







# An international, multistakeholder survey about metadata awareness, knowledge, and use in scholarly communications

Kathryn A. Kaiser<sup>1</sup> , Michelle Urberg<sup>2</sup> , Maria Johnsson<sup>3</sup> , Jennifer Kemp<sup>4</sup> ,  
Alice Meadows<sup>5</sup> , and Laura Paglione<sup>6</sup> 

<sup>1</sup>Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA

<sup>2</sup>Independent Researcher and Maverick Publishing Specialists

<sup>3</sup>Lund University, Sweden

<sup>4</sup>Crossref

<sup>5</sup>National Information Standards Organization (NISO)

<sup>6</sup>Metadata 2020, Spherical Cow Group

an open access  journal



Citation: Kaiser, K. A., Urberg, M., Johnsson, M., Kemp, J., Meadows, A., & Paglione, L. (2021). An international, multi-stakeholder survey about metadata awareness, knowledge, and use in scholarly communications. *Quantitative Science Studies*, 2(2), 454–473. [https://doi.org/10.1162/qss\\_a\\_00133](https://doi.org/10.1162/qss_a_00133)

DOI:  
[https://doi.org/10.1162/qss\\_a\\_00133](https://doi.org/10.1162/qss_a_00133)

Peer Review:  
[https://publons.com/publon/10.1162/qss\\_a\\_00133](https://publons.com/publon/10.1162/qss_a_00133)

Received: 18 December 2020  
Accepted: 10 April 2021

Corresponding Author:  
Laura Paglione  
[lpaglione@sphericalcowgroup.com](mailto:lpaglione@sphericalcowgroup.com)

Handling Editor:  
Ludo Waltman

Copyright: © 2021 Kathryn A. Kaiser, Michelle Urberg, Maria Johnsson, Jennifer Kemp, Alice Meadows, and Laura Paglione. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



**Keywords:** FAIR principles, interoperability, metadata, researchers, standards, workflows

## ABSTRACT

The Metadata 2020 initiative is an ongoing effort to bring various scholarly communications stakeholder groups together to promote principles and standards of practice to improve the quality of metadata. To understand the perspectives and practices regarding metadata of the main stakeholder groups (librarians, publishers, researchers, and repository managers), we conducted a survey during summer 2019. The survey content was generated by representatives from the stakeholder groups. A link to an online survey (17 or 18 questions depending on the group) was distributed through multiple social media, listserv, and blog outlets. Responses were anonymous, with an optional entry for names and email addresses for those who were willing to be contacted later. Complete responses ( $N = 211$ ; 87 librarians, 27 publishers, 48 repository managers, and 49 researchers) representing 23 countries on four continents were analyzed and summarized for thematic content and ranking of awareness and practices. Across the stakeholder groups, the level of awareness and usage of metadata methods and practices was highly variable. Clear gaps across the groups point to the need for consolidation of schema and practices, as well as broad educational efforts to increase knowledge and implementation of metadata in scholarly communications.

## 1. INTRODUCTION

Classifying and/or descriptive information about knowledge resources (what we now call *meta-data*) has been in existence as long as libraries have. Callimachus of Cyrene created in 245 BCE the Pinakes (Jones, 2018), which is often characterized as the first library catalog and was comprised of a collection of bibliographic (meta)data about each of the works in the Library of Alexandria. With the advent of computers, Stuart McIntosh and David Griffel (1967) created a similar cataloging resource in the 1960s to that of the Pinakes. It was called *ADMINS* and listed of the library contents from several points of view (by author, by subject, etc.). However, they instead codified this information by attaching a set of “meta data” to the resources themselves, and then using a computer system to produce a list of resources from the desired point of view. As

metadata became incorporated into various scholarly communications systems (publishing, data repositories, cataloging, etc.), it maintained its status as a characteristic of the resources and systems themselves, often optimized for the use of the organization or system codifying the metadata rather than for the consumers of this information. A consequence of this evolution is that metadata maintenance and distribution has not really evolved as an industry of its own. Instead, it relies on sets of disparate standards and loose agreements for metadata exchange, which leaves some of the more ambitious promises of metadata unfulfilled as a mechanism for discovery, innovation, and knowledge distribution. Individual stakeholders thus reliably invest resources to optimize metadata use to address their own needs and goals.

### 1.1. Previous Work That Informs This Project

In 2019, the participants of the Metadata 2020 initiative published a review that examined the published literature regarding the use of metadata in scholarly communications (Gregg, Erdmann et al., 2019). The aim of this literature review was to address “... 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” (Gregg et al., 2019). Many challenges, gaps, and opportunities were revealed. This survey project follows the review’s structure, asking questions of similar stakeholder groups to gain additional insights into metadata knowledge, use, standards, practices, and other factors that not only describe the present landscape but point to other potential opportunities to improve industry practices.

Another project of Metadata 2020 participants’ work was focused on developing personas for those involved in metadata work, based on their metadata roles rather than their job titles or organization type—metadata creators, curators, custodians, and consumers (Metadata 2020, Kaiser et al., 2021; Metadata 2020 Personas, 2020). Metadata creators provide descriptive information (metadata) about research and scholarly outputs; curators classify, normalize, and standardize this descriptive information to increase its value as a resource; custodians store and maintain this descriptive information and make it available for consumers; and consumers knowingly or unknowingly use the descriptive information to find, discover, connect, and cite research and scholarly outputs. In analyzing the survey responses, the benefits of this approach become apparent in addressing metadata challenges. We consider the application of a persona approach in our discussion and concluding remarks.

### 1.2. About the Present Work

The survey provides insights from four key stakeholder groups for the metadata associated with scholarly objects—researchers, librarians, repository managers, and publishers. Respondents to the survey hailed from four continents (23 countries) and represented individuals at all stages of their careers. Additional details about respondents can be found in the methods section (Section 2). This paper shares the survey results and our insights organized by stakeholder group, mapped against their persona(s). Deidentified raw data, detailed survey result summaries, and the original survey questions are made available for further study by the reader. Details on accessing these resources can be found in the Appendix.

### 1.3. About the Metadata 2020 Initiative

Initiated by a diverse group of volunteers in 2017 (About Metadata, 2020), Metadata 2020 expanded quickly to develop into an international community of stakeholders from across scholarly communications. It has functioned as a collaboration advocating for “richer,

connected and reusable, open metadata for all research outputs in order to advance scholarly pursuits for the benefit of society” (About Metadata, 2020). While many efforts have been made to address challenges in single communities, few have extended solutions that are targeted to be applied across them. Recognizing this situation as a strength, the Metadata 2020 conveners agreed on an agenda driven by the interests and needs of this broad community. This approach allowed for unrestricted interactions among the subgroups participating in Metadata 2020 as determined by its volunteer community members and needs for additional expertise and review. As part of this initiative, to better understand how and the degree to which various stakeholders in scholarly communications understand and use metadata, and using insights gained from the aforementioned literature review, we deemed it necessary to test our personal impressions with a broader audience in the form of a survey. The survey was conducted by a subgroup called Research Communication within the Metadata 2020 project.

## 2. METHODS

Questionnaire content by stakeholder groups (librarians, publishers, repository managers, and researchers) was drafted by the first author, first for researchers, and then circulated to other project teams in the Metadata 2020 community to customize the questions for the other stakeholder groups. More specialized selection lists or modifications to questions were done by Metadata 2020 team members who identified with that stakeholder group and posted for review and comment. Final versions of question sets for the four stakeholder groups were then harmonized to allow for cross-stakeholder answer comparisons. An online survey tool (SurveyGizmo, Boulder, Colorado, United States) was used to create the question sets and branching structure based on the roles of stakeholders. The questions and answer choices were presented in English. The survey was released for public participation on June 5, 2019 and closed on July 15, 2019. Participants were recruited via direct emails of personal contacts, social media (Twitter, Facebook), blogs (e.g., *The Scholarly Kitchen*, listservs, and online newsletters). The questions and response options may be found in the Appendix. The first author summarized and analyzed all data, with input from stakeholder group members on creating categories of job roles/titles. All authors reviewed the analysis and raw data and discussed how to compare within and between the groups. The survey questions and protocol for data collection and handling were reviewed by the University of Alabama at Birmingham Institutional Review Board.

