Document generated on 08/04/2025 6:43 a.m.

Evidence Based Library and Information Practice

Do Institutional Repository Deposit Guidelines Deter Data Discovery?

Shawn W. Nicholson and Terrence B. Bennett

Volume 16, Number 3, 2021

URI: https://id.erudit.org/iderudit/1082670ar
DOI: https://doi.org/10.18438/eblip29913

See table of contents

Publisher(s)

University of Alberta Library

ISSN

1715-720X (digital)

Explore this journal

Cite this article

Nicholson, S. & Bennett, T. (2021). Do Institutional Repository Deposit Guidelines Deter Data Discovery? *Evidence Based Library and Information Practice*, *16*(3), 2–17. https://doi.org/10.18438/eblip29913

Article abstract

Objective – This study uses quantitative methods to determine if the metadata requirements of institutional repositories (IRs) promote data discovery. This question is addressed through an exploration of an international sample of university IRs, including an analysis of the required metadata elements for data deposit, with a particular focus on how these metadata support discovery of research data objects.

Methods – The researchers worked with an international universe of 243 IRs. A codebook of 10 variables was developed to enable analysis of the eventual randomly derived sample of 40 institutions.

Results – The analysis of our sample IRs revealed that most had metadata standards that offered weak support for data discovery—an unsurprising revelation in view of the fact that university IRs are meant to accommodate deposit and storage of all types of scholarly outputs, only a small percentage of which are research data objects. Most IRs seem to have adopted metadata standards based on the Dublin Core schema, while none of the IRs in our sample used the Data Documentation Initiative metadata that is better suited for deposit and discovery of research datasets.

Conclusion – The study demonstrates that while data deposit can be accommodated by the existing metadata requirements of multi-purpose IRs, their metadata practices do little to prioritize data deposit or to promote data discovery. Evidence indicates that data discovery will benefit from additional metadata elements.

© Shawn W. Nicholson and Terrence B. Bennett, 2021

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

Evidence Based Library and Information Practice

Research Article

Do Institutional Repository Deposit Guidelines Deter Data Discovery?

Shawn W. Nicholson
Associate Dean for Digital Initiatives
Michigan State University Libraries
East Lansing, Michigan, United States of America
Email: Nicho147@msu.edu
https://orcid.org/0000-0002-2144-3578

Terrence B. Bennett
Business / Economics Librarian
R. Barbara Gitenstein Library
The College of New Jersey
Ewing, New Jersey, United States of America
Email: tbennett@tcnj.edu
https://orcid.org/0000-0002-1469-0271

Received: 20 Jan. 2021 **Accepted**: 5 July 2021

© 2021 Nicholson and Bennett. This is an Open Access article distributed under the terms of the Creative Commons-Attribution-Noncommercial-Share Alike License 4.0 International (http://creativecommons.org/licenses/by-nc-sa/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly attributed, not used for commercial purposes, and, if transformed, the resulting work is redistributed under the same or similar license to this one.

DOI: 10.18438/eblip29913

Abstract

Objective – This study uses quantitative methods to determine if the metadata requirements of institutional repositories (IRs) promote data discovery. This question is addressed through an exploration of an international sample of university IRs, including an analysis of the required metadata elements for data deposit, with a particular focus on how these metadata support discovery of research data objects.

Methods – The researchers worked with an international universe of 243 IRs. A codebook of 10 variables was developed to enable analysis of the eventual randomly derived sample of 40 institutions.

Results – The analysis of our sample IRs revealed that most had metadata standards that offered weak support for data discovery—an unsurprising revelation in view of the fact that university IRs are meant to accommodate deposit and storage of all types of scholarly outputs, only a small percentage of which are research data objects. Most IRs seem to have adopted metadata standards based on the Dublin Core schema, while none of the IRs in our sample used the Data Documentation Initiative metadata that is better suited for deposit and discovery of research datasets.

Conclusion – The study demonstrates that while data deposit can be accommodated by the existing metadata requirements of multi-purpose IRs, their metadata practices do little to prioritize data deposit or to promote data discovery. Evidence indicates that data discovery will benefit from additional metadata elements.

Introduction

Just as every designer knows that form follows function, data professionals adhere to the dictum that documentation drives discovery. University based institutional repositories (IRs) continue to play an evolving and expanding role in the scholarly communication ecosystem, including the collection, organization, and dissemination of digital data objects. To remain relevant within this evolving ecosystem, university IRs need to support a common language that advances data discovery—not only across academic institutions, but throughout the wider research data network. A first and crucial step in promoting this common language is the design of deposit forms and guidelines for the metadata that accompanies digital data deposit, which is essential for discovery, reuse, and interoperability—the fundamental elements of the FAIR (**f**indable, **a**ccessible, **i**nteroperable and **r**eusable) Guiding Principles first articulated by Wilkinson et al. (2016). While these principles advocate machine action, human readability and full understanding of means of access remains important.

