Journal of eScience Librarianship

Volume 8
Issue 2 *Special Issue: Research Data and Preservation (RDAP) Summit 2019*

Article 7

2019-12-23

Peer Review of Research Data Submissions to ScholarsArchive@OSU: How can we improve the curation of research datasets to enhance reusability?

Clara Llebot
Oregon State University

Et al.

ISSN 2161-3974

JeSLIB 2019; 8(2): e1166
doi:10.7191/jeslib.2019.1166

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Peer Review of Research Data Submissions to ScholarsArchive@OSU: How can we improve the curation of research datasets to enhance reusability?

Clara Llebot[1] and Steven Van Tuyl[2]

[1] Oregon State University, Corvallis, OR, USA
[2] Academic Data Science Alliance

Abstract

Objective: Best practices such as the FAIR Principles (Findability, Accessibility, Interoperability, Reusability) were developed to ensure that published datasets are reusable. While we employ best practices in the curation of datasets, we want to learn how domain experts view the reusability of datasets in our institutional repository, ScholarsArchive@OSU. Curation workflows are designed by data curators based on their own recommendations, but research data is extremely specialized, and such workflows are rarely evaluated by researchers. In this project we used peer-review by domain experts to evaluate the reusability of the datasets in our institutional repository, with the goal of informing our curation methods and ensure that the limited resources of our library are maximizing the reusability of research data.

Methods: We asked all researchers who have datasets submitted in Oregon State University's repository to refer us to domain experts who could review the reusability of their data sets. Two data curators who are non-experts also reviewed the same datasets. We gave both groups review guidelines based on the guidelines of several journals. Eleven domain experts and two data curators reviewed eight datasets. The review included the quality of the repository record, the quality of the documentation, and the quality of the data. We then compared the comments given by the two groups.

Correspondence: Clara Llebot: clara.llebot@oregonstate.edu
Keywords: data curation, data peer-review, data publication

Disclosures: The substance of this article is based upon a panel presentation at RDAP Summit 2019. Additional information at end of article.

JeSLIB 2019; 8(2): e1166
doi:10.7191/jeslib.2019.1166

Abstract Continued

Results: Domain experts and non-expert data curators largely converged on similar scores for reviewed datasets, but the focus of critique by domain experts was somewhat divergent. A few broad issues common across reviews were: insufficient documentation, the use of links to journal articles in the place of documentation, and concerns about duplication of effort in creating documentation and metadata. Reviews also reflected the background and skills of the reviewer. Domain experts expressed a lack of expertise in data curation practices and data curators expressed their lack of expertise in the research domain.

Conclusions: The results of this investigation could help guide future research data curation activities and align domain expert and data curator expectations for reusability of datasets. We recommend further exploration of these common issues and additional domain expert peer-review project to further refine and align expectations for research data reusability.

JeSLIB 2019; 8(2): e1166
doi:10.7191/jeslib.2019.1166

Introduction

Managing research data with care, and sharing research outcomes with the public (including publications and datasets) is becoming an expectation in academia. Several governments and funders emphasize this expectation with policies and regulations to increase access to federally funded research, such as the 2013 White House Office of Science and Technology Policy memorandum in the United States (Holdren 2013) or the Open Research Data Pilot, part of the European Union Horizon 2020 program (Directorate-General for Research & Innovation 2016). Journals that require publication of data underlying the research are also becoming increasingly common (Dearborn et al. 2018). The goal of these policies and recommendations is clear: maximize access to research data to ensure that research is more reproducible, and make data that is more reusable. Requiring a data management plan with research grant proposals and requiring or recommending sharing the data generated during the research will usually achieve this. The success of the implementation of these practices, however, is not clear. Research suggests that funders and journals are ineffective at convincing researchers to share their data (Savage and Vickers 2009), and that when data is shared, it often has issues (e.g., lack of documentation, inadequate formats) that make it difficult to reuse the data (Van Tuyl and Whitmire 2016, Naudet et al. 2018).