## 3. ANALYSIS

A total of 222 submitted responses were received and organized in Microsoft Excel™ by the first author. Responses were evaluated for completeness, and any ( $n = 11$ ) that were mostly blank responses were not further analyzed. The analyzed set total was 211 respondents.

Complete responses among the four stakeholder groups consisted of:

- 87 librarians,
- 27 publishers,
- 48 repository managers, and
- 49 researchers.

Overall, the sample consisted of respondents from four continents, from which 23 countries were represented (Table 1). Sixty-four respondents voluntarily provided their names and contact information to clarify free responses to questions as needed.

**Table 1.** Survey respondents by country and stakeholder group

# Respondents	Librarian	Publisher	Repository Manager	Researcher	Grand total	% Total
United States	48	14	22	10	94	45%
United Kingdom	18	6	12	9	45	21%
Canada	6	1	1	5	13	6%
Australia	3	1	7	2	13	6%
Germany	2	2	4	3	11	5%
Sweden	6			3	9	4%
Netherlands	1	1	1	2	5	2%
France			1	2	3	1%
India		1		1	2	1%
Italy				2	2	1%
Saudi Arabia				1	1	<1%
Brazil		1			1	<1%
Botswana				1	1	<1%
Anonymous				1	1	<1%
Venezuela				1	1	<1%
Malaysia				1	1	<1%
Serbia				1	1	<1%
Croatia				1	1	<1%
Switzerland				1	1	<1%
Nigeria				1	1	<1%
Ghana	1				1	<1%
Norway	1				1	<1%
Austria	1				1	<1%
Portugal				1	1	<1%
Grand Total	87	27	48	49	211	

As the question sets were specific to each target stakeholder group, each group's data is summarized separately. Where possible and with care to avoid risk of losing meaning, free responses were coded into categories to aid in the identification of themes or classes of responses by Metadata 2020 team members who were part of the respective stakeholder groups. Note that numbers reported in the results section in each stakeholder response set may not add up to the group total when a response may have covered more than one category, or some items were not selected.

The author team viewed the data for summary insights based on the questions from three perspectives:

- Information about the respondent and whom this person interacts with in terms of metadata used or workflows.
- Information about how this person interacts with metadata from internal or external sources.
- Metadata specifics.

All survey questions, as well as visualized raw responses and coded interpretations, are available in the Appendix.

## 4. SUMMARIZED RESULTS

### 4.1. Librarians

#### 4.1.1. About librarian respondents

The highest response rate of the survey was from the librarians group. Most librarians ( $n = 87$ ) responded via the direct email; 47% were from the United States, 18% from the United Kingdom, 6% from Canada, and 6% from Sweden. The concept of “metadata” is highly central and well known in the library sector—librarians are primarily curators, custodians, and consumers of metadata—so the comparatively high engagement and interest of this group was not surprising. Moreover, research on metadata for scholarly communications is quite dominated by the library and information sector, which was also reflected in the literature review by Gregg et al. (2019).

The respondents were asked to describe in free text their roles and subject areas, and there were a variety of responses to this question. The responses were grouped into four areas of common roles for library services, with a fifth category of miscellaneous “other” roles. Most respondents have roles related to “Cataloging and metadata.” See Figure 1 for the relative distribution of roles for Librarian respondents.

Generally, most of the respondents come from very traditional libraries with a common offering of services. This is also reflected in the results of the remaining questions, available in the Appendix.

#### 4.1.2. Library insights

Librarians focus on offering researcher education about metadata in different ways (Library Survey Q4). While the question was posed requesting a free-text answer, offerings for metadata education and support could be grouped into several broad categories:

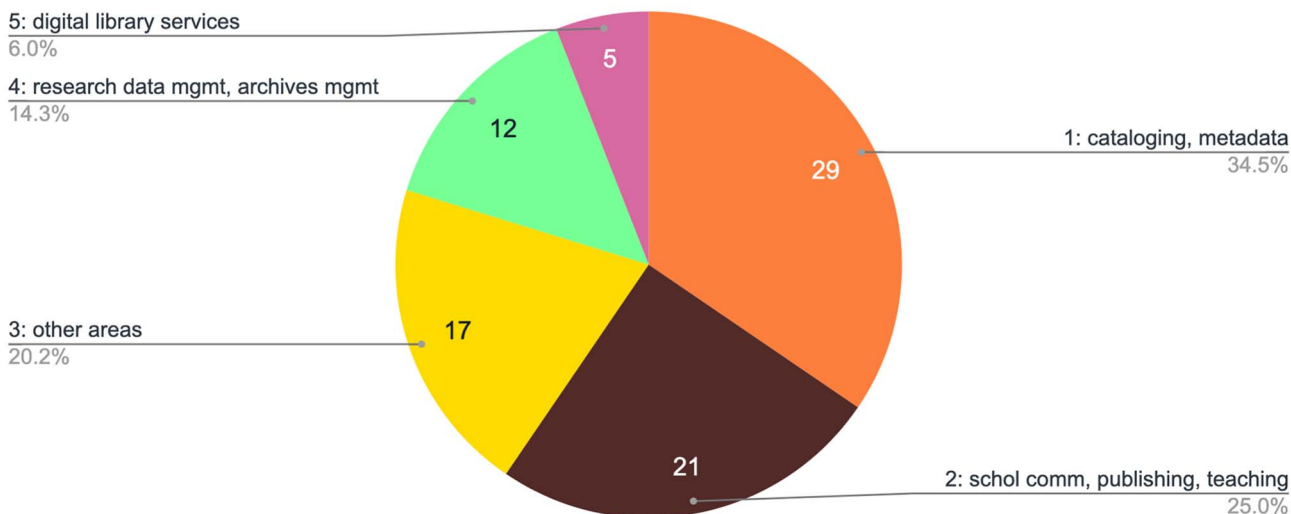
- Group: Classes and workshops (34.9%,  $n = 44$ )
- Personalized or individual: Individual services and consulting (30.2%,  $n = 38$ )
- Self-serve: Web-based user guides and software tools (14.3%,  $n = 18$ )
- Enhancement to research assets (8.7%,  $n = 11$ )

However, supporting and educating research is only a component of librarian activities. When asked about the role of their services and support of metadata (Library Survey Q5), creating and analyzing cataloging metadata is a slightly more common role:

- Create/analyze cataloging metadata (29.2%,  $n = 21$ )
- Support researchers with metadata production (22.2%,  $n = 16$ )

## Library/ information science speciality or subject area

Summarized and grouped answers



**Figure 1.** Library/information science specialty or subject area (write-in answers have been summarized and grouped into broader categories).

- Create institutional repository metadata (13.9%,  $n = 10$ )
- Assist with discovery of research (9.7%,  $n = 7$ )

Responses to these two questions, moreover, revealed a strong correlation between description of metadata support methods—education vs. enhancements—and the job title identified in the Library Survey (Library Survey Q2). For example, catalogers and metadata librarians focus on creating enhancements with bibliographic records, while reference librarians tend to focus on education through group instruction, individual consults, and self-serve library guides.

