with another individual: if a person publishes under the name Michaela Schaal in 2019 and then publishes an article under the name Michaela Petsch in 2020, it will remain clear through use of an ORCID that these names belong to the same person. Similarly, Michaela could publish as Michaela M. Petsch, M. M. Petsch, or any other configuration while remaining unambiguously associated with all her work. Searching by ORCID, researchers can be more certain that they are retrieving all the works of a given author and none from unintended authors.

A similar effort applied to the citation and sharing of data could yield powerful results, though not perhaps without shifting away from the model of traditional citation (Kelly 2015, p. 12). Efforts to improve researcher metadata should not be limited to researchers. Poole (2016) finds that researchers show an interest in creating shareable metadata but are often unaware that help is available from metadata experts, most often in the form of librarians at their home institutions (p. 969). Metadata experts in academic libraries, research data management institutions, and data publishing organizations all have a role to play, especially in developing a practice for research data sharing. Further discussion is offered in another section of this review, Data Curators and Repositories.

Gaps

A robust collection of literature profiles the problems that researchers encounter when managing research data, supplementing research on similar issues in the context of published books or articles. However, it remains difficult to attain a bird's eye view of the subject given the variety in the ways that researchers approach metadata: Metadata for research data varies based on funder requirements, data management software, data repository specifications, and research discipline and training. Likewise, author-supplied metadata for published works will vary according to the fields required by journals/publishers, type of metadata submission form or tool, and, research discipline. For both

research data and publications, researchers' level of involvement in creating metadata may vary based on the requirements of their institutions or publishers.

Studies that employ methodologies such as that of Wilson (2007) -- surveying a large number of publications to evaluate metadata supplied by authors and contributors -- would make it possible to more accurately depict a state of practice for researcher metadata. Given that community efforts for evaluating research data quality already exist (such as the widely deployed FAIR principles), readers would also benefit from an evaluation of research data quality across repositories or platforms. This evaluation could resemble a study by Neumaier et al. (2016) which developed criteria to measure metadata quality for open government/civic data across multiple sites and platforms. There is also a lack of understanding of the degree of consistency between institutions when it comes to practices for collecting metadata. A study of such practices would illuminate the researchers' training needs.

Though research policy makers (mostly in government institutions) support the accessibility of research data (Harvey et al. 2017, p. 19), more could be done to understand "the discipline-specific barriers and enablers for data sharing in academia in order to make informed policy decisions" (p. 20). More generally, there is a concern that "the ways in which data travel among scholarly fields remains understudied" (Poole 2016, p. 966).

Funders

Organizations that fund research range from governments to corporations to foundations. Common sources of government funding in the United States include, among others, the National Science Foundation (NSF), the National Institutes of Health (NIH), the National Endowment for the Humanities (NEH), the National Endowment for the Arts (NEA), and the Andrew W. Mellon Foundation. As research funders, these organizations have power to mandate how research data is maintained and distributed, both of which impact the researchers and institutions who generate and maintain metadata.

Challenges

Government funders have strengthened requirements for researchers to make their research publicly available. For example, in 2008, NIH began to require researchers to submit the peer-reviewed, manuscript versions of their papers to PubMed Central, an open access database for medical research (Dietrich et al. 2012, p. 1). In 2011, NSF began requiring that all researchers include a data management plan in their funding applications to facilitate data sharing (p. 2). While data management requirements in humanities-focused funding bodies may be less stringent, the NEA, for example, requires that grant proposals include a data management plan (National Endowment for the Arts 2019). Despite these and similar funder requirements, one survey found that data sets were recoverable in only 26% of cases where sharing was required, or 81 of 315 projects. Of these 81, the authors found that 60 data sets contained sufficient metadata for formal

archiving while 21 did not (Couture et al. 2018, p. 5). Though this low rate can be explained in part by the fact that the studies considered spanned 1989-2010, Poole (2016) finds that funder policies still lack sufficiently clear language and that they focus on access at the cost of preservation (p. 972). Mentioned above in the Researchers section, the finding of Sallans and Lake (2014) that "current practices appear fragmented largely because funding agencies propagate broad requirements" also applies. There is a consensus that funders acknowledge data sharing to be valuable but leave researchers without resources to ensure that their data is published and findable through quality metadata.