To solve the problem of quality in data sharing, the research community continues to develop best practices. Two examples of discipline agnostic data sharing guidelines can be found in the FAIR data principles (Wilkinson et al. 2016) and the DATA (Discoverable, Accessible, Transparent and Actionable) rubric (Van Tuyl and Whitmire 2016). There are also many examples of discipline specific recommendations, such as the work of White et al. (2013) for ecology, Griffin et al. (2017) for life sciences, and Leberg and Neigel (1999) in the field of population genetics.

The attitudes and practices of researchers with respect to data management and data sharing has been studied at length by many (cf. Van Tuyl and Michalek 2015, Akers and Doty (2013), Whitmire et al. (2015), Rolando et al. (2013), Borghi and Van Gulick (2018), Johnston and Jeffryes (2013), Whiley and Mischo (2016), Federer et al. (2015), and Carlson and Stowell-Bracke (2013)). However, several authors have expressed concerns related to data sharing, like ethical considerations (Merson et al. 2016), the availability of data (Vines et al. 2014), and the complexity of data sharing in general (Borgman 2012). The question about data quality has been noted by some (Merson et al. 2016), but very few studies have investigated it. A notable exception is the study by Naudet et al. (2018), where they tested the reproducibility of data published in two biomedical journals with a strong data sharing policy. Naudet and coauthors found that when datasets were available, it was possible to reproduce the original results for most of them. They also found that the procedures used to share the data were very heterogeneous, and that authors did not use any standard to share their data, which complicated the reuse of the data. Another study addressing data quality (Van Tuyl and Whitmire 2016) created a rubric to evaluate the quality of shared data. They saw that almost none of the data produced by researchers at their institution scored highly on discoverability, accessibility, transparency, and actionability (DATA).

The heterogeneity of research data and of science domains adds a level of complexity to the problem. Different disciplines have varying expectations regarding data sharing, and value it differently (Tenopir et al. 2011). They also interpret “quality” data differently. In addition,

standards and best practices differ across scientific disciplines. Formal peer review of datasets has been proposed as a possible solution to the challenge of producing quality data that complies with discipline specific standards (Lawrence et al. 2011, Costello et al. 2013). Research on data peer review indicates that researchers do not expect data to be peer reviewed, but that data that is peer reviewed is recognized by researchers as being of higher quality and trustworthy (Kratz and Strasser 2015). Formal peer review of data sets is performed in data journals (Candela et al. 2015), but it is uncommon in journals that focus on publication of scientific discoveries, even if they require the data to be publicly available. At the repository level, a common form of review is the data curator review. This process varies depending on the repository and the institution (Johnston et al. 2017, Koshoffer et al. 2018), but it usually involves a data curator who is commonly not a domain expert, with a focus on reviewing the metadata and documentation associated with the dataset, and on data formats. Finally, a large proportion of repositories and journals allow self deposit and do not require any form of review.

The best practices mentioned above are helpful, but for the most part are domain agnostic and focused on the data curator, rather than on the domain expert who will potentially reuse the data. This article evaluates how domain experts and data curators interpret data quality differently. By becoming aware of these differences we expect to learn if the guidelines and criteria the research community has been developing are useful for dataset reuse. We do that by asking domain experts to peer review datasets in our institutional repository, and compare them with evaluations done by data curators.

Materials and Methods

Datasets in ScholarsArchive@OSU have not been curated uniformly over the past decade due to personnel, program, and policy changes across that time period. Over the past decade, our methods for curating dataset deposits to ScholarsArchive@OSU have changed as our program has grown and we have gained experience and insight into the needs of our stakeholders. An initial set of dataset deposits in ScholarsArchive@OSU, dating from 2009, were deposited prior to the launch of our formal Research Data Services program, and did not undergo any explicit or systematic curation.

Between 2010 and 2016 data curators reviewed datasets submitted to ScholarsArchive@OSU for overall coherence and cleanliness (see Tidy Data, below) including removing extraneous files, ensuring that included files were actionable by users (i.e. the files could be opened), and clearly documenting the dataset either in a readme file or other documentation. The data curators requested changes to researchers and, while not required, they strongly encouraged creation of documentation.