Researchers' interactions with librarians were revealed to be targeted interactions about planning research projects (48%), including asking for help finding scholarship and data sets, and about post-project publication and reporting help (47%) (Library Survey Q7). When researchers seek consulting services about where to find scholarship or data sets, librarians point them towards services that support rich, accurate, and elaborated metadata: first to abstracting/indexing databases (59.8%), then to discovery layer keyword searches (46%) and to discovery layer subject searches (41.4%) (Library Survey Q8). When librarians discuss searching strategies with patrons, they report title (53%), year and date of publication (44%), and author name(s) (44%) as the three most important pieces of information for finding relevant scholarship or data sets (Library Survey Q12). These three fields correspond with the three fields librarians identify as most important in scholarly publications (Library Survey Q11). Librarians also report that researchers often do not ask directly about metadata; rather, they ask questions related to metadata in some way ( $n = 54$ ). When researchers do specifically inquire about metadata, they often want to learn *how to create metadata* (41%) (Library Survey Q9).

### 4.1.3. Key librarian themes

The responses indicated that librarians have a strong motivation to instruct and educate users about metadata and are highly engaged with maintaining good metadata records in systems—

as they understand metadata's importance for knowledge discovery and scholarly communication. Responses show that:

- Many librarians are educating and instructing users about metadata through classes and workshops, by individual consulting, and by web-based user guides.
- Quality control of metadata in various systems is a major and time-consuming task for many librarians.
- Librarian metadata work is spread across multiple systems including, library catalogs and discovery layers for item-level cataloguing or in Current Research Information Systems (CRIS) or other repositories where they check researchers' publications.
- Metadata is important: "... we aim to ensure that metadata is fit for purpose, well maintained and widely disseminated"—one of many statements indicating the importance of metadata according to librarians.

In several previous publications, authors discuss the problems that libraries experience with poor and incomplete metadata received from publishers and system vendors, and the needs and opportunities for collaboration between these groups (Bascones & Staniforth, 2018; Flynn, 2013; Kemp, Dean, & Chodacki, 2018; Poole, 2016).

## 4.2. Publishers

### 4.2.1. About publisher respondents

Surprisingly, although publishers may be metadata creators, curators, custodians, and consumers, they were least well represented in the survey results, with just 27 completed responses. Most respondents were U.S.-based (52%,  $n = 14$ ) or U.K.-based (22%,  $n = 6$ ), with the remainder of respondents coming from six other countries. The most well represented job functions among respondents were editors (30.8%,  $n = 8$ ) and/or director-level strategic managers (23.2%,  $n = 6$ ), including two production/content management respondents, perhaps indicating an understanding of how better metadata will help increase their organization's success in the future. Other job functions represented included content management, data management, and information discovery/library support (Publisher Survey Q2). Despite the low response rate from publishers in this survey, previous research shows that publishers are engaged in improving their metadata and willing to invest in new technologies to process it (Gregg et al., 2019; Imbue Partners, 2017; Kemp et al., 2018).

### 4.2.2. Publishing insights

Nearly half of respondents (48.5%,  $n = 16$ ) either felt that there was little or no support for metadata education of users by their organizations, or they were unclear about the level of support (Publisher Survey Q3). Likewise, about half of respondents (48%,  $n = 13$ ) felt that the critical issues with metadata for their publications reside with the researchers—and yet educating researchers does not seem to be a priority for most of the respondents (Publisher Survey Q9). A significant number of respondents (34.5%,  $n = 10$ ) did not know what changes, if any, should be made to their organization's approach to metadata education for end users (Publisher Survey Q4).

Lack of controlled metadata is seen as a problem by many respondents. It was ranked the second most critical issue in workflows (48%) (Publisher Survey Q9), and related issues, such as the need for standardization and consistency, were mentioned several times in free text comments identifying critical issues for metadata submission (Publisher Survey Q6). Manual metadata entry exacerbates this problem. Metadata was identified to be entered at least partly

manually in most organizations represented, by internal experts (70%), authors (59%), or external experts (26%) (Publisher Survey Q5). Challenges of manual entry may also account for why publisher respondents ranked internal cleanup lower than the lack of controlled metadata as a critical issue (Publisher Survey Q8). Over half of publishers reported a combination of manual and automatic metadata workflows, but only four reported use of automated processes exclusively (Publisher’s Survey Q5). As shown in Figure 2, when scholarly publishers export their metadata, it can be found in many places, particularly Crossref (Crossref, 2020), library service platforms, aggregators, and PubMed (PubMed, 2020) (Publisher Survey Q8).

There was a clear consensus about the most important metadata elements, with (in order) title (93%), year and date of publication, DOI (each 89%), author name(s) (81%), and abstract (70%) noticeably more highly ranked than any of the other options (Publisher Survey Q11). Unsurprisingly, most of these were also the top elements that respondents said were required in internal publisher systems and their content platforms, though abstracts were ranked slightly lower and, instead, volume and issue number were seen as more important (Publisher Survey Q12). The NISO JATS schema (Journal Article Tag Suite, 2020) was by far the most used metadata schema used by publishers (63%) (Publisher Survey Q14).

4.2.3. Key publisher themes

Through their interactions with researchers as both creators and consumers of content, publishers are well placed to play a larger role in improving metadata workflows than they appear to do currently based on the responses to this survey. For example:

4.2.3.1. Communication need-priority mismatch Publisher respondents saw a lack of understanding of the benefits of metadata for end users (59%) and for discovery (70%) as the most

Metadata destination when exported

% respondents that selected location. Multiple answers encouraged.

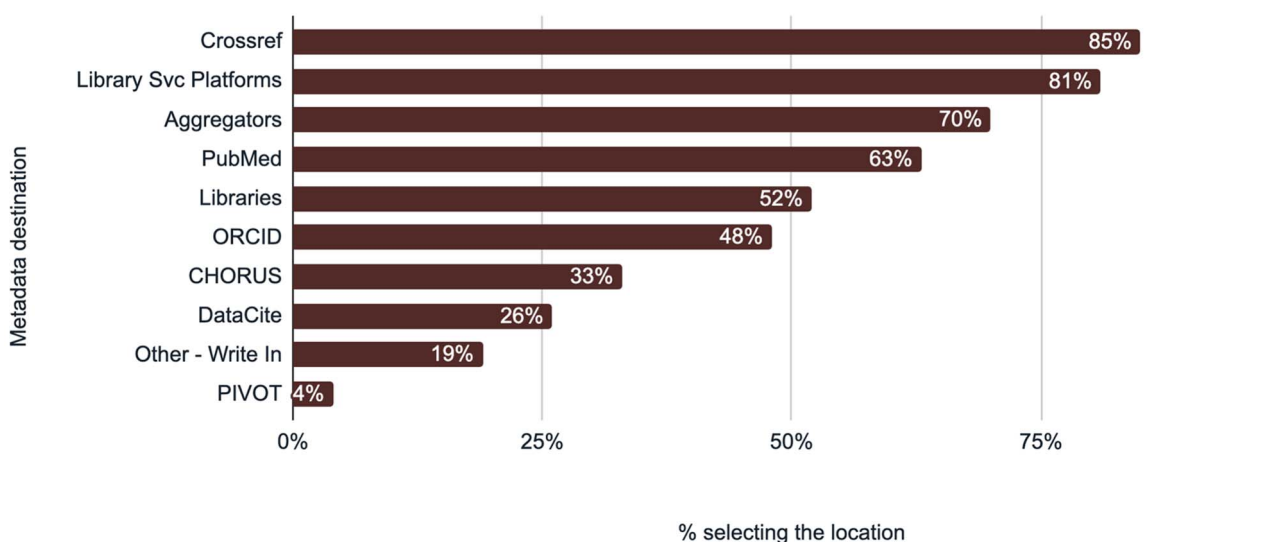


Figure 2. Metadata destination when exported (figure shows the percentage of respondents that selected the destination. Multiple answers were encouraged).



critical issues for authors (Publisher Survey Q6), although they de-emphasized promotion and guidance as an area for education support (20.7%,  $n = 6$ ) (Publisher Survey Q4). Where support was provided by publishers, it was as a mix of e-resources (18.2%,  $n = 6$ ) and live resources (21.2%,  $n = 7$ ). Others indicated support, but did not specify the type (12.1%,  $n = 4$ ) (Publisher Survey Q3). Most respondents (34.5%) had no suggestions for changing the support they provide. Of those that suggested changes, most favored promoting its benefits (20.7%), providing technical support (17.2%), or providing guidelines and templates (13.8%). (Publisher Survey Q4).