This paper reports on the results of our empirical exploration across an international sample of university IRs to analyze the required metadata elements for data deposit. Specifically, we examined IR deposit forms and guidelines to determine comparable fields as mapped against the Dublin Core schema, with a particular focus on how these guidelines support the requirements and expectations for data discovery within and across diverse academic disciplines.

Literature Review

A diverse and rich body of literature exists across the three lines of inquiry—institutional repositories, open access movements, and metadata requirements—that wonderfully intertwine to undergird our motivations for this research. Lynch (2003) got right to the heart of the matter by observing, "The development of institutional repositories emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication" (p. 327). The article clearly and importantly framed the IR as a service that is greater than the sum of its software and hardware parts. Additional points of emphasis included the importance of standards development and targeted metadata handling, with the latter clearly signaling to IR developers then, as now, the need for broad organizational commitment. Writing only a few years after Lynch, Green and Guttmann (2007) reminded readers of the relatively long history of discipline/domain specific digital repositories emanating from the social science data

community. They promoted the idea that the IR and discipline/domain repository developers, despite different missions and roles, must find ways to work in concert so as to fully support the research community while concomitantly optimizing data stewardship.

Shortly thereafter, Salo (2008) caused a stir with a frank assessment of IRs by plainly stating, "Most repositories languished understaffed and poorly-supported, abandoned by library and institutional administrators, scoffed at by publishers, librarians, and open-access ideologues" (p. 99). Salo painted a grim picture of self-archiving practices, and dimmed the bright promise of authors agreeably inputting needed metadata. The article helpfully and hopefully concluded with a series of ideas to advance IR goals that would engender success. Chief among these was advice for easing deposit through simplifying the input forms and metadata requirements.

When considered as a whole, this early IR literature might easily be mapped to the Hype Cycle made popular by the information technology firm Gartner, specifically tracking the curve that begins with the "innovation trigger" and continues through toward the end of 2010 when the idea of IRs seemed to rest in the "trough of disillusionment" (Gartner, n.d.).

In parallel to and at times clearly intertwined with the IR discussions are the arguments for and against open access (OA). OA not only promotes unfettered access to content, but also offers broad benefits to scholarly practice by enabling replication, thus reducing duplicate data collection efforts, and accelerating scientific progress. The early 2000s witnessed a series of OA statements and pronouncements from broad coalitions; notable among these are the Budapest Open Access Initiative and the Organisation for Economic Co-operation and Development (OECD) Final Communique. The Budapest statement (2002) advocated for "free and unrestricted online availability" (para. 1) to peer reviewed journal literature and, importantly, pointed up the value of self-deposit. The OECD Communique (2004) illuminated the importance of documentation to make data available and accessible internationally. Investigating the development of open access journal publishing, Laakso et al. (2011) documented a rapid growth in article output from the early 1990s through 2009 while also observing a demonstrable increase in the number of journals providing OA.

With OA journals filling up a growing number of IRs, an emerging need arose to facilitate efficient discovery and dissemination of the content. The Open Archives Initiative (OAI) gained prominence (Lagoze & Van de Sompel, 2003). The OAI's chief aim was to promote interoperability through standards-based exposure and exchange of metadata. While libraries have had a long history of generating metadata records for catalogues and indexes, it was becoming clear that the role of researchers and practitioners also had to be taken into account for the growing number of digital repositories to be successful. Robertson (2005) noted the importance of including "skills from computer science and learning technology as well as LIS, together with enthusiasts and from a great diversity of other disciplines as well" (p. 296).

Moulaison Sandy and Dykas (2016) surveyed a random sample of administrators of US-based IRs included in the OpenDOAR (Directory of Open Access Repositories) registry. The authors concluded that staffing, standards, and systems combined to enable quality metadata. Above all, qualified staff proved to be most crucial. Many authors were beginning to investigate the impact of quality metadata upon discovery across a broad range of fields. Giuliani et al. (2016) declared that "metadata production is still perceived as a complex, tedious and time-consuming task. This typically results in little metadata production and can seriously hinder the objective of facilitating data discovery" (p. 239). Amplifying this observation, Radio et al. (2017) noted that

> The proliferation of research datasets and their availability in various repositories require metadata that provides sufficient context and organizational clarity to enable their use. However, datasets come in myriad forms, structures, and relationships. As characteristics of datasets vary across disciplines, it is reasonable to suggest that the methods by which they are discoverable by metadata should be informed by the considerations unique to differing research areas. (p. 161)

Recent years have not witnessed any dwindling of scholars, funders, and national policy-making bodies investigating how open access literature, IRs, and interoperable metadata interplay. Plan S, the Europe-backed program, is a set of principles that ensure open and immediate access to funded research publications. It was first launched by cOAlition-S in 2018 (cOAlition-S, n.d.). The US Office of Science and Technology Policy (2020) call for characteristics of data repositories "ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves" (p. 3086).