After 2016, Oregon State University's research data services underwent methodological changes, resulting in a somewhat altered approach to review of research data deposits, including the following items:

- Tidy data: datasets are expected to be reasonably organized. This loosely defined requirement involves checking that there isn't duplicated data, or several versions of some data. We check that data includes a consistent structure (e.g. variables in each column) throughout all the files.

- Formats: data must be in actionable formats, cross platform formats, and open formats when possible. The data curator discusses with the researcher he best formats for their data that will maximize preservation. The most common example is transforming excel files to csv files.

- Documentation: it is mandatory that all datasets in ScholarsArchive@OSU include a documentation file. A template is suggested as guidance, but researchers are free to create the documentation that makes the most sense for their dataset.

Review Guidelines

We created a set of peer-review guidelines based on an amalgamation of review guidelines provided by a number of academic data journals. We included a variety of journals, with different peer review models, and we included discipline specific journals as well as discipline agnostic journals. Table 1 shows the journals that we relied upon, open access journals which had detailed review guidelines specific for datasets.

Table 1: Characterization of the main journals that we used to develop review guidelines.

Journal Name	Discipline specific	Peer review	Publisher	Citation
Scientific Data	No	Single blind peer review	Nature	Scientific Data, 2019
Geoscience Data Journal	Yes	Single blind peer review	Royal meteorological Society and Wiley	Geoscience Data Journal, 2019
Earth System Science Data	Yes	Interactive public peer review	Copernicus Publications	Earth System Science Data, 2019
F1000Research	No	Open peer review	F1000Research publishing platform	F1000Research, 2019
Open Archaeology Data	Yes	Single blind peer review	Ubiquity press	Open Archaeology Data, 2019

Discipline specific: journal accepts only articles within a particular discipline, e.g. geosciences (Yes) or accepts papers in all disciplines (No). **Peer review**: peer review model followed by the journal. Single blind peer review: Anonymous reviewers review the content of the paper. Authors are not anonymous. If paper passes peer-review, it gets published. Interactive public peer review: Initial review by a topic editor. Citeable preprint of paper is published. Article goes through discussion phase where interactive comments can be posted by designated referees (anonymous and non anonymous) and by all interested members of the scientific community (non anonymous). Authors revise manuscript and resubmit. Topic editor accepts/rejects. Accepted paper is published, with a reference to all the review materials, that are openly available. Open peer review: Article is published after submission. Non anonymous reviewers are invited. Registered users and authors can also submit comments. Authors are encouraged to publish revised versions. Articles that pass peer-review are indexed. All the versions and comments are publicly available. **Publisher**: Publisher of the Journal. **Citation**: citation of the source, to be found in the reference list.

The review guidelines were divided into four blocks. We asked domain experts to consider the quality of the repository record (Is the repository metadata sufficiently descriptive? Are keywords and/or abstract helpful? Are there elements that could be added?), the quality of documentation associated with the dataset (Is there sufficient documentation? Is there robust contact information? Is there enough information describing the content and format of the data? Are methods described in sufficient detail? Can all internal references be resolved to real entities? Is there information about your rights to reuse the dataset?), the quality of the dataset itself (Are the data recorded in a useful format? Are error limits, spatial and temporal coverage good enough? Are values physically possible and plausible? Are there missing data?), and to score the overall usability of the dataset.

We created a questionnaire based on the review guidelines, that can be found as supplementary material for this article. Most of the questions required a written response, but we also asked reviewers to rate the quality of the record, the documentation, the data, and the overall dataset in a scale from 0 (low quality) to 10 (high quality). The interpretation of the data provided below relies heavily on qualitative data, but quantitative data are used on occasion to reinforce our arguments. For both the domain expert scores and the data curator scores, we calculated the median and standard deviation of each score to summarize the quantitative information provided via reviews.

Peer Review of Datasets

We selected each research dataset that had been deposited to Oregon State University's institutional repository, ScholarsArchive@OSU, and collected the names and contact information of the depositors of the data, either from readme files, documentation, or by searching the university directory for contact information. Datasets for which we were not able to identify an appropriate contact e-mail were excluded from the study. Of a total of 86 datasets, 59 were considered for the study. The datasets that were not included consist of one dataset in embargo, a dataset published by one of the authors of this study, and 25 datasets for which we were unable to find contact information for the author.