**4.2.3.2. Limited support for author-driven metadata correction** The publisher respondents reported limited or no opportunities for authors to update their own metadata after submitting it. Only one respondent indicated that their organization allows authors to update all fields themselves (Publisher Survey Q7). Where metadata can be updated this appears to be a manual or publisher-controlled process that would be seen as a barrier by researchers, or a sign that publishers do not value their input on metadata creation or curation for their own works. If authors are to take ownership of their metadata, this clearly needs to be addressed.

**4.2.3.3. Inconsistent quality verification methods** Because quality control is highly variable, metadata quality and completeness varies widely. In addition to the challenges introduced by manual data entry, combinations of manual and partially-automated quality control checking exist in most publishing workflows. Over half of respondents have some sort of automated metadata checking process in place (59%), but most also use other manual forms of checking—either in their own systems (48%) or other systems (33%), or via curated support from others (44%), with 11% reporting no checking at all (Publisher Survey Q10).

### 4.3. Repository Managers

#### 4.3.1. About repository manager respondents

The repository managers group has the most diverse set of roles within the groups surveyed. Some respondents are in charge of units, divisions, or organizations that manage data in repositories (e.g., manager, director, head); some are working directly with data ingestion or manipulation (e.g., analyst, curator, archivist); some are helping researchers use a repository effectively or to plan for managing their data and scholarship in a repository (e.g., research data librarian, digital collections librarian); and many work in a variety of contexts within the same organization. These respondents work with a wide variety of data types, with a variety of workflows, and manage their metadata using many different standards.

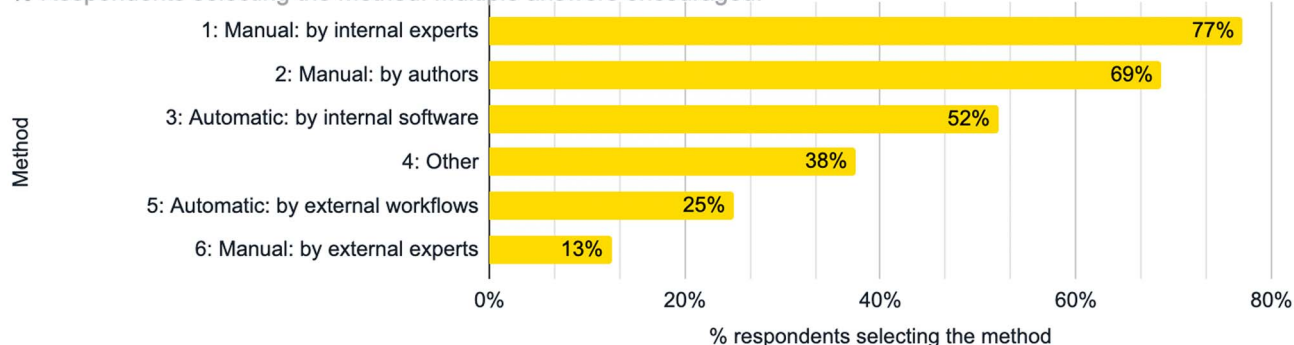
Repository managers are typically curators and custodians of metadata. The professional duties of this group vary widely, with most working in a university setting (73%,  $n = 35$ ), and others in an industry setting (21%,  $n = 10$ ) or public archival role (6%,  $n = 3$ ) (Repository Manager Survey Q2). Of the 48 responses received, nearly half work in the United States (45%,  $n = 22$ ), a quarter work in the United Kingdom (25%,  $n = 12$ ), fewer work in Australia (15%,  $n = 7$ ), and the remainder work in Germany, Canada, France, or the Netherlands (15%,  $n = 7$ ).

#### 4.3.2. Repository metadata insights

Most organizations—university or industry—use a combination of manual and automatic methods to create data in their system about repository content (Repository Manager Survey Q7, Figure 3). Several respondents indicated other methods (38%), for example, using CRIS, Symplectic Elements, or ETDs (Electronic Theses and Dissertations) to harvest metadata, which are then used in conjunction with manual repository workflows.

## Methods of data entry into repository systems

% Respondents selecting the method. Multiple answers encouraged.



**Figure 3.** Methods of data entry into repository systems (figure shows the percentage of respondents that selected the method).

Half of the respondents identified two or three workflow challenges (50%,  $n = 24$ ). The top three challenge areas identified by this group are associated with the researcher (80%), internal data clean up (58%), and lack of controlled vocabulary (46%) (Repository Manager Survey Q11).

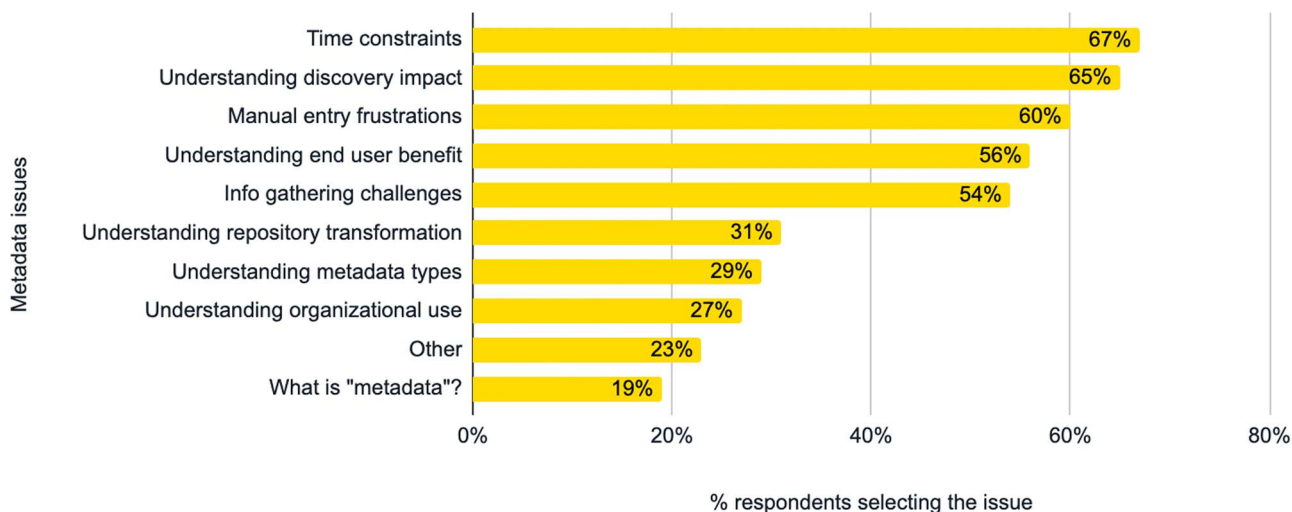
Most of the repositories not only ingest metadata but export it as well. DataCite (DataCite, 2020) (50%) and Library Service Platforms (44%) were the two most commonly reported places to deposit repository data (Repository Manager Q9). Unexpected was the wide variety of options not included in the survey (other, 44%). The written-in responses revealed that repositories deposit their data into a very different set of tools than librarians or publishers. Google Scholar, CORE, and Trove were common tools for this group not originally identified in the survey options. The evolution of institutional repositories as well as discipline-based repositories for research data has been discussed in the literature recently (Gregg et al., 2019). Many organizations with institutional repositories have spent a lot of resources reviewing their workflows for populating their repositories with data and metadata (Bull & Quimby, 2016; Jennings, 2017). The expansion of repositories also has to do with the increasing interest in open data and the development of robust systems to share research data in recent years.