As higher education institutions, and especially libraries, continue to explore ways to best provide research data services to their research communities, the use of digital repositories for storing and providing access to datasets continues to evolve. With a focus only on institutions deploying the Digital Commons IR software, Manninen (2018) sought to assess the ease of locating datasets in these IRs, and the metadata standards employed in depositing the datasets. Finding no universally applied standard and noting that "56 unique fields were identified from the 15 example data items," the study concludes by reminding the reader that through "robust metadata, curated research data repositories will be discoverable, usable, and interoperable into the future independent of the repository platform" (p. 10). The study by Kim et al. (2019) of metadata practices across 20 repositories in three academic disciplines provided the initial impetus for our present study of academic IRs. We borrow heavily from their framework and methodology, particularly, the idea that documentation—data deposit forms and attendant guidelines—"performs the dual purposes of defining a contract between depositors and repositories and gathering information about the deposited data" (p. 843). We are especially drawn to the manner in which these authors promote the concept of data reuse. We want to further the conversation by exploring aspects of discovery—a necessary precursor to reuse.

In sum, the rich and varied contemporary scholarly conversation around open access, IRs, and metadata suggests that these ideas have moved along Gartner's Hype Cycle, through the "slope of enlightenment" and heading toward the "plateau of productivity" (Gartner, n.d.). The purpose of our project is to advance this conversation with the addition of an empirically-based descriptive component.

Aims

This paper reports on the results of our exploration across an international sample of university-based IRs to analyze the required metadata elements for data deposit. Specifically, we examine IR deposit forms and guidelines to determine comparable fields as mapped against the Dublin Core schema, with a particular focus on how these guidelines support the requirements and expectations for data discovery within and across diverse academic disciplines. Our aim was to explore the following questions:

1. Is a deposit form that allows or enables author-supplied metadata present, and how does the deposit guidance describe metadata requirements?
2. What is the prevalence and uniformity of metadata standards?
3. Is there use of controlled vocabulary, and evidence of mediated deposit?

4. Is there evidence of deposited research data?

Methods

The study sought to analyze IRs from an international cohort. To generate a universe of IRs, we first made use of the OpenDOAR API to target Singapore, Australia, the United Kingdom, New Zealand, and Hong Kong—an English-language cohort. OpenDOAR is a directory of open access repositories that adhere to specific criteria for inclusion, such as comprising academic content and being freely available. The results (institution name, IR URL, country, and software name) were exported as a JSON structured file, and then the OpenRefine tool was used to format as a comma separated value (CSV) for ready analysis. The US universe was drawn directly from the Moulaison Sandy and Dykas (2016) article. Those 50 US-based IRs were simply copied and pasted into the master CSV file. The resulting universe of IRs by country of origin is shown in Table 1.

Table 1
Universe, Institutional Repositories (IRs) by Country

Country	**Number of IRs**	**Percentage of IRs**
Australia	83	34%
Hong Kong	7	3%
New Zealand	17	7%
Singapore	6	2%
United Kingdom	80	33%
United States	50	21%
Totals	243	100%

After using MS Excel to assign a random number to each row of the master CSV, we sorted the universe in ascending numerical order. Using the methodology from Kim et al. (2019) as broad inspiration as well as some targeted elements specific to our inquiry, we developed a coding scheme. Beginning with two IRs randomly selected from the universe, we performed time trial and simplified interrater agreement testing to determine variance in code assignments. Our aim was to arrive at appreciable consistency, not an absolute correct code for each case. Only minor modifications were necessary to finalize the coding scheme (see Appendix). Using the randomized numbers, we identified the first 40 results to arrive at a 16.5% sample. The sample tracks reasonably well to the originating universe, as shown by comparing Table 1 to the sample 40 repositories by country in Table 2.

Once the sample was created, including a direct URL associated with each IR, we systematically browsed each IR's author and user guidelines, depositor instructions, FAQs and other supporting documentation in search of key terms, recording coded variables in a shared Google Sheet. Google Sheets was chosen for ease of use and shared quality control capabilities. Author one coded IRs 1-20 and author two coded IRs 21-40, with regular discrepancy checks for discussion and resolution throughout the coding period.