We asked the authors of these 59 datasets to provide contact information for 3 or more possible domain specialist reviewers for their dataset, and received contact information of 61 potential reviewers for datasets (29% response). We contacted the domain experts, and asked if they would be willing to review the dataset that they had been recommended for. Of the 61 potential domain expert reviewers, 22 responded, 11 of them indicating willingness to review 8 different datasets (3 datasets were reviewed by 2 different domain experts). 18% of the domain expert reviewers had e-mail addresses that indicated affiliation with Oregon State University. We then provided each domain expert with a set of review guidelines in the form of an online questionnaire (see supplementary materials), and a link to the dataset in question as published in ScholarsArchive@OSU. Names and e-mails of domain expert reviewers were removed from the dataset to protect their confidentiality. No other personal information was collected from the domain expert reviewers. The reviewers were aware of the identities of the dataset creators, as this information is part of the record that they had to evaluate.

To provide a library data curator review of the dataset, the authors also reviewed each dataset using the same guidelines. We then used our review, alongside the domain expert reviews to make an overall determination about the quality of the dataset and to evaluate differences

between expectations of domain experts and data curators. The authors performed the review before reading the results of the domain expert reviews.

Results

The 8 datasets reviewed are listed in Table 2. Three of the datasets were reviewed by two reviewers, so we have 11 reviews by domain experts and 8 reviews by data curators.

Table 2: Summary table of datasets reviewed for this project.

Title	Year publication	URI	Domain	Domain expert reviews
Host and Habitat Index for Phytophthora Species in Oregon: Supplemental Spreadsheet.	2014	dx.doi.org/10.7267/N99G5JR4	Botany and plant pathology	1
Dataset for Status of the European Green Crab, Carcinus maenas, in Oregon and Washington coastal Estuaries in 2017	2018	doi.org/10.7267/N9VD6WM4	Integrative Biology	2
First Leaf Emergence Force of Winter Wheat	2018	doi.org/10.7267/N91834NH	Crop and Soil Science	2
Dataset for Finding Submerged Sites: An Exploration of Shoreline and Environmental Change in Oregon's Yaquina River Basin using GIS Predictive Modeling	2018	doi.org/10.7267/zybt-bt36	Anthropology	1
Methodology for comparing wood adhesive bond load transfer using digital volume correlation	2018	doi.org/10.7267/N9QV3JQF	Wood Science and Engineering	1
Blended sea level anomaly fields with enhanced coastal coverage along the U.S. West Coast. Version 3	2017	doi.org/10.7267/N9639MWJ	Oceanic and Atmospheric Science	2
Remarkable solar panels Influence on soil moisture, micrometeorology and water-use efficiency - database	2017	doi.org/10.7267/N9639MWJ	Biological and Ecological Engineering	1
Benthic Habitat Mapping: eCognition Rule Set	2018	doi.org/10.7267/N9GF0RP1	Civil and Construction Engineering	1

The range of overall quality for the datasets included in this study was generally high from domain expert (median 9.0, standard deviation 1.5; Table 3) and data curator (median 9.0, standard deviation 1.4). Quality of specific dataset records and documentation in the repository followed a similar pattern of high scores. While these overall scores suggest broad agreement on usability of these datasets between domain and curator reviews, a number of common areas of difference are worth highlighting here. They have been grouped in 5 themes: insufficient description (datasets are not adequately described by the record or documentation), links (regarding information being conveyed via links), duplication of effort (when the same information is recorded in more than one location), domain expertise (issues concerning the different levels of expertise of domain experts and data curators), and repository functionality (topics related to the way ScholarsArchive@OSU works).

Table 3: Statistics calculated for the quantitative questions of our questionnaire. DE: Domain Experts. DC: Data Curators. The DE statistics are calculated over all the 11 reviews. The DC statistics are calculated over 8 datasets, because data curators only reviewed each dataset once. Data Curators did not feel qualified to evaluate the quality of the data, according to the reviewer guidelines, blank value indicated by a N/A.