Most of the repositories require metadata to be present at the time of deposit (77%), which aligns with what repository managers hear from their clients and patrons about their challenges with entering data (Repository Manager Survey Q10). The top three critical issues identified by the repository managers for those entering data in their system are the time needed to add required information (67%); lack of knowledge about the impact of metadata in discovery (65%); and frustrations about manual entry for different types of research objects (60%) (Repository Manager Survey Q8, Figure 4).

The standards, schema, specifications, and element sets used by this group vary widely, and many repositories use at least one standard for recording metadata about their repository objects. Dublin Core (DC) is the most-used element set (49%), with DataCite XML (19%), and MODS (17%) used third and fourth-most (the catch-all Other category returned the second most (33%)). The wide use of Dublin Core is not surprising given that many of the respondents are embedded in a university setting. The Other category also revealed a weakness in the survey instrument: DDI, ISO 9115, and internal schema should have been included. As one respondent points out, the answer options for this question were very library (and publisher) centric (Repository Manager Survey Q15).

## Critical issues for metadata submitters

% Respondents selecting the issue. Multiple answers encouraged



**Figure 4.** Critical issues for metadata submitters (figure shows the percentage of respondents selecting the issue. Multiple answers were encouraged).

The pieces of metadata identified by repository managers as being “required” (Repository Manager Survey Q13) versus “most important” (Repository Manager Survey Q14) reveal a couple of telling pieces of information. Direct Object Identifiers (DOIs) are identified as the fourth most important pieces of metadata (65%) behind Title, Author Name(s), and Year and Date of Publication, but they are ninth most important in the metadata required for deposit (40%). Author ORCID(s) (ORCID, 2020), likewise were identified as the sixth most important piece of metadata (46%), but ORCID(s) do not appear in the top 10 required pieces of metadata.

### 4.3.3. Key repository manager themes

A high-level analysis of the responses reveals a few key things about repository managers:

- Repositories are found in a lot of different corners of the scholarly communications life cycle and no two repositories are alike.
- They have different metadata needs with respect to what pieces of content are necessary and important for a repository to record and make accessible deposited content.
- Content intervention by humans is still required in repositories to address gaps in controlled vocabularies, but opportunities exist for automation to assist content and metadata ingestion.

## 4.4. Researchers

### 4.4.1. About researcher respondents

Researchers are primarily creators and consumers of metadata (e.g., keywords for articles, data set descriptions). However, they reported very little knowledge about broader types of metadata ( $n = 49$  respondents). As the survey was short and simple, we can only speculate as to whether researchers may be motivated to learn more or do more to support better metadata for creation and curation of research outputs that support their own needs. Researchers need to be educated

in ways that creation of richer metadata will enhance the utility of the scientific record, including supporting findability and potentially increasing citations (a metric important to many researchers).

Most researchers responded via email distribution links, and the top three countries represented by respondents were the United States (20%), the United Kingdom (18%), and Canada (10%). The researchers reported specific fields of study or research, which were then more generally categorized (Researcher Study Q2, Figure 5).

**4.4.2. Researcher insights**

Respondents reported 94% actively publishing their research (Researcher Study Q4), and 91% reported collaborating with other researchers (Researcher Study Q5). When asked about their concept of metadata (When you think of the term “metadata” what comes to mind?), their free-text responses (summarized in Figure 6) reflected the wide variety of experiences, uses, and understanding about metadata among researchers (Researcher Survey Q6).

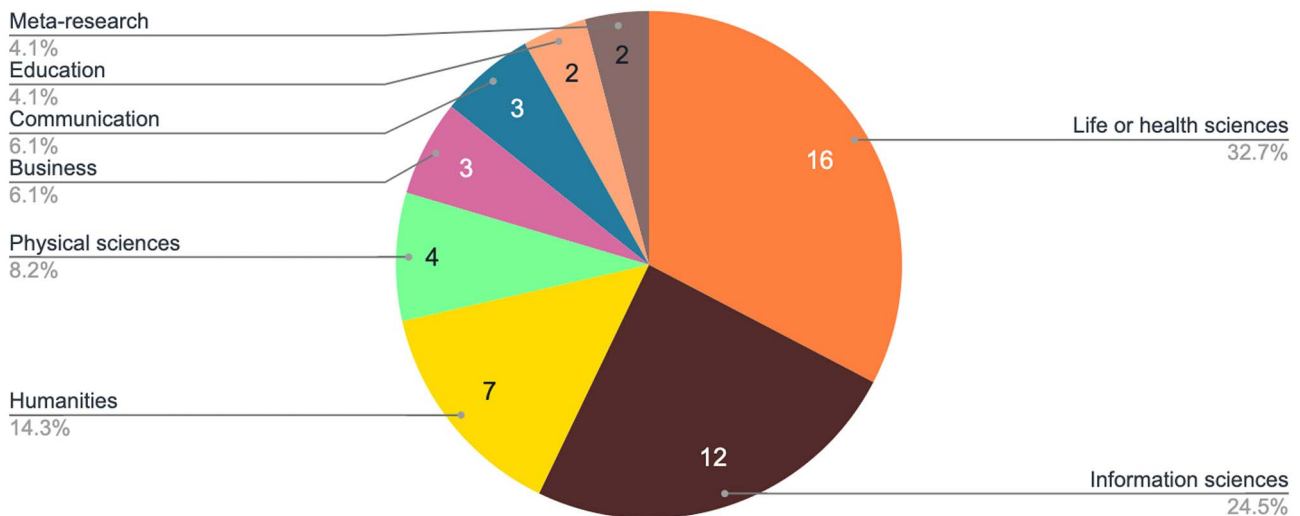
When asked about the most important pieces of information about research objects to make them accessible and useful to others in their field in future, the top five selected responses were title (90%), author name (90%), date of publication (86%), digital object identifier (DOI, 72%), and abstract (56%) (Researcher Survey Q8). When asked what pieces of information are most helpful in finding articles sought, the top five selected responses were title (88%), author name (70%), abstract (54%), date of publication (50%), and DOI (36%) (Researcher Study Q9).

When seeing an interesting title or scanning through a table of contents, researchers indicated the following information as most important when deciding to read an article or a book chapter in detail: full text access (86%), date published (44%), population and design details (42%), availability in an online format (33%), and the authority of the authors (32%) (Researcher Survey Q11).

Researchers indicated what they feel to be the most critical evidence and artifacts to know to evaluate credibility and usefulness of a data set. The top information selected includes the