Results

Metadata Standards in Deposit Forms, Guidelines and Sample Records

We first looked for the presence of a deposit form (or self-deposit form) that allows or enables author-supplied metadata to accompany deposits into an IR. Such a form was easily discoverable in only four of the IRs in our sample, while more than half did not include a form. The remaining IRs required potential depositors to log in with a username and password, so if a deposit form was available, we

Table 2
Sample, Institutional Repositories (IRs) by Country

Country	Number of IRs	Percentage of IRs
Australia	9	23%
Hong Kong	2	5%
New Zealand	3	7%
Singapore	2	5%
United Kingdom	14	35%
United States	10	25%
Totals	40	100%

could not view it. However, supplementary guidelines for depositors (apart from a form) were found for more than half of the IRs (including half of those with password-protected access for depositors), thereby providing us with a fuller picture of the deposit process. Among these guidelines, more than two-thirds discussed or described metadata requirements for materials submitted to the IR. Ultimately, we discovered some mention of metadata requirements, either directly within a deposit form or described in supplemental guidelines, for just under half of the 40 IRs examined. Finally, we looked at one or more sample records from each IR to glean additional information about metadata standards. From IRs for most of the institutions at which deposit forms or guidelines yielded no metadata information, we were able to discern something about metadata standards by examining the sample record. And among the institutions whose IR deposit forms or guidelines provided some initial guidance about metadata requirements, our examination of the sample record offered additional clarification of metadata standards at nearly half of these.

Institution Size and Inclusion of Datasets in IR Content

While *our* focus was specifically on the presence of research data in IRs, such data storage is unlikely to be the primary focus of an IR. With the exception of repositories that are created for the specific purpose of research data storage and management (one of which was included in our sample), IRs tend to emphasize deposit of articles or preprints, working papers, book chapters, reports, and other text-based scholarly outputs. An examination of deposit forms or guidelines (or other descriptive information) from the IRs in our sample revealed that just over half included any mention of data or datasets among the types of material to be stored in the IR.

To complement Table 2, we gathered enrollment information for the hosting institution of each of the IRs in our sample, as shown in Table 3. We used this information to examine the association (if any) between the size of the institution and the apparent acknowledgment of research data as a type of record that could be found in the institution's IR. We discovered that among the

21 institutions in our sample whose IRs noted the specific presence of datasets (about half of the total sample), less than one-third of these were small- or medium-sized institutions, while the remaining two-thirds were large institutions with enrollments greater than 15,000.

Table 3
Sponsoring Institution Enrollment Size

Total Enrollment	Category	Number of Institutions
<5000	Small	6
5001-15,000	Medium	7
>15,001	Large	27

Metadata Schema

Among the IRs in our sample, nearly all had records that clearly reflected Dublin Core (DC) metadata. Of these, only a small number reflected weak or moderate use of DC, represented by six or fewer identifiable DC elements; the remaining majority represented full adoption of DC, as evidenced by the presence of more than six DC elements (and more than half of these had ten or more identifiable DC elements).

It is noteworthy to mention that from our examination of deposit forms, guidelines, and sample item records, we found no evidence that any of the IRs in our sample had adopted the standards for describing research data specified by the Data Documentation Initiative (DDI), "an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences" (DDI Alliance, n.d.).

From our analysis, we were able to determine that just over half of the IRs in our sample maintained uniformity across item records by imposing some sort of controlled vocabulary onto their metadata elements. Of the remaining IRs, a few allowed for metadata tagging that did not adhere to a controlled list of elements, while there was insufficient evidence from the remaining IRs [n=15] for us to determine if a controlled vocabulary was in use.

Mediated Deposit

Deposit of items into the IR was either fully controlled by librarians or other IR staff, or partially mediated (self-deposit with administrator review) at more than half of the IRs in our sample. Not surprisingly, two-thirds of the institutions that had a mediated deposit process were large institutions (Table 3) where, presumably, IR staff was likely to be larger and, therefore, able to take on the responsibility of overseeing the deposit of records into the IR. Only three of the IRs in our sample appeared to support total self-deposit with no administrative review, while we were unable to determine if mediated deposit was present at nearly one-third of the IRs in our sample.

Discussion

As evidenced from the results of the studies cited above, the consensus conclusion is that IRs are doing an inadequate job of promoting and enabling data discoverability; our study offers additional support for this conclusion. Integral to discoverability is the presence of descriptions and metadata. Metadata matters, and as Meadows asserts in a 2019 *Scholarly Kitchen* post, it could very well help save the world! While it is easy to delight in this hyperbolic blog post title, the importance of descriptive elements that enable findability and interoperability should be evident. Meadows' crucial contention that metadata creates efficiencies is illuminated through four key actors: creators, curators, custodians, and consumers. In Figure 1, we provide a graphic representation of Meadows' actors and their roles, and find resonance with Meadows' behavioral nudge that each actor