Opportunities

Beyond the wider benefits to the scientific community from access to open data, both private and public funders stand to benefit from more rigorous metadata requirements. For example, to measure the output of its Very Large Telescope (VLT) instruments, the European Southern Observatory (ESO) mandates that any article using data gathered from a VLT instrument include the ID of the project under which it was gathered in a standardized format ((ESO) 2018, 2018). This requirement allows ESO administrators to determine the number of citations ("science impact") generated by their instruments on the whole as well as how many citations are generated by each of their 12 VLTs (Leibundgut et al. 2017, p. 11). Over time, this data will allow the ESO to better understand the impact of its instrumentation, what kinds of research are performed on which instruments, and to make more informed decisions on future instruments. The judicious use of metadata in this case, which has an impact on the fiscal health and long-term success of an organization, could be extended to other organizations, working in different disciplines, which have an interest in measuring the impact of their labs or instruments.

Gaps

Foremost, the literature would benefit from studies which provide a systematic comparison of funder requirements for research data metadata as well as any guidelines provided, schemas recommended, or platforms specified. Additionally, government funding agencies are better studied than corporate funders even though, at least in the United States, corporate entities were responsible for 72% of research and development funding in 2015 (Boroush 2017, p. 1). Corporate funding bodies should be studied in the same manner as governmental bodies.

Librarians

Libraries are the places where end-users with varying levels of expertise often seek out resources and thus encounter metadata. The challenges that librarians face in making their resources accessible are often the problems faced by the metadata supply chain as a whole, with each stakeholder group having an impact on metadata quality for library resources.

Challenges

In “How libraries use publisher metadata,” Steve Shadle demonstrates how library discovery systems are impacted by publisher-generated metadata. (Shadle 2013) In the current environment, major academic libraries are paying third party suppliers, or library system vendors, for access to hundreds of thousands of electronic resources. Metadata records are not generated ‘by hand’ for most of these resources. Instead, metadata, most often created by publishers and content aggregators and reused by library software vendors and provide online access (via either direct linking or OpenURL link resolvers), is also used by librarians who download or receive MARC records in bulk for the packages they have purchased. For an electronic journal article, this metadata might include title, author, date, page numbers, ISSN, issue number, volume number, issue title, and DOI. In this way, writes Shadle, the choices of a publisher impact the librarian and end user: “In order for link resolvers to function properly, citation metadata coming from a source must be accurate. In most cases, this data is originally coming from the publisher” (Shadle 2013, p. 293). Kemp et al. (2018) echo these issues, writing that “... libraries face multiple external obstacles such as insufficient metadata provision upon the delivery of content from publishers (e.g. data about article license information), prohibitive costs of vendor systems and services, the absence of single-service options that provide multiple solutions” (p. 208).

Poor metadata quality brings into question the value of working with large numbers of publisher- and vendor-supplied metadata records. Records are generated by a variety of automatic methods, all of which are aimed at increasing efficiency when working with a large number of resources. Overall efficiency is somewhat diminished, however, if librarians are left without reliable ways to access a resource and if users of library databases see search results that are less than optimal. Wiersma and Tovstiadi (2017), for instance, found widely varying inconsistencies for search results based on which paid platform was used to search for ebooks. Librarians can encounter roadblocks when trying to confront vendors, library software vendors, or publishers about these problems (see the example in Bascones and Staniforth 2018, 2018, p. 13), and, when they do have the ability to make improvements to metadata themselves, “enriching existing records can be difficult and time consuming” (Kemp et al. 2018, p. 2018). Compounding the problem, Flynn argues that libraries lack entry-level cataloging positions (Flynn 2013, p. 30) at a time when they are very much needed. Given this and other problems, Bascones & Staniforth ask, “how many institutions are paying for items end-users will not be able to find in the library catalogues/discovery systems? How many disappointed end users are there?” (2018, p. 19).

Opportunities

Given the nature of these challenges, some see librarians as the best advocates for the end user and the health of the system overall because they “represent end-users who experience specific resources not ‘value for money’ packages” (Bascones and Staniforth 2018, p. 15). Furthermore librarians should not shy away from an advocacy role due to the

fact that “Libraries account for a majority of vendors’ business, giving librarians leverage for demanding high-quality, full-level cataloging records” (Flynn 2013, p. 30).