**Research field and/or topic**

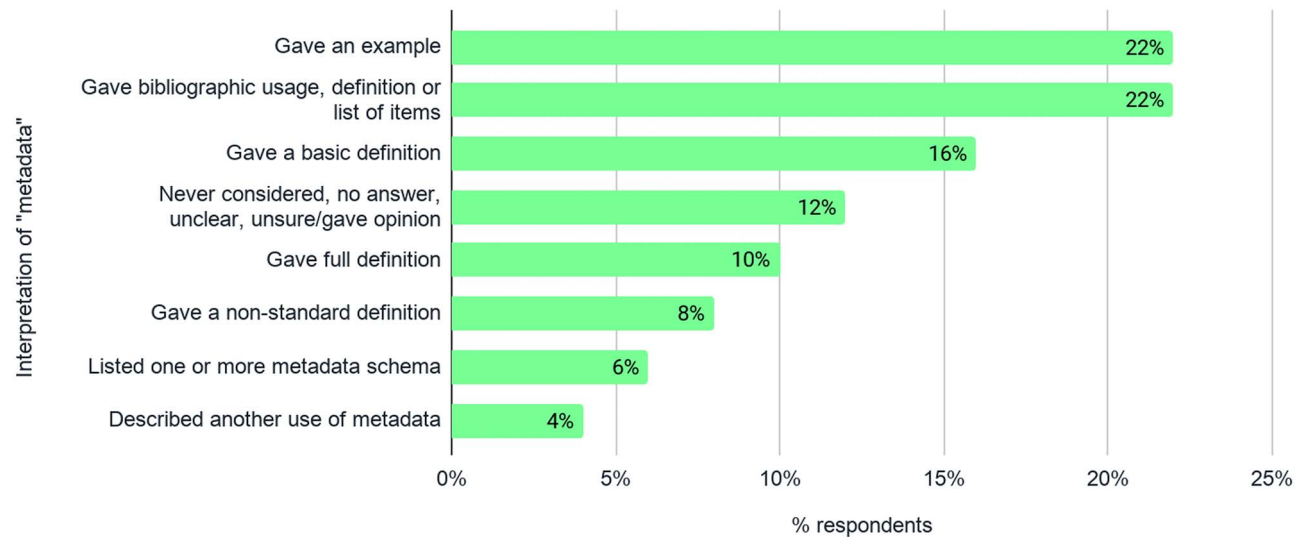
Summarized and grouped answers



**Figure 5.** Researcher’s field and/or topic (write-in answers have been summarized and grouped into broader categories).

## What comes to mind when you hear the term "metadata"?

Summarized and interpreted answers.



**Figure 6.** Categorized conceptions of metadata among researcher respondents.

collection creation description (72%), scope of the sample set (48%), limitations of scope (48%), number of study subjects (44%), and use description (40%) (Researcher Study Q12). Related to this evaluation, the following aspects were in the top five traits that might prevent use of a data set or piece of evidence: usage restrictions (56%), limitations of scope (44%), financial costs (38%), file formats (30%), and collection creation process (26%) (Researcher Study Q13).

Researchers were asked where they store data sets and other scholarly artifacts when done using them. Many store them on their personal computer (64%), though for many, this information is also stored elsewhere. Just over half reported storing artifacts in institutional storage (52%). Other locations include personal cloud storage (38%) and public repositories or websites (36%) (Researcher Survey Q14).

There was not a high degree of consensus among the surveyed researchers about the standardized metadata schemas that they use when publishing research outputs. About a third of respondents indicated that they did not know what standards were used (32%), and nearly as many indicated that they use none of the schema (28%) or something other than the choices provided (22%). The most popular schema from which they could select were Dublin Core (22%), other (22%), and JATS (Journal Article Tagging Suite, 10%).

### 4.4.3. Key researcher themes

Several key themes emerged about researchers:

- The response patterns reflect a limited perspective of researchers as consumers of metadata, or specific awareness or expressed needs about metadata for other researcher purposes.
- The reported common uses of metadata reflect the position of researcher as a creator of content, but mostly for human consumption in small volumes rather than for machine readability or large volume work.

- Full access rights and information about usage restrictions is top of mind for researchers. Although few questions asked about this directly, free-text and provided comments frequently mentioned these topics.

#### 4.5. Cross-Stakeholder Insights

All stakeholder groups were asked which metadata fields they feel are most important, and which metadata schemas they use. These questions highlighted stark differences between and—particularly in the case of repository managers—within each stakeholder group.

##### 4.5.1. Most important metadata fields: Different priorities across stakeholders

Respondents were provided a list of 28 metadata fields commonly found in scholarly publishing and were asked to select the set of them that they felt were the “most important.” (Cross-Stakeholder A1) The list of fields was developed during a cross-stakeholder workshop that listed the metadata fields that were most important to the participants. On average, publishers selected more fields (11.4 on average) than the other stakeholders, and nearly 40% more fields than librarians, who on average selected 6.8 of the fields listed (Cross-Stakeholder A2).

##### 4.5.2. Metadata fields more important to publishers than other stakeholder groups

In addition to selecting more fields, publishers were significantly *more* likely to select 11 specific fields than other stakeholders, particularly librarians (Cross-Stakeholder A3) (Table 2).

##### 4.5.3. Metadata fields significantly different in importance to librarians

The fields that librarians selected as most important also varied significantly from other stakeholder groups, with librarians selecting several fields as important significantly *less* often than other stakeholder groups (Table 3). Differences for data in bold indicate metadata fields where

**Table 2.** Percentage of respondents indicating that a metadata field is important, compared to publisher responses (**publishers are more likely to select certain metadata fields as important compared to other stakeholder groups. Differences greater than 20% in bold**)

Field	Publishers	Librarians		Repository		Researchers	
	%	%	Difference from publishers	%	Difference from publishers	%	Difference from publishers
Year. & date of publication	89	57	<b>-32%</b>	69	-20%	86	-3%
DOI	89	47	<b>-43%</b>	63	<b>-26%</b>	72	-17%
Abstract	71	25	<b>-46%</b>	43	<b>-29%</b>	56	-15%
Author(s) ORCID	54	23	<b>-31%</b>	45	-9%	48	-6%
Volume / issue	46	27	-19%	20	<b>-26%</b>	34	-12%
Author(s) affiliation	50	10	<b>-40%</b>	37	-13%	30	-20%
ISSN	43	18	<b>-25%</b>	27	-16%	30	-13%
Page numbers	39	15	<b>-25%</b>	16	<b>-23%</b>	24	-15%
Grant ID(s)	36	8	<b>-28%</b>	37	1%	10	<b>-26%</b>
Funder Name(s)	36	9	<b>-27%</b>	22	-13%	18	-18%
Funder ID(s)	32	6	<b>-26%</b>	24	-13%	16	-18%

**Table 3.** Percentage of respondents indicating that a metadata field is important, compared to librarian responses (**metadata fields that are significantly different in importance to librarians; % respondents selecting the field; multiple selections encouraged; differences greater than 20% in bold**)

Field	Librarians	Publishers		Repository		Researchers	
	%	%	Difference from librarians	%	Difference from librarians	%	Difference from librarians
Year & date of publication	57	89	<b>32%</b>	69	13%	86	<b>29%</b>
DOI	47	89	<b>43%</b>	63	17%	72	<b>25%</b>
Abstract	25	71	<b>46%</b>	43	18%	56	<b>31%</b>
Author(s) ORCID	23	54	<b>31%</b>	45	<b>22%</b>	48	<b>25%</b>
Volume / issue	27	46	19%	20	-7%	34	7%
Author(s) affiliation	10	50	<b>40%</b>	37	<b>27%</b>	30	20%
Grant ID(s)	8	36	<b>28%</b>	37	<b>29%</b>	10	2%
Language of content	26	21	-5%	22	-4%	20	-6%
Local controlled subject vocab	20	25	5%	24	4%	10	-10%
LCSH subjects	28	11	-18%	2	<b>-26%</b>	6	<b>-22%</b>
MeSH Subjects	14	18	4%	2	-12%	10	-4%

librarians were less likely to indicate that the field was important. Data in italic indicate that the librarian group was more likely to indicate the field was important (Cross-Stakeholder A4).

## 5. DISCUSSION AND RECOMMENDATIONS

### 5.1. Metadata Now and Then: The View of Metadata Throughout the Years

Libraries have dominated the application and development of modern metadata and systems for cataloging metadata, since before the adoption of the Dewey Decimal Classification in 1876 and the Universal Decimal Classification in 1895, and continue to drive the adoption of new standards and best practices for a wide variety of content types. Publishers have long needed supply chain metadata to help them sell their content, but this metadata has traditionally been proprietary and not interoperable. Repositories are not streamlined either and lack a systematic approach to metadata creation, although Dublin Core Metadata may be the most common schema to follow. Researchers have, more or less, been guided in their creation of metadata by the other three groups and need to collaborate with the other three groups to create research documentation and create metadata for access and discovery.