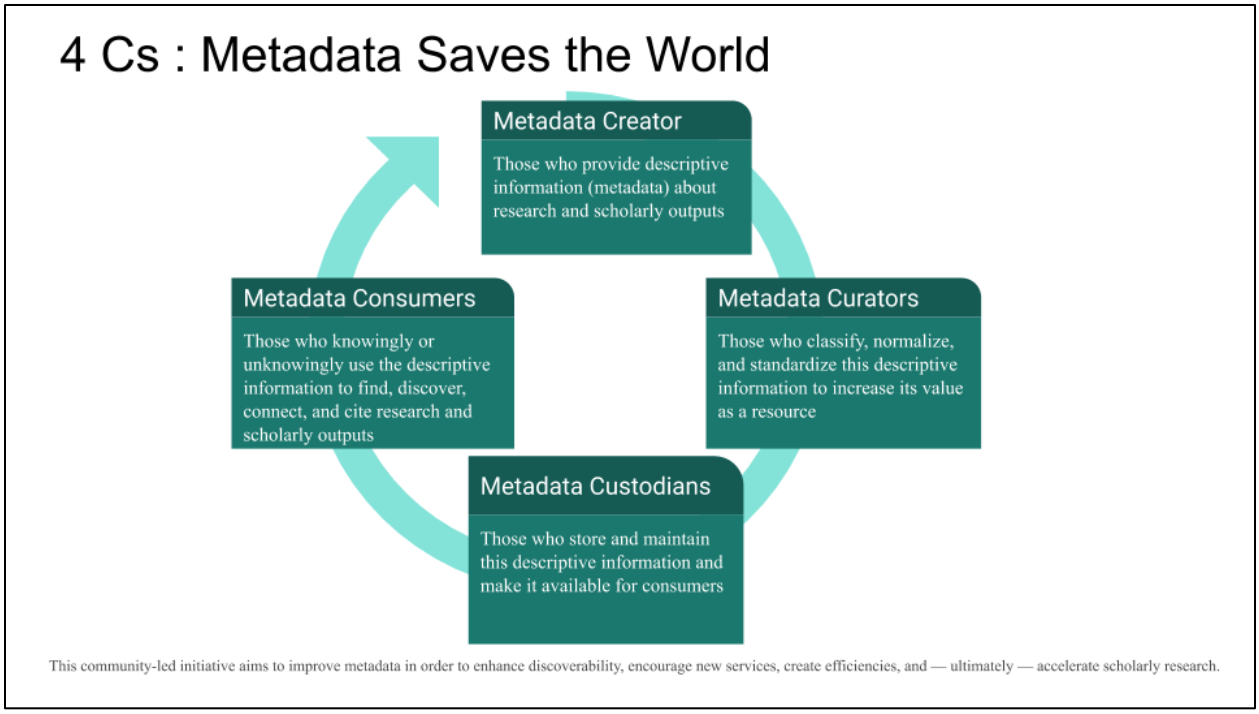

Figure 1
Four Cs of metadata, adapted from Meadows (2019).

shares "collective responsibility to ensure that [metadata] is the best it can be" (2019).

Whereas Kim et al. (2019) focused on the role of metadata in data-deposit requirements to enable reuse of data from discipline-based repositories, we sought to build upon their study by examining the role of metadata to promote data discovery in multipurpose (and multidisciplinary) academic IRs. In their analysis of data deposit requirements compiled from previous studies, Kim et al. noted that "methodology" is included among the five most common data deposit metadata elements in discipline-specific repositories (p. 845). Since methodology is not a DC metadata element, it is reasonable to intuit that the metadata employed by the non-discipline-repositories in our study is even less effective in enabling data discovery.

From their study of the data curation practices in IRs, Lee and Stvilia (2017) noted "a dearth of research, on identifier metadata quality, uses and practices for research data in the context of IRs" (p. 2), as these IRs relied too heavily on simple DC metadata elements, which fail to accommodate the complexity and diversity of datasets. The lack of dataset-specific metadata is connected to issues with reusing, sharing, and searching the data. Our observations reinforced these findings, particularly, as noted, none of the IRs in our sample used a more granular schema such as DDI. As Garnett et al. (2017) state, DC is in wide use because "by default, most OAI portals serve Dublin Core metadata" (p. 208). The Open Archives Initiative Protocol for Metadata Harvesting standard's chief aim is to promote interoperability through standards-based exposure and exchange of metadata (Open Archives Initiative, n.d.). The reliance on DC elements for discovery, as our results suggest, omits important data-specific metadata. Garnett et al. (2017) refer to these as "structural metadata elements that describe dataset variables, questions, concepts, categories, values, etc. that accompany the physical dataset" (p. 206).