Librarians are also perceived as experts who can benefit other stakeholders. Kemp et al. (2018) argue for the value of librarians who connect with publishers to create taxonomies and to researchers in describing research data sets (p. 209). Similarly, Poole (2016) calls for librarians to create direct relationships with researchers who might not think of the library as a collaborator (p. 969). Mercer discusses her direct experience in acting as an advisor to two nascent journal editorial boards as a model for others to do the same (Mercer and Dyas-Correia 2011, p. 239). Librarians can also channel their expertise into advocacy for new initiatives. Librarians were “instrumental” in the development of ORCID (Kemp et al. 2018, p. 2019). The Shared Access Research Ecosystem (SHARE), an initiative “to maximize research impact by making research widely accessible, discoverable, and reusable,” is a collaboration between the Center for Open Science and the Association of Research Libraries (Association of Research Libraries In press). Librarians have also had a significant role to play in research data management. Poole (2016) writes optimistically of their potential impact on the quality and preservation of data sets, and is most optimistic about the potential of archivists in particular “...archivists are increasingly well-placed to address pressing digital curation needs” (p. 969).

Gaps

A useful contribution to the literature in the confluence of librarianship and metadata would be a study that takes up the aforementioned question of Bascones and Staniforth (2018), “how many libraries are paying for items that end users will not be able to find…?” Indeed, the question raised in their title, “Is wrong metadata really bad for libraries and their end users,” though answered with a resounding yes by the authors, is only answered from the perspective of the limited cases they were able to compile. More surveys like that performed in Wiersma and Tovstiadi (2017) would not only put librarians in a better position to make their case with publishers and vendors, but also serve to establish what in particular is wrong with the metadata in most cases.

As with researchers, publishers, and funders, librarians are a large group working in many different kinds of institutions with different subject knowledges. Hearkening back to the alternative metadata roles first mentioned in the Stakeholders section of this review, librarians are metadata creators, consumers, and custodians all rolled into one. While the differences between a cataloger, a health science librarian, and a reference librarian may be clear to those within the profession, the metadata needs of these various roles are not immediately apparent to those coming from the outside. Use cases and narrative examples written as far as possible in plain language would serve well in elaborating metadata problems for other stakeholders. Organizations like Metadata 2020, the Dublin Core Metadata Initiative (DCMI), and the North American Serials Interest Group (NASIG) each attempt to build our understanding of the metadata needs of different roles within the profession.

Data Curators and Repositories

As can be seen in the sections on researchers and funders, much recent literature is focused on managing metadata for research data. As of yet, this function is not performed by one kind of institution. Though 'data curators' are in some cases professionals with a specific title and a particular skill set, data is curated by many who do not carry that title. As far as possible, this section attempts to describe the challenges and opportunities faced by all of those who work with metadata for research data.

The focus on research data management in scholarly communications literature is in part due to increased interest in the problems of reproducibility and accountability in research. Without access to data, the validity of a claim in an article rests on the professionalism of the authors and the reputation of the journal. Concerns about integrity have reached the general public through TEDx*2 and public radio.*3 Kratz and Strasser (2015) write that "the reproducibility problem plaguing science in the scholarly and mainstream press could be addressed, in part, by opening underlying data to scrutiny" (p. 2).

Others have expressed concern that the traditional journal article does not model the scientific process correctly or is not the most effective way to facilitate new discoveries. "Papers provide an account that excludes research aspects and outputs that are not directly relevant to the arguments at hand," writes Leonelli (2016), "such as experiments or models that failed, data that proved irrelevant or neutral with regards to the hypotheses at stake, and procedural details that do not fit existing formats" (p. 393). Scholarly publishing was originally developed to suit printed material; our current technological capabilities may be able to better model the scholarly process by allowing researchers to "disseminate scholarship as it happens, erasing the artificial distinction between process and product" (Priem 2013, p. 437).