For all of these stakeholders in the scholarly communications lifecycle, metadata processes rely on heavily manual or semi-manual processes for creation and dissemination of metadata. No single powerful best practice, standard, or software system drives the metadata ecosystem. The environment reflects a fragmented and organic growth that was originally designed around print-based workflows. Workflows for metadata do not always produce desired discovery; unclear ownership over metadata creation can lead to gaps in coverage; and very few fields are consistently required or used for discovery, challenging the promises often made about metadata. In many cases, answers in this survey reflect the haphazard development of systems creating and using metadata; nevertheless, several issues are common to all stakeholder groups who replied

to the survey. Everyone reported problems with the large amount of time needed for submitting metadata, inadequate knowledge about the impact of metadata for discovery, and frustration about manual entry of metadata into repositories, catalogs, and publication systems.

### 5.2. Researcher Metadata Considerations

One goal of this survey, and the charge of the Researcher Communications Project within Metadata 2020, was to identify levels of researcher knowledge about metadata for publication, discovery of research, and data sets. The answers in the researcher survey reflect a limited awareness of metadata in all of these areas (Researcher Survey Q6 and Q15). However, when asked about how to evaluate different types of research outputs, responses point to fields with rich metadata (Researcher Survey Q11 and Q12).

The other audiences surveyed maintain metadata for their own uses, which can leave out the researchers' needs. For example, researchers might not receive adequate support from publishers to increase their competency with creating and using metadata for data sets, publications, or other research projects (Publisher Survey Q3). Librarians, meanwhile, appear to have many demands on their time to deal with incomplete metadata as provided to them by publishers (Library Survey Q4 and Q5). They are often underutilized by researchers as resources to help them fully leverage metadata for publications or data sets.

Answers in the repository manager survey reflect a mixed picture of metadata issues in repositories. The number of repositories is growing and technologies for repository administration are developing fast; however, repository metadata is the least standardized of any metadata required by the stakeholder groups surveyed here.

In the cross-stakeholder analysis, the respondents' views on the most important metadata fields reflect different priorities than expected by the Project. The respondents were asked to select the most important metadata fields in a list of 28 fields. Of the top 11 fields identified by the four stakeholder groups, publishers, researchers, and repository managers all strongly endorsed the DOI and year and date of publication as the two most important pieces of metadata, with the abstract and author ORCID being third and fourth most important pieces. This means that two of the most important pieces of metadata are persistent identifiers (PIDs). Librarians also identified the DOI and year and date of publication as the two most important pieces of metadata, but not nearly with the same amount of consensus. Moreover, librarians' third and fourth most important pieces of metadata were the volume/issue number and the language of the content; abstracts and author ORCIDs were fifth and sixth respectively. It is unclear exactly why there is a discrepancy about the value of abstracts and ORCIDs between librarians and the other three groups, but given librarians' strong need to educate researchers about metadata, it seems like there are opportunities for librarians to better understand the needs of their researcher patrons.

### 5.3. Next Steps: Future Studies

Having completed this survey on metadata with the different stakeholder groups, we believe there are several avenues for further studies about improving metadata in the scholarly communications life cycle. We must acknowledge the limitations of the convenience sample of respondents who were predominantly from English-speaking countries and represent the organizations that control much in the scholarly communications ecosphere. Future work would benefit from targeted recruitment from underrepresented areas of the world and stakeholders



and should include questions about less common but important items such as nontraditional works and indigenous knowledge.

### **5.3.1. *Persona perspectives are important and multidimensional***

As work progressed with Metadata 2020's project groups, it became clear that the original stakeholder groups encompassed multiple roles. The work completed by other projects within the Metadata 2020 effort resulted in identifying four types of personas—creators, custodians, curators, and consumers—that cut across all of the stakeholder groups (Metadata 2020 Personas, 2020). These personas better encapsulate how metadata is handled in the scholarly communications life cycle and these lenses can better situate how researchers interact with metadata. Future work on metadata from the perspective of personas will align roles with metadata life cycles that play a role for each stakeholder group.

### **5.3.2. *Metadata workflows***

The survey reinforced Metadata 2020 workshop discussion about metadata workflows. Metadata moves through various systems in complex and idiosyncratic ways, but the various workflows have not been identified or studied in ways that can help researchers learn more about where their metadata goes once it enters a platform or a publisher's website. This survey did not address the receipt of metadata for further use, but this is another crucial piece of understanding metadata pipelines.

### **5.3.3. *Metadata as an essential part in applying the FAIR Principles***

The survey revealed that more work can be done to provide a vision of an ecosystem that supports the FAIR principles (Wilkinson, Dumontier et al., 2016) and that also streamlines certain disjointed and nonautomated workflows across stakeholder groups. Researchers will benefit from publishers, librarians, and repository managers having more interoperability in metadata.

### **5.3.4. *Specific metadata elements***

The survey includes a lot of data on which metadata elements are rated as important by the groups, and many of these findings are interesting for further studies. ORCIDs, for example, could be explored for different types of uses by researchers if further functionalities could be added. The "Abstract" metadata element is also an interesting theme for further study, and the Initiative for Open Abstracts (I4OA; Initiative for Open Abstracts, 2021) is already leading the way with expanding the availability of abstracts.

### **5.3.5. *Increased standardization of metadata***

There may also be a need for the Journal Article Tagging Suite (JATS), which was intentionally developed as a descriptive model, to be revisited to make JATS terms more prescriptive (Journal Article Tag Suite, 2020). This would enable content to be tagged in a more consistent way, making it more consistent and predictable for improved machine-readability.

### **5.3.6. *The importance of controlled vocabularies for metadata***

Controlled vocabularies have an impact on the visibility and discoverability of researcher outputs. This study focused less on controlled vocabularies and more on the schema, such as Dublin Core, that use them. More work can be done to understand the perceived value of and usage of controlled vocabularies, as well as examining tools for enriching research output with controlled vocabularies and ontologies. We are aware of the ongoing evolution of tools

and techniques for ontologies and controlled vocabularies, such as the SPAR Ontologies (Peroni & Shotton, 2018).

### **5.3.7. Streamlining metadata deposits for researchers**

Researchers are increasingly struggling with administrative demands associated with creating metadata for funding applications, publications, and data sets and yet they are the most familiar with the best way to describe these objects to support findability and use by other researchers. More work is needed to see what types of partnerships researchers can foster with data repositories and academic libraries to streamline metadata creation. Additionally, as more researchers become aware of the FAIR Principles (Wilkinson et al., 2016) and their utility, the practices of asking for new metadata during the publishing process should be evaluated to increase ease of use and compliance. Further, only the researchers mentioned bibliographic management tools such as Endnote, Refworks, Zotero, and Mendeley. There are opportunities to integrate future metadata components to enhance such tools, one such opportunity is to use semantic web technologies to enrich relationships across scholarly domains.

### **5.3.8. Persistent identifiers (PIDs) not specifically addressed in our survey**

While the use of PIDs continues to increase and the application of standards to control their use across communities gains more acceptance, new opportunities continue to arise. Some PIDs do have metadata embedded in them, for example, more work can be done to enrich PIDs with more metadata.

### **5.3.9. Semantic web technologies not addressed in the survey**

Massive opportunities exist to use metadata created in the publishing life cycle to create semantic metadata. There is evidence that repository managers are the stakeholder group most familiar with the metadata schema and encoding languages that facilitate ontological metadata (JSON, RDF, Schema.org), but metadata is only beginning to be used for these purposes in the research lifecycle. Researchers do not usually operate in this space, meaning that education is the first step toward developing metadata intended for a semantic web platform.