	Record		**Documentation**		**Data**		**Overall**	
	DE	DC	DE	DC	DE	DC	DE	DC
Median	8.5	9.0	8.5	9.0	9.0	N/A	9.0	9.0
Standard Deviation	1.2	0.9	1.3	2.6	1.5	N/A	1.5	1.4

Insufficient Description

Domain experts stated that descriptive information about the dataset is critical to a user's ability to understand what the data is and whether it is potentially useful for their application. Several domain experts made it clear that the need for high quality description applies to the metadata record, the documentation, and to the dataset itself, and deficiencies in any of these areas can result in confusion about whether and how the data may be relevant to their research.

Various domain experts described abstracts, keywords, and title as useful to determine the interest of the dataset (i.e. a domain expert commented "The record and especially the abstract clearly outline the nature and contents of the dataset. The title and keywords are complementary and should capture most searches."), but their usefulness diminished when they were too short, lacked details (e.g. "it could be helpful to include a full species name in the abstract"), or if they were not specific to the dataset (i.e. the abstract described related research, but not the dataset itself). One domain expert identified extra keywords that should have been used to describe one of the datasets.

Lack of information in the documentation file(s) accompanying the dataset was a common complaint. A number of domain experts reported missing methodology (e.g., “QC [quality control] is referred to but without much elaboration.”), information about the authors and their contact information (“No, there is no author contact information given.”), about licenses (“This information [information to determine your rights to reuse the dataset] wasn't apparent.”), and url (“The listed DOI didn't link to the publisher.”) about the dataset. One of the domain experts gave a very thorough description about how this lack of information regarding, in this case, the sampling effort, limits the interpretation that can be done of this data, and the calculation of derived datasets. One domain expert also noted areas where the documentation did not include description of data values that were in the dataset (“I might add brief descriptions of the input data layers”). A few domain experts also expected information about how to use the dataset. For example, the domain expert who looked at a dataset including a model noted a lack of information about the input data which made the task of testing and using the model for other applications difficult. An example of input data was not included in the dataset, and had to be generated by the domain expert to test and review the model.

Imprecise geographic information was highlighted a number of times as a barrier to potential reuse. One domain expert suggested, when possible, to use GPS or latitude and longitude, instead of more general locations like, for example, county level location. Links to polygons for sampling location (e.g. geographic placenames from services like geonames.org) were described by one domain expert as “very helpful”, but in some cases insufficiently resolved to allow a user to understand the spatial elements of the data. One domain expert explained that the lack of detail about where the sampling stations were located could hinder the possibility of reusing the data in another context.

In some cases, domain experts noted that these areas of description work together to create a more complete description of the dataset. For example, in one case, the domain expert noted that the abstract contained methodological information that was not clear from the dataset documentation.

Data curators noted most of the issues related to lack of information in the record that the domain experts identified, such as short or missing abstracts and titles. Unlike the domain experts, data curators described too many keywords in one of the datasets as possibly confusing. Data curators noticed datasets that lacked documentation, but they had a harder time identifying when the information was present but insufficient, or when the precision of the data was inadequate. Data curators valued the presence of scripts used in the analysis of data and considered those to be part of the research methods.

Links

Information about datasets is often provided via links. The most common type of link leads to a related published article, but in our sample, links are also used to indicate locations of sampling stations. Three domain experts noted that some of the information necessary to understand the dataset was available in a related paper, not in the dataset itself. These pieces of information included contextual information (e.g. “the abstract is a little brief but a link is contained to the full article with more than sufficient contextual information”); methodological instructions (“the methods are clearly outlined in the report”) and data processing steps (“I believe [that data processing steps are described with enough detail], at least within the

referenced and linked report"); comprehensive information about all the data present in the record ("Yes [there is a comprehensive description of all the data that is there]—again in the paper"); and contact information of authors ("The associated paper contains contact info."). One domain expert mentioned that the lack of information in the dataset abstract is balanced by the presence of a good abstract in the published article. Most comments suggest that the authors are comfortable using related articles as a source of information to understand the dataset, and that they even expect it. This was confirmed by a domain expert who noted that when they found a lack of methodological information in a dataset, they looked for references to external papers to see if the information was there ("I could not find a reference to a paper or a detailed methods description.") One of the three domain experts described the need to look for information in a related article as something negative, and they do it in the context of finding contact information for the authors of the dataset ("No, there is no author contact information given. I assume I could have looked at the cited papers and found the corresponding author information.").