Challenges

A movement toward public access to data has followed and has attempted to create best practices for the infrastructure necessary to make data available. Data infrastructure is the collection of hardware and software used to support the preservation and retrieval of data, and relies heavily on good metadata. For instance, storing data with an eye toward long-term preservation requires technical metadata elements such as file inventory, file format (FITS, SPSS, HTML, Stata, Excel, tiff, mpeg, 3D, Java, CIF), file structure including "organization of the data file(s) and layout of the variables," version information including a date/time stamp for each version, checksum values, and the software and hardware in which the data were created (DMPTool In press). The Metadata Encoding and Transmission Standard (METS) is one metadata standard suited to capturing this information and is maintained by the Library of Congress (Library of Congress 2019). In terms of retrieving data, metadata is required which can accurately describe the data with subject-specific language while still facilitating access for the general public. For these reasons, "… activities involved in developing a data infrastructure are highly sophisticated

and require various types of expertise (including discipline-specific knowledge and information science)" (Leonelli 2016, p. 394).

A closely related issue is the challenge of citing data. As referenced in the Researchers section, proper metadata plays a role in preserving the most important incentive for researchers to publish data, that of certainty that their data will be cited. In Kratz and Strasser (2015), 83% of respondents indicated that a citation was the best way of giving credit for use (p. 6), but rates of citation fall well below this benchmark now as in the past (p. 16). Echoing this finding, Fecher et al. (2015) write that researchers donating their data to open access repositories "do not receive enough formal recognition to justify the individual efforts and that a safeguard against uncredited use is necessary" (p. 11). Moreover, traditional citation formats do not leave authors with enough options for citing all of the people, software, and tools (often discipline-specific) that contribute to the research process.

Opportunities

Proper methods for storing and citing research data are still in development, but a number of initiatives have emerged with best practices to guide researchers and publishers when creating robust metadata for citation by others. With regard to citation, The Research Data Alliance's Data Citation Working Group published a list of recommendations for citing data*4 (latest revision 2015) while DataCite assists with creating unique identifiers in the form of DOIs for data sets and data set updates (Kelly 2015, p. 15). CRediT from Casrai has brought nuance to the way that people are given credit, if not cited traditionally, for their contributions to research: authors and publishers can now choose between 14 different roles ranging from data curation to project administration to visualization (Casrai 2018). The Research Resource Identification Initiative (RRID) from PLOS is providing researchers, primarily biologists and geneticists, with a means for citing previously neglected sources of information such as models of organisms, antibodies, software, and databases (Hill 2015). Finally, the Software Citation working group at Force11 has developed recommendations for citing software used in academic research and software developers (Smith et al. 2016) at a time when authors of research software are increasingly looking to document, preserve and cite their software: At the time of writing, over 72,000 software DOIs have been minted via Zenodo, an open access repository (DataCite In press, accessed 2019-04-11), since the publication of "Making Your Code Citable,"*5 a github guide.

With regard to creating metadata for research data management, Force11 published the Findable, Accessible, Interoperable, and Re-usable (FAIR) principles in 2016 with associated metrics for evaluating machine actionable metadata and data quality. DataCite has created a metadata schema for data sets used by many repositories such as Elsevier's Mendeley-based repository (Kelly 2015, p. 16). Still, with over 2,000 data repositories worldwide (Witt 2018, 2018) a consensus has yet to form on the most appropriate platform for hosting data.

In “A metadata-driven approach to data repository design,” Harvey et al. outline the process for creating a repository in alignment with the FAIR (Findable Accessible Interoperable Re-usable) principles. DOIs are first assigned to collections of data, e.g. groups of "logically related material" (DSpace In press), as well as to data sets, individual data entries, and even individual instances when data was queried by a researcher. The authors recommend starting by assigning a DOI to the data collection which, in DSpace, will automatically inherit the ORCID of its creator. Another step in the workflow employed by the authors is to register the data via r3data, a global registry of research data repositories that enables one to describe the repository content and obtain a persistent identifier (DOI) for the repository itself (Harvey et al. 2017). The principles for research data management have been built from the ground up and cross institutional and disciplinary lines. The collaboration around this issue serves as a model for improving the rest of scholarly communications metadata.