## **ACKNOWLEDGMENTS**

The authors would like to thank all respondents to the survey, as well as the project teams for Metadata 2020 who helped to develop and refine the question sets. We are also grateful to Rachael Lammey of Crossref for generating the online survey forms. And, finally, many thanks to the Metadata 2020 instigators and maintainers who have supported this project, including Metadata 2020's project managers Clare Dean and Laura Paglione.

## **AUTHOR CONTRIBUTIONS**

The article has been a common project by all the authors, but specific author contributions have been the following: Kathryn A. Kaiser: conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing—original draft, writing—review & editing; Michelle Urberg: conceptualization, formal analysis, investigation, methodology, visualization, writing—original draft, writing—review & editing; Maria Johnsson: formal analysis, investigation, methodology, writing—original draft, writing—review & editing; Jennifer Kemp: investigation, writing—review & editing; Alice Meadows: conceptualization, formal analysis, investigation, writing—original draft, writing—review & editing; Laura Paglione: formal analysis, project administration, visualization, writing—review & editing.

### COMPETING INTERESTS

The authors have no competing interests.

### FUNDING INFORMATION

This study was performed with funding support from the MetaData 2020 project. The use of the survey instrument was made possible by Crossref.

### DATA AVAILABILITY

The data used in the study leading to this paper are available with supporting materials to facilitate their use. A list of these materials can be found in the Appendix.

### REFERENCES

- About Metadata. (2020). <https://www.metadata2020.org/about/>. Accessed December 4, 2020.
- 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. <https://doi.org/10.1629/uksg.441>
- Bull, S., & Quimby, A. (2016). A renaissance in library metadata? The importance of community collaboration in a digital world. *Insights*, 29(2), 146–153. <https://doi.org/10.1629/uksg.302>
- Crossref. (2020). <https://www.crossref.org/>. Accessed December 4, 2020.
- DataCite. (2020). <https://datacite.org/>. Accessed December 4, 2020.
- Flynn, E. A. (2013). Open access metadata, catalogers, and vendors: The future of cataloging records. *Journal of Academic Librarianship*, 39(1), 29–31. <https://doi.org/10.1016/j.acalib.2012.11.010>
- Gregg, W. J., Erdmann, C., Paglione, L. A. D., 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>
- Harvey, M. J., McLean, A., & Rzepa, H. S. (2017). A metadata-driven approach to data repository design. *Journal of Cheminformatics*, 9(4). <https://doi.org/10.1186/s13321-017-0190-6>, PubMed: 28184255
- Imbue Partners. (2017). Industry leaders' perspectives on the digital transformation. *Journey in Publishing*. <https://www.buchmesse.de/files/media/pdf/whitepaper-industryleaders-perspectives-frankfurter-buchmesse.pdf>. Accessed December 4, 2020.
- Initiative for Open Abstracts. (2021). <https://i4oa.org/>. Accessed June 5, 2021.
- Jennings, L. (2017). Metadata for research discovery and management. *Catalogue and Index*, 187, 5–8. [https://purehost.bath.ac.uk/ws/portalfiles/portal/155978250/ci187\\_metadata\\_for\\_research\\_data\\_discovery\\_and\\_management\\_lizz\\_jennings.pdf](https://purehost.bath.ac.uk/ws/portalfiles/portal/155978250/ci187_metadata_for_research_data_discovery_and_management_lizz_jennings.pdf). Accessed December 4, 2020.
- Jones, J. (2018). The rise and fall of the Great Library of Alexandria: An animated introduction. *Open Culture*. <https://www.openculture.com/2018/08/rise-fall-great-library-alexandria-animated-introduction.html>. 2, 28. Accessed December 4, 2020.
- Journal Article Tag Suite. (2020). <https://jats.nlm.nih.gov/index.html>. Accessed December 4, 2020.
- Kaiser, K., Urberg, M., Johnsson, M., Kemp, J., Meadows, A., & Paglione, L. (2021). Metadata 2020 metadata usage survey methods and results summary (Version 0.1.0). *Zenodo*. <https://doi.org/10.5281/zenodo.4666086>
- Kemp, J., Dean, C., & Chodacki, J. (2018). Can richer metadata rescue research? *The Serials Librarian*, 74(1–4), 207–211. <https://doi.org/10.1080/0361526X.2018.1428483>
- Kratz, J. E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLOS ONE*, 10(2), e0117619. <https://doi.org/10.1371/journal.pone.0117619>, PubMed: 25706992
- Leonelli, S. (2016). The disruptive potential of data publication. *Notes and Records*, 70, 393–395. <https://doi.org/10.1098/rsnr.2016.0036>, PubMed: 30124274
- McIntosh, S., & Griffel, D. (1967). ADMINS – for computer based library management. Center for International Studies, M.I.T. <https://dspace.mit.edu/bitstream/handle/1721.1/85122/826372845.pdf>. Accessed December 4, 2020.
- Metadata 2020, Kaiser, K., Urberg, M., Johnsson, M., Kemp, J., Meadows, A., & Paglione, L. (2021). Metadata 2020 metadata usage survey questions (Version 0.1.0). *Zenodo*. <https://doi.org/10.5281/zenodo.4666192>
- Metadata 2020 Personas. (2020). <https://www.metadata2020.org/resources/metadata-personas/>. Accessed December 4, 2020.
- ORCID. (2020). <https://orcid.org/>. Accessed December 4, 2020.
- Paglione, L., Kaiser, K., Urberg, M., Johnsson, M., Kemp, J., & Meadows, A. (2021). Raw data: An international, multi-stakeholder survey about metadata awareness, knowledge, and use in scholarly communications [Excel spreadsheet]. <https://doi.org/10.5061/dryad.4f4qrj8s>
- Peroni, S., & Shotton, D. (2018). The SPAR ontologies. *Proceedings of the 17th International Semantic Web Conference (ISWC 2018)*. [https://doi.org/10.1007/978-3-030-00668-6\\_8](https://doi.org/10.1007/978-3-030-00668-6_8). Accessed March 29, 2021.
- Poole, A. H. (2016) The conceptual landscape of digital curation. *Journal of Documentation*, 72(5), 961–986. <https://doi.org/10.1108/JD-10-2015-0123>
- PubMed. (2020). <https://pubmed.ncbi.nlm.nih.gov/>. Accessed December 4, 2020.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>, PubMed: 26978244

## **APPENDIX: SUPPLEMENTARY MATERIALS**

### **A.1. Survey Methods and Summary Results**

This appendix represents an analyzed data set of the raw data collected for the Metadata 2020 survey. The purpose of the data summary is to draw attention to the unique responses of each stakeholder group. The authors of this article jointly analyzed and developed this analysis (Kaiser, Urberg et al., 2021).

### **A.2. Survey Questions**

This document records the finalized version of the survey that was distributed, following IRB approval, to a broad number of communities. The questions were answered by participants self-identifying as researchers, publishers, librarians, and repository managers. The purpose of these questions was to assess how metadata is understood by stakeholders in the scholarly communications life cycle. To the designers' best knowledge, no other survey had previously attempted to analyze knowledge about, and perceptions of metadata associated with publication. The survey was conceived as a primary output by the Metadata 2020 Researcher Communications Project and the survey instrument was developed with the assistance of stakeholder groups active in Metadata 2020 (Metadata 2020 Survey Questions 2021).

### **A.3. Raw Survey Data**

This document contains the Excel spreadsheet with all of the raw, unanalyzed responses to the Metadata 2020 Survey. Each stakeholder group is represented with its own tab. In addition, one tab contains all of the responses, another the demographics for the survey. The responses in these tabs may include incomplete responses. Two other tabs contain comparisons of how each stakeholder group has self-prioritized metadata fields and metadata schema, which are discussed in this paper further (Paglione et al., 2021).