From the results of our study, it seems unlikely that data-seeking researchers will develop any reliance on serendipitous searching for data discovery. Using the definition of discovery put forth by Schonfeld (2014) as "the process and infrastructure required for a user to find an appropriate item" (p. 3), we would have to concede that while IRs may support limited discoverability of text-based academic outputs (depending on how well they are configured for search-engine optimization), scholars in pursuit of existing datasets to support their research are likely to be disappointed by irrelevant or nonexistent search results within IRs. This would perpetuate the practice of general data-seeking only within discipline-based data repositories, with researchers only occasionally turning to an academic IR in pursuit of a known item (or at least an expected-to-exist item).

And it is probably unfair to consider this an unacceptable result. To remain relevant, an academic IR also has to remain current, so the first priority of its administrators is to ensure ongoing deposit (with a focus on more commonly occurring text-based scholarly outputs) in order to avoid the dormancy and abandonment predicted by Salo (2008) and subsequently reinforced by others. For example, in 2011, Giesecke observed that IRs still can appear to be "a solution in search of a problem," and that continued evidence of nonparticipation by faculty scholars "would make one ask why should institutions reallocate resources to create and maintain an institutional repository" (p. 540). And as recently as 2017, Tillman observed that "Salo's assessment of the flaws in repository strategies as then practiced ... and her recommendations for next steps ... remain relevant nine years later" (p. 2). With support devoted to storage and access, there may be few resources (including professional staff) remaining to focus on other considerations such as detailed metadata mark-up. Here again we are girded by Meadows' contention that good metadata requires actors in multiple roles; the custodians and curators might, ideally, effectively engage data creators (and possibly consumers) towards the goal of improving metadata. Except for IRs at some large and well-resourced institutions, there will always be constraints from resource limitations. Thus, for most IRs, there remains little expectation that staff expertise will be devoted to the creation of a robust discovery overlay—especially for data, which, from what we've observed, might only comprise a handful of total records in the IR.

While it would require a separate study to gain evidence of researcher behavior, it nevertheless seems unreasonable to infer that these circumstances are likely to provoke potential data discoverers into rattling the IR administrators with demands for improved metadata to support data discovery, as most researchers probably know that they can turn to their discipline-based data repositories when discovery needs to happen. As these disciplinary repositories have already established themselves as a primary venue for data storage, it would be difficult (and probably pointless) for academic IRs to compete in that space. As Pirolli (2016) notes, "information foraging theory assumes that people adapt their information-seeking behavior to maximize their rate of gaining useful information to meet their ongoing goals" (para. 1). Thus, the existing low expectations for data discovery in IRs is likely to create a self-perpetuating cycle of inaction: if there is no expectation that researchers will turn to IRs for data discovery, then there is no impetus for IR administrators to make enhancements to metadata standards and practices that will enable data discovery. The small population of IR data depositors (compared with depositors of text-based research outputs) are also not poised to raise an outcry, because they're likely only using the IRs for data deposit when it is a convenient way to meet publishers' data management requirements. So the only remaining cage-rattlers are the authors of studies such as ours—and we are unlikely to capture the attention of IR administrators in a way that would effect significant changes in policy or practice.

Notwithstanding this state of current practice, the community continues to grapple with roles and goals, though not without divergent opinion. Writing in late 2020 as we were completing this manuscript, Shearer and Kingsley, representing the Confederation of Open Access Repositories, addressed the evolving discussion around criteria for selecting data repositories. Their blog post was in direct response to the Data Repository Selection document produced by representatives from journals and publishers who were already working as part of the FAIRsharing Community (FAIRsharing.org, n.d.). A crucial point of contention rests on the degree to which the community can accept concentration of repository services. It is our hope that all of the key actors mentioned by Meadows (2019)—creators, curators, custodians, and consumers—will continue to engage in this discussion.

Limitations

While our sample was large enough to afford meaningful observations, a larger sample size is always desirable. Our results show, for example, a strong relationship between institution size and a propensity to acknowledge directly that research data may be stored (if not discovered) in an IR. Specifically: the larger the institution, the more likely the mention of datasets. While further study would be necessary before we could suggest that these results reflect causation, it is nonetheless reasonable to infer that research institutions with higher enrollments also have more faculty (and others who conduct research), which increases the likelihood that *any* type of material (including data) makes its way into the IR. A related unsurprising observation is that mediated deposit is more likely to occur at larger institutions. Again, an interrelationship does not prove causation; further study would be needed to arrive at any stronger conclusion.

Of necessity, we had to limit our analysis to IRs at academic institutions in English-speaking countries. It's possible that the most innovative and forward-thinking advances in IR development are happening at non-English-language institutions, or that these institutions have developed a more advanced culture of data deposit for reuse (and therefore greater advances in data discovery). Moreover, if we had used different criteria to generate a sample—such as, for example, starting with a universe of IRs in which we had first identified existing records for datasets—then our observations would have focused on practices among IRs that are known to accommodate data deposit (and may therefore be expected to enable data discovery by employing data-specific metadata).