Gaps

The promise of new initiatives has generated excitement but leaves questions as to their effectiveness which can only be answered as time passes. There is an opportunity for future research to quantify the adoption rates and impact of practices recommended by these initiatives. For instance, while Harvey et al.’s recommendation that DOIs be assigned to data collections, data sets, and individual data points would usher in a new level of transparency in scholarly research, a significant question remains as to whether researchers and data curators have the resources to implement it. The world of research data management would benefit from a service such as Crossref Participation Reports, which provides member organizations with an instant evaluation of their metadata quality and suggests ways to make improvements (Bartell 2018).

Moreover, it has yet to be seen whether adoption of initiatives by some researchers and organizations, but not others, will create a discordant state of practice that is itself a barrier to access for researchers who seek to study scholarly outputs in the aggregate. It may be difficult to ascertain which citation practices were in effect for whom and at what time.

The metadata supply chain

The literature detailed above, in addition to profiling the issues surrounding individual stakeholders, also demonstrate the ways in which stakeholders depend on one another. It is clear from the literature on librarians, for instance, that library discovery systems and institutional repositories are impacted by publisher and vendor metadata. Publishers, in turn, derive metadata from researchers, whose actions are guided by their personal choices and requirements of funders. The whole process impacts the ability of the end user to find the data and publications they need, from which point the process begins again.

There is a small body of literature dedicated to understanding the entirety of these complex interactions. Within the time scope of this review, attempts to diagram the scholarly

communications lifecycle begin with the development of a hierarchical model, consisting of multiple diagrams, which detail the stages of research (Björk 2007): formulation, performing the research (p. 15), informally communicating results (p. 17), publishing the results (pp. 18-24), and facilitating dissemination and retrieval (pp. 25-34). A somewhat less detailed attempt was also made in Regazzi (2015) *Scholarly communications: A history from content as king to content as kingmaker*. The book includes diagrams of the research lifecycle overall (p. 3), informal research communication (p. 4), journal publication (p. 23), and open peer review and publication (p. 32).

Diagrams provide a useful overall understanding of the scholarly communications ecosystem but, as of yet, do not include any particular focus on metadata. (One diagram does outline the integration of metadata into “search services,” (Björk 2007, p. 28)). None of the diagrams attempt to detail the responsible parties behind the creation of metadata or how and by whom it is used. A possible explanation for this gap is the granularity of metadata and the expertise required to understand its uses in a particular context. Metadata takes many forms and is most easily grasped by those with expertise, making it “difficult for those working with specific metadata in specific local environments to see how they can influence the bigger metadata workflow” (Bascones and Staniforth 2018, p. 3). The desire to speak to all parties which interact with metadata may guide literature in the future, as represented by this call from Bull and Quimby (2016): “Can we present a clear business case for engagement with metadata that speaks to all individual stakeholders – around discovery, usage, best value, impact, and relationship management? These should be central values to all of us regardless of whether we are a cost centre or a revenue business” (p. 152). This literature review and other efforts by Metadata 2020, such as a list of metadata best practices, seek to answer this call.

Acknowledgements

The author would like to thank those who lent their generous support as contributors to this literature review. Christopher Erdmann, Clare Dean, and Laura Paglione provided guidance as to the overall structure and arrangement of the review and helped the author to make contact with experts in the field. Julianne Schneider and T. Scott Plutchak edited a draft of the review and provided valuable comments. Michelle Urberg provided extremely thorough comments on the final draft of the review and, with Alice Meadows, co-chairs Metadata 2020’s Researcher Communications project which gave rise to this review. The author benefitted from a workshop graciously hosted by Henning Schoenenberger of Springer Nature. The review draws on resources created by the various project groups of Metadata 2020 and thus owes a debt to all of its participants. Finally, the author would like to thank individuals who responded for requests to lend their subject expertise including Stacy Konkiel of Altmetric and Dimensions and Nena Moss of DOE/OSTI.