Data curators recognized that some of the information can be found in related journal articles, and took that into account when assessing the thoroughness of the documentation of the datasets. However, data curators value the presence of dataset specific documentation higher than most domain experts. In particular, the datasets that lacked documentation but had related articles were scored much higher by domain experts than by data curators. Data curators discovered citations in one of the datasets that resolved to inconsistent locations and in some cases did not resolve at all, something missed by the domain experts. Data curators also took into consideration whether the linked documents were accessible and open, as some of the related articles pointed to articles behind paywalls. None of the domain experts commented on that.

Duplication of Effort

Another challenge encountered by domain experts, data curators, and data submitters, was the duplication of effort to provide some required content, and the duplication of location of some of this content. Our template readme file, which is used by most data submitters, requires information such as Authors, Abstract, and Title, which is also required for the metadata record for the dataset deposit. In addition, many data depositors referenced research articles, linked to the dataset record, as a place for data users to find methodological information (as referenced above). It is not entirely clear how this duplication of effort impacts data submission quality, as the combination of information provided in the record, documentation, and linked articles typically was enough to allow the domain expert or data curator to understand the dataset in detail. Some domain experts did note that this basic descriptive information as part of the metadata record was important for their initial understanding of the relevance of the dataset, indicating that providing basic descriptive information on the record is of value.

Domain Expertise

We noted that domain expertise, or lack thereof, was important across all areas of review for datasets. The need for the domain expert to be familiar with the research area of the dataset was apparent when domain experts identified such anomalies as a lack of specificity in the abstract, missing methodology in documentation, and insufficient metadata embedded in the

file-specific metadata. Two specific examples of this need for domain-specific knowledge were the mis-application of a data standard in a specific file format (geographic coordinates in a file "I think the dimension order for the data variables should be T, Y (lat), X (lon) not T, X (lon), Y (lat)") and that one dataset's documentation did not provide sufficient detail for the domain expert to understand the methodology used. In all of these areas, the data curators do not have sufficient domain expertise to properly evaluate the quality of the data, or documentation, at a level of detail that could catch these problems. Conversely, there was some confusion on the part of domain experts in the areas of licensing, rights statements, persistent identifiers, and where specific types of information belong in the metadata record, all areas that tend to fall in the data curator's domain.

Possibly the most important aspect of the need for domain expertise in these reviews was the inability of data curators to evaluate the contents of specific datafiles. Very often, the details of the dataset are opaque to non-domain experts, making it very challenging for data curators to review a dataset for completeness, accuracy, and whether it meets domain standards for research data. Large divergences in the evaluation of datasets by domain experts and data curators were mostly the result of differing expectations for where and how to document the datasets. Domain experts seem more willing than data curators to turn to published articles for details about methods and results generated by the dataset.

Repository Functionality

While not directly related to documentation of datasets, there were a number of issues that arose related to repository functionality that made evaluation of datasets more difficult for a few domain experts. Where core repository functionality was not working well, domain experts were usually not able to separate the repository's behavior from their review of the dataset, generating a lowered opinion of the quality of the dataset. For example, a change in the behavior of a linked open data program in our repository (Geonames) confused some domain experts about why they were not receiving the information they expected from the system. Some domain experts were unable to locate licensing and rights information in the metadata record, making it hard for them to determine their rights to reuse the dataset for research. This highlights a need for thorough user experience testing in the repository.

Discussion

Our study reveals that domain experts and data curators evaluate the reusability of datasets in a similar way. This is demonstrated by the responses of data curators and domain experts, and reflects criteria at the record, documentation, and dataset level. The challenges to the reusability of our particular sample of datasets are described similarly. Insufficient description, especially short abstracts, lack of methodology and contact information in the documentation, and low precision of measurements, are the most common issues. Domain experts and data curators also coincide in describing difficulties when data curation guidance requires that the same information will be duplicated (e.g. abstract needed in the record and in the documentation) and when some of the information is given via links. Data users may rely on a link to a related publications to provide information on contact information, methodology of data collection and data processing, or context of the study. This agreement is good news, as it suggests that the work that data curators are doing, and the standards and best practices that we are using, is probably meeting most of the needs that researchers have, and enabling and

JeSLIB 2019; 8(2): e1166
doi:10.7191/jeslib.2019.1166

facilitating data reuse.