Conclusions

There is an extensive body of literature around academic IRs as enablers of open access, including access to datasets for replication and reuse. The purpose of our study is to present an empirical analysis of the role of metadata in promoting data discoverability, a necessary precursor to reuse. By examining the deposit forms, guidelines, and output of an international sample of academic IRs, our observations reinforce the findings of others: while data deposit can be *accommodated* by the existing metadata requirements of these multi-purpose IRs, their metadata practices do little to *prioritize* data deposit or to promote data discovery.

It is reasonable to expect that this status quo will perdure. Academic IRs have to address many competing priorities, most of which are skewed toward deposit, storage, discovery and retrieval of text-based scholarly outputs. Therefore, the effort and expertise that must be expended to maintain an IR will inevitably continue to favor articles and other texts. Within this reality, academic IRs will perpetuate their established, and expected, function as convenient containers for a limited subset of research-related datasets.

Author Contributions

Shawn Nicholson: Conceptualization, Data curation, Investigation, Methodology, Software,

Writing - original draft, Writing - review & editing **Terrence Bennett:** Formal analysis, Investigation, Writing - original draft, Writing - review & editing

Acknowledgments

The authors wish to thank Heather Moulaison Sandy and Felicity Dykas for sharing their previously compiled list of US-based IRs, which proved to be a time-saving starting point for deriving our sample for this study. We are also very grateful to the two anonymous reviewers for their thoughtful and thorough comments on an earlier draft of this article.

References

Budapest Open Access Initiative. (2002, February 14). *Budapest open access initiative*. https://www.budapestopenaccessinitiative.org/read

cOAlition-S. (n.d.). *Plan S: Making full and immediate open access a reality.* https://www.coalition-s.org/

DDI Alliance. (n.d.). *Data Documentation Initiative (DDI).* https://ddialliance.org/

FAIRsharing.org. (n.d.). *FAIRsharing.org*. https://fairsharing.org/

Garnett, A., Leahey, A., Savard, D., Towell, B., & Wilson, L. (2017). Open metadata for research data discovery in Canada. *Journal of Library Metadata*, *17*(3-4), 201-217. https://doi.org/10.1080/19386389.2018.1443698

Gartner, Inc. (n.d.). *Gartner hype cycle.* https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

Giesecke, J. (2011). Institutional repositories: Keys to success. *Journal of Library Administration*, *51*(5-6), 529-542. https://doi.org/10.1080/01930826.2011.589340

Giuliani, G., Guigoz, Y., Lacroix, P., Ray, N., & Lehmann, A. (2016). Facilitating the production of ISO-compliant metadata of geospatial datasets. *International Journal of Applied Earth Observation and Geoinformation*, *44*, 239-243. https://doi.org/10.1016/j.jag.2015.08.010

Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services: International Digital Library Perspectives, 23*(1), 35-53. https://doi.org/10.1108/10650750710720757

Kim, J., Yakel, E., & Faniel, I. M. (2019). Exposing standardization and consistency issues in repository metadata requirements for data deposition. *College & Research Libraries, 80*(6), 843-875. https://doi.org/10.5860/crl.80.6.843

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PLoS ONE, 6*(6), e20961. https://doi.org/10.1371/journal.pone.0020961

Lagoze, C., & Van de Sompel, H. (2003). The making of the open archives initiative protocol for metadata harvesting. *Library Hi Tech, 21*(2), 118-128. https://doi.org/10.1108/07378830310479776

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE, 12*(3), e0173987. https://doi.org/10.1371/journal.pone.0173987

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *portal: Libraries and the Academy*, *3*(2), 327-336. https://doi.org/10.1353/pla.2003.0039

Manninen, L. (2018). Describing data: A review of metadata for datasets in the digital commons institutional repository platform: Problems and recommendations. *Journal of Library Metadata, 18*(1), 1-11. https://doi.org/10.1080/19386389.2018.1454379

Meadows, A. (2019, June 11). Better metadata could help save the world! *The Scholarly Kitchen.* https://scholarlykitchen.sspnet.org/2019/06/11/better-metadata-could-help-save-the-world/

Moulaison Sandy, H., & Dykas, F. (2016). High-quality metadata and repository staffing: Perceptions of United States-based OpenDOAR participants. *Cataloging & Classification Quarterly, 54*(2), 101-116. https://doi.org/10.1080/01639374.2015.1116480

Office of Science and Technology Policy. (2020). Draft *desirable characteristics of repositories for managing and sharing data resulting from federally funded or supported research*. *Federal Register*, *85*(12), 3086-3087. https://www.govinfo.gov/content/pkg/FR-2020-01-17/pdf/2020-00689.pdf