Hosting institution

Metadata 2020

References

- Association of Research Libraries (In press) SHARE. https://www.arl.org/focus-areas/shared-access-research-ecosystem-share#. Accessed on: 2019-4-11.
- Barone L, Williams J, Micklos D (2017) Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLoS Computational Biology 13 (10): e1005755. [In English]. https://doi.org/10.1371/journal.pcbi.1005755
- Bartell A (2018) Member participation. https://www.crossref.org/participation/. Accessed on: 2019-4-13.
- Bascones M, Staniforth A (2018) What is all this fuss about? Is wrong metadata really bad for libraries and their end-users? Insights 31 (41): 1-24. [In English]. https://doi.org/10.1629/uksg.441
- Björk BC (2007) Scientific communication life-cycle model. Information Research 12 (2): paper 307. [In English]. URL: http://InformationR.net/ir/12-2/paper307.html
- Boroush M (2017) U.S. R&D Increased by $20 Billion in 2015, to $495 Billion; Estimates for 2016 Indicate a Rise to $510 Billion. National Science Foundation InfoBrief December 2017[In English]. URL: https://www.nsf.gov/statistics/2018/nsf18306/nsf18306.pdf
- Bull J, Schultz TA (2017) Harvesting the academic landscape: Streamlining the ingestion of professional scholarship metadata into the institutional repository. Librarianship and Scholarly Communication 6 (1): p.eP2201. [In English]. https://doi.org/10.7710/2162-3309.2201
- Bull S, Quimby A (2016) A renaissance in library metadata? The importance of community collaboration in a digital world. Insights 29 (2): 146-153. [In English]. https://doi.org/10.1629/uksg.302
- Casrai (2018) CRediT. https://casrai.org/credit/. Accessed on: 2019-4-10.
- Confederation for Open Access Repositories (COAR), Scholarly Publishing and Academic Resources Coalition (SPARC) (2019) Good practice principles for scholarly communication services. https://sparcopen.org/wp-content/uploads/2019/01/Sparc-Good-Practice-Principles-v4.pdf. Accessed on: 2019-4-10.
- Couture JL, Blake RE, McDonald G, Ward CL (2018) A funder-imposed data publication requirement seldom inspired data sharing. PLOS One 13 (7): e0199789. [In English]. https://doi.org/10.1371/journal.pone.0199789
- DataCite (In press) DataCite Search. https://search.datacite.org/works?resource-type-id=software. Accessed on: 2019-4-11.
- DataCite Metadata Working Group (2016) DataCite metadata schema documentation for the publication and citation of research data. 4.1. DataCite. URL: https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf
- de Moor T, van Zanden JL (2008) Do ut des (I give so that you give back): Collaboratories as a a new method for scholarly communication and cooperation for global history. Historical Methods 41 (2): 67-80. [In English]. https://doi.org/10.3200/HMTS.41.2.67-80

- Dietrich D, Adams TMA, Steinhart G (2012) De-mystifying the data management requirements of funders. Issues in Science and Technology Librarianship 70 (1): 1-16. [In English]. https://doi.org/10.5062/F44M92G2
- DMPTool (In press) Data management general guidance. https://dmp.cdlib.org/general_guidance#metadata-data-documentation. Accessed on: 2019-4-10.
- DSpace (In press) DSpace: an open source dynamic digital repository. https://duraspace.org/wp-content/uploads/dspace-files/DSpace_Diagram.pdf. Accessed on: 2019-4-10.
- Enis M (2014) EBSCO opens metadata to third-party discovery services. Library Journal 139 (9). [In English]. URL: https://www.libraryjournal.com/?detailStory=ebsco-opens-metadata-to-third-party-discovery-services
- (ESO) ESO (2018) Publications based on ESO data. http://www.eso.org/sci/observing/policies/publications.htm. Accessed on: 2019-4-11.
- ExLibris (2017) What are the common causes of full text linking problems, and how can linking be improved. https://knowledge.exlibrisgroup.com/360_Services/360_Link/Knowledge_Articles/What_Are_the_Common_Causes_of_Full_Text_Linking_Problems%2C_and_How_Can_Linking_Be_Improved%3F. Accessed on: 2019-4-10.
- Fecher B, Friesike S, Hebing M (2015) What drives academic sharing? PLoS ONE 10 (2): e0118053. [In English]. https://doi.org/10.1371/journal.pone.0118053
- Flynn EA (2013) Open access metadata, catalogers, and vendors: The future of cataloging records. The Journal of Academic Librarianship 39 (1): 29-31. [In English]. https://doi.org/10.1016/j.acalib.2012.11.010
- Harvey MJ, McLean A, Rzepa HS (2017) A metadata-driven approach to data repository design. *Journal of Cheminformatics* 9 (4). [In English]. https://doi.org/10.1186/s13321-017-0190-6
- Hill C (2015) Introducing the Research Resource Identification Initiative at PLOS Biology & PLOS Genetics. https://blogs.plos.org/biologue/2015/01/29/introducing-research-resource-identification-initiative-plos-biology-plos-genetics/. Accessed on: 2019-4-10.
- Imbue Partners (2017) Industry Leaders' Perspectives on the Digital Transformation Journey in Publishing. https://www.buchmesse.de/files/media/pdf/whitepaper-industry-leaders-perspectives-frankfurter-buchmesse.pdf. Accessed on: 2019-4-10.
- Information Technology Infrastructure (ITI) Planning Committee (2015) IHE Information Technology Infrastructure (ITI) white paper: Health IT standards for health information management practices. Revision 1.1. Release date: 2015-9-18. URL: http://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_WP_HITStdsforHIMPratices_Rev1.1_2015-09-18.pdf
- Jennings L (2017) Metadata for research discovery and management. Catalogue and Index 187: 5-8. [In English]. URL: https://purehost.bath.ac.uk/ws/portalfiles/portal/155978250/ci187_metadata_for_research_data_discovery_and_management_lizz_jennings.pdf
- Kelly MC (2015) The ongoing challenges of citing the results of scholarly research. Information Standards Quarterly 27 (2/3): 12-19. [In English]. URL: https://groups.niso.org/publications/isq/v27no2-3/Kelly/
- Kemp J, Dean C, Chodacki J (2018) Can richer metadata rescue research? The Serials Librarian 74 (1-4): 207-211. [In English]. https://doi.org/10.1080/0361526X.2018.1428483