Despite general agreement, there were several issues that data curators and domain experts perceived differently. A better understanding of these may help curators fine tune data curation processes to make them more effective. Domain experts showed they are very comfortable with the idea of using published journals as a source of information to understand the data, none talked about whether the linked documents were open access or not. This is not entirely surprising, as research articles are a product researchers work with continuously, but it may be a symptom of the lack of familiarity that many researchers still have with the concept of sharing data (the facts) separately from research articles (the interpretation of the facts; Brewer 2017). This could also be showing the privilege of the domain experts who belong to institutions where access to publications is not usually a problem. Discussion of these issues of privilege are somewhat common among data curators considering scholarly communications issues in academia, but may be less familiar to domain experts. This highlights a need for libraries to continue to communicate the breadth of issues related to access to resources when discussing data sharing best practices with our domain expert colleagues.

Another difference that appeared in many of the reviews was how domain expertise gave domain experts a very different perspective to evaluate datasets. For example, some domain experts, familiar with best practices in the field, were able to recommend metadata standards of which data curators were not aware. Domain experts were also capable of envisioning specific examples for reusing data, and thus identify information that should have been shared to make a dataset more reusable. Domain experts were able to evaluate when data was not precise enough. From the point of view of the data curator and data curation processes, lack of research domain expertise is a difficult problem to solve, especially for curators of domain-agnostic repositories.

It is unlikely that research data curators will have research data-related expertise at the level of a domain researcher, and it is important to keep this in mind as we set standards and best practices for our data curation and data repositories. While this may seem obvious, it highlights a few specific needs: First, it is important for data curators to honor the limits of their ability to curate a dataset and to set expectations for data depositors that communicate the boundary between what work the curators can/should do and what work the depositor must do. Often, as we see with this study, the curator's ability to engage meaningfully with the dataset can be limited by a lack of expertise and the dialog around ensuring reusability of the dataset needs to shift from the curator back to the domain expert. Likewise, the data curator should assert their expertise into conversation with the researcher, to ensure that broader research data curation considerations are taken into account and so that the expertise of each enhances the quality of the dataset. Asking researchers to envision possible future reuse of one's dataset is one way data curators can encourage the inclusion of information that helps make a dataset reusable.

Second, research data curators should create and engage with communities of practice around data curation in order to share knowledge and benefit from the knowledge of others. The Data Curation Network (https://datacurationnetwork.org; Data Curation Network 2018) is an example of such community building and knowledge sharing, with the aim of using a community network to dive deeper into domain-specific data curation needs. Expansion of this and similar projects can only help define best practices for data curation and help examine the limitations of research data curation not conducted by domain researchers.

Both domain experts and data curators expressed a concern for the validity and longevity of dataset author contact information. While this may seem a minor point, the ability to track down a person who is knowledgeable about a dataset is of high value to dataset users, especially when documentation for a dataset is lacking. Many of the dataset reviews suggested changes to documentation or the dataset that could only be made by someone intimately familiar with the deposited data. Given the above mentioned limitations of curating dataset deposits for a general purpose repository, it is inevitable that datasets will be deposited that require additional information from the depositor or a designated responsible party.

This issue is not new to academic publishing or academic research, and academia has made significant improvements in this area over the past decade. Notably, ORCID (https://orcid.org), and other stable identifiers for researchers, provide tools for stabilizing contact information for researchers. In theory this is a valuable piece of the solution for this problem; it is not clear how widely ORCID has been adopted in research communities or at academic institutions, though adoption is certainly uneven. ORCIDs and other identifiers also must be maintained, often by the researchers themselves, and thus can serve to replicate the problem of keeping contact information up to date for dataset deposits—if the researcher doesn't update their ORCID profile with new contact information, it becomes much more difficult to track them down to answer questions. Indeed, one of the datasets reviewed for this project had this specific issue—the dataset documentation provided links to ORCID profiles for all dataset authors, but the profiles were not up to date.