Open Archives Initiative. (n.d.). *Protocol for Metadata Harvesting (PMH).* https://www.openarchives.org/pmh/

Organisation for Economic Co-operation and Development. (2004, January 30). *Science, technology and innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at ministerial level, 29-30 January 2004—final communique.* https://www.oecd.org/sti/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm

Pirolli, P. (2016). Information foraging. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems*. Springer. https://doi.org/10.1007/978-1-4899-7993-3_205-2

Radio, E., Rios, F., Oliver, J. C., Hickson, B., & Wallace, N. (2017). Manifestations of metadata structures in research datasets and their ontic implications. *Journal of Library Metadata, 17*(3-4), 161-182. https://doi.org/10.1080/19386389.2018.1439278

Robertson, R. J. (2005). Metadata quality: Implications for library and information science professionals. *Library Review, 54*(5), 295-300. https://doi.org/10.1108/00242530510600543

Salo, D. (2008). Innkeeper at the roach motel. *Library Trends, 57*(2), 98-123. https://doi.org/10.1353/lib.0.0031

Schonfeld, R. C. (2014). Does discovery still happen in the library? Roles and strategies for a shifting reality. *Ithaka S+R Briefing.* https://sr.ithaka.org/wp-content/uploads/2014/10/SR_Briefing_Discovery_20140924_0.pdf

Shearer, K., & Kingsley, D. (2020, November 24). Input to "Data repository selection: Criteria that matter." *Confederation of Open Access Repositories (COAR) News-Updates.* https://www.coar-repositories.org/news-updates/input-to-data-repository-selection-criteria-that-matter/

Tillman, R. K. (2017). Where are we now? Survey on rates of faculty self-deposit in institutional repositories. *Journal of Librarianship and Scholarly Communication, 5*(1), Article eP2203. https://doi.org/10.7710/2162-3309.2203

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Appendix

Code	Category	Label/Description
1.0	**Current status of IR**	Is directory info up to date?
1.1	Yes -- IR found	If OpenDOAR URL (especially from Moulaison Sandy & Dykas list) is broken/outdated; check OpenDOAR for updated URL
1.2	No IR found	Use this code only if search/browse yields no IR
2.0	**Deposit Form**	Where user inputs metadata
2.1	Y	
2.2	N	
2.3	DK / can't determine	likely password protected/controlled
3.0	**Guidelines Doc**	Text (separate from deposit form) describing self-deposit process and needed metadata
3.1	Y	
3.2	N	
3.3	Y, but	does not specifically describe metadata fields
3.4	DK / can't determine	Don't know / password protected
4.0	**Deposit form or guidelines mention data**	Specifically mentions dataset (as standalone item, or as supplement to article, book chapter, working paper, etc.)
4.1	Y	
4.2	N	
4.3	N, but data deposit noted elsewhere	
4.4	DK / can't determine	No sample "data" record
5.0	**Sample Record**	Use of sample output to determine metadata fields
5.1	Still nothing	Nothing in form or guidelines and record provides inadequate additional detail
5.2	At least something	Nothing in form or guidelines yet record provides a modicum additional detail

5.3	Adds nothing additional	Metadata adequately described elsewhere, sample output added nothing new
5.4	Adds more	Some metadata described elsewhere, sample output offers more detail
5.5	Muddies	Metadata adequately described elsewhere, but sample output muddled/contradicted guidelines
6.0	**DC Elements Explicit**	Dublin Core https://www.dublincore.org/specifications/dublin-core/dces/
6.1	Title	
6.2	Creator	
6.3	Subject	
6.4	Description	
6.5	Publisher	
6.6	Contributor	
6.7	Date	
6.8	Type	
6.9	Format	
6.10	Identifier (DOI, ORCID)	
6.11	Source	
6.12	Language	
6.13	Relation	
6.14	Coverage	
6.15	Rights	
6.16	No apparent DC element	
6.17	Unknown	Can't see form, no documentation
7.0	**DDI Mentioned**	Is DDI specifically noted for data description?
7.1	Y	
7.2	N	
8.0	**Controlled Vocabulary**	Is controlled vocabulary used for some descriptors?

8.1	Y	
8.2	N	
8.3	Unknown	
9.0	**Institution Student enrollment**	
9.1	<5000	Small
9.2	5,000-15,000	Medium
9.3	>15,000	Large
10.0	**Metadata Mediated**	Are descriptors vetted, corrected before IR record is published?
10.1	No	Totally self-deposit
10.2	Sort of	Self-deposit with administrator review
10.3	Totally	The entire deposit process is handled by librarians or other IR staff
10.4	DK	Can't tell / unknown