- Kim J (2017) The impact of author name disambiguation on knowledge discovery from large-scale scholarly data. *University of Illinois at Urbana-Champaign School of Library and Information Science* [In English]. URL: http://hdl.handle.net/2142/98269
- Kratz JE, Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. PLoS ONE 10 (2): e0117619. [In English]. https://doi.org/10.1371/journal.pone.0117619
- Leibundgut B, Bordelon D, Grothkopf U, Patat F (2017) Scientific returns from VLT instruments. The Messenger (European Southern Observatory) 169: 11-15. [In English]. https://doi.org/10.18727/0722-6691/5032.
- Leonelli S (2016) The disruptive potential of data publication. Notes and Records 70: 393-395. [In English]. https://doi.org/10.1098/rsnr.2016.0036
- Library of Congress (2019) METS: Metadata Encoding & Transmission Standard. http://www.loc.gov/standards/mets/. Accessed on: 2019-7-06.
- Lowndes JS, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CO, Jiang N, Halpern BS (2017) ur path to better science in less time using open data science tools. Nature Ecology & Evolution 1 (article 0160): 1-7. [In English]. https://doi.org/10.1038/s41559-017-0160
- Marsh RM (2014) The role of institutional repositories in developing the communication of scholarly research. OCLC Systems and Services: International Digital Library Perspectives 31 (14): 163-195. [In English]. https://doi.org/10.1108/OCLC-04-2014-0022
- McNutt MK, Bradford M, Drazen JM, Hanson B, Howard B, Jamieson KH, Kiermer V, Marcus E, Pope BK, Schekman R, Swaminathan S, Stang PJ, Verma IM (2018) Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. Proceedings of the National Academy of Sciences of the United States of America 115 (11): 2557-2560. [In English]. https://doi.org/10.1073/pnas.1715374115
- McQuilton P, Gonzalez-Beltran A, Rocca-Serra P, Thurston M, Lister A, Maguire E, Sansone S (2016) Biosharing: Curated and crowd-sourced metadata standards, databases and data policies in the life sciences. Database: The Journal of Biological Databases and Curation 2016 (1): baw075. [In English]. https://doi.org/10.1093/database/baw075
- Mercer H, Dyas-Correia S (2011) Metadata value-chain for open-access e-journals. The Serials Librarian 60 (1-4): 234-240. [In English]. https://doi.org/10.1080/0361526X.2011.556040
- Michener WK (2015) Ecological data sharing. Ecological Informatics 29 (1): 33-44. [In English]. https://doi.org/10.1016/j.ecoinf.2015.06.010
- National Endowment for the Arts (2019) Research: Art Works FY19 Grant Application Form Instructions. https://www.arts.gov/sites/default/files/FY19-Research-Art-Works-Step2-Instructions-August2018-revised.pdf. Accessed on: 2019-7-05.
- Neumaier S, Umbrich J, Polleres A (2016) Automated quality assessment of metadata across open data portals. ACM Journal of Data and Information Quality 8 (1): article 2. [In English]. https://doi.org/10.1145/2964909
- Perrier L, Blondal E, Patricia Ayala A, Dearborn D, Kenny T, Lightfoot D, Reka R, Thuna M, Trimble L, MacDonald H (2017) Research data management in academic institutions: A scoping review. PloS ONE 12 (5): : e0178261. [In English]. https://doi.org/10.1371/journal.pone.0178261
- Poole AH (2016) The conceptual landscape of digital curation. Journal of Documentation 72 (5): 961-986. [In English]. https://doi.org/10.1108/JD-10-2015-0123

- Priem J (2013) Beyond the paper: the journal and article are being superseded by algorithms that filter, rate and disseminate scholarship as it happens. Nature 495 (7442): 437-440. [In English]. https://doi.org/10.1038/495437a
- Regazzi JJ (2015) Scholarly communications: A history from content as king to content as kingmaker. 1st edition. Rowman & Littlefield, London, 27 pp. [In English]. [ISBN 978-0-8108-9087-9]
- Sallans A, Lake S (2014) Data management assessment and planning tools. In: Ray J (Ed.) Research data management: Practical strategies for information professionals. 1st. Purdue Universtity Press, West Lafayette. [In English]. [ISBN 1557536643].
- Shadle S (2013) How libraries use publisher metadata. Insights 26 (3): 290-297. [In English]. https://doi.org/10.1629/2048-7754.109
- Silverchair Information Systems (2018) Platform Strategies 2018 - Search and Discovery. https://youtu.be/BFKAUOUNIlE?t=1191. Accessed on: 2019-4-11.
- Smith AM, Katz DS, Niemeyer DE, Group F1SCW (2016) Software citation principles. PeerJ Computer Science 2 (e86). [In English]. https://doi.org/10.7717/peerj-cs.86
- Society for Scholarly Publishing (2018) 4D. https://youtu.be/4A4_Vx8ySeE?t=978. Accessed on: 2019-4-11.
- Tan T (2018) A quick check on metadata: Tracking its issues, progress, and real-world application. Publisher's Weekly July 2: 28-30. [In English].
- Walter D (2016) Nielsen Book US study: The importance of metadata for discoverability and sales. http://www.nielsenbookdata.co.uk/uploads/10666%20Nielsen%20Book%20US%20Study%20The%20Importance%20of%20Metadata%20for%20Discoverability%20and%20Sales_DIGI_D5.pdf. Accessed on: 2019-4-11.
- Wiersma G, Tovstiadi E (2017) Inconsistencies between academic e-book platforms: A comparison of metadata and search results. Portal: Libraries and the Academy 17 (3): 617-648. [In English]. https://doi.org/10.1353/pla.2017.0037
- Wilson AJ (2007) Toward releasing the metadata bottleneck: A baseline evaluation of contributor-supplied metadata. Library Resources & Technical Services 51 (1): 16-28. [In English]. https://doi.org/10.5860/lrts.51n1.16
- Witt M (2018) 2,000 Data repositories and science Europe's framework for discipline-specific research data management. DataCite Blog https://doi.org/10.5438/jeag-2v54
- World Economic Forum (2016) Davos 2016 - The State of Artificial Intelligence. https://youtu.be/VBceREwF7SA?t=461. Accessed on: 2019-4-11.

Endnotes

*1 http://www.metadata2020.org/

*2 https://www.youtube.com/watch?v=dXKbkpilQME

*3 https://www.npr.org/sections/health-shots/2017/04/11/523406710/top-scientists-revamp-standards-to-foster-integrity-in-research

*4 https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

*5 https://guides.github.com/activities/citable-code/

*6 Alice Meadows and Fiona Counsell of Metadata 2020 have defined these roles, or personas, on the Metadata 2020 website: http://www.metadata2020.org/blog/2019-06-27-the-metadata-cast/