One key recommendation from this study is that the research data services community create a shared, domain agnostic data curation checklist that defines required, preferred, and optional elements for data curation and documentation. This checklist should include an indication of the boundary (as fuzzy as it might be) between what information/curation the curator can provide and what information/curation the depositor must provide. Research data curation communities have converged on some shared standards over the past decade (e.g. Wilkinson et al. 2016, Van Tuyl and Whitmire 2016) and it seems the field is in a position to collectively build common guidelines and expectations for research data curators and depositors. Of course, shifting to a common set of curation practices may not be immediately practical for all domains, and a tiered or phased approach may be necessary. For example, this study found many researchers are comfortable linking to the methods sections of published research articles as part of their data documentation. In the near term, this type of linking is probably sufficient, especially given the widespread acceptance of the research paper methods section as a reliable and complete source of information. However, we might collectively consider what it would take to meaningfully shift away from this method of documentation, to a method that allows for data documentation independent of published journal articles.

This study suffers from a number of limitations, most of which were unavoidable. First, the number of datasets for which we were able to receive reviews was small. While this makes broad generalizations difficult, this small sample provides useful insight. It would be helpful for other institutions to conduct similar investigations of institutional repositories that host research datasets. Second, almost all of the datasets in this study had been curated at some level before deposit. This may create an environment where domain expert and data curator evaluation of deposited data were more similar than they might have been without prior curation. More widespread investigation of researcher perspectives on the usability of deposited datasets could help create a broader understanding of the quality of existing shared

data and recommended improvements to curation practices. Moreover, investigation of the reusability of datasets deposited to large, generalist repositories (e.g. figshare, Zenodo) and domain specific repositories could bring this discussion to a broader audience and help answer questions about differences in dataset usability across these types of repositories.

Conclusions

This study reveals the value and shortcomings of engaging domain experts in the curation process for research data. Through this investigation, some of the assumptions of our data curation practices, and of broader research data curation communities, have been validated, while others have been thrown into question. We highlight a number of differences in perspective and expectations provided by data curator versus domain expert reviews of deposited datasets.

Ultimately, data curation activities can be insufficient or incomplete if either the data curator or the data producer are wholly responsible for curation activities. Balancing the curation requirements of repositories and data curators against the expectations of data depositors and data users is critical to finding a path forward for research data curation and sharing.

Supplemental Content

Questionnaire
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2019.1166 under “Additional Files”.

Data Availability

The data for this article are the reviews provided by 11 reviewers. It would be impossible to anonymize these reviews, so we are not sharing the data. The article is accompanied by additional materials, that include the questionnaire we used.

Acknowledgements

Thank you to the anonymous reviewers who gave their time to this article and to the datasets discussed therein, as well as the data depositors who agreed to participate in this study. Thank you to all the colleagues at Oregon State University that reviewed the text to make this article stronger and better.

Disclosures

The substance of this article is based upon a panel presentation on ‘Researcher Perspectives’ at RDAP Summit 2019: “Peer Review of Research Data Submissions Study. How can we improve the curation of research datasets to enhance reusability?” available at https://osf.io/uyq6b.

References

Akers, Katherine G., and Jennifer Doty. 2013. “Disciplinary Differences in Faculty Research Data Management Practices and Perspectives.” *International Journal of Digital Curation* 8(2): 5-26. https://doi.org/10.2218/ijdc.v8i2.263

Borghi, John A., and Ana E. Van Gulick. 2018. “Data Management and Sharing in Neuroimaging: Practices and Perceptions of MRI Researchers.” *PLOS ONE* 13(7): e0200562. https://doi.org/10.1371/journal.pone.0200562

Borgman, Christine L. 2012. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63(6): 1059-1078. https://doi.org/10.1002/asi.22634

Brewer, Peter. 2017. “‘Do You Expect Me to Just Give Away My Data?’” *Eos* 98(September). https://doi.org/10.1029/2018EO081175

Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. 2015. “Data Journals: A Survey: Data Journals: A Survey.” *Journal of the Association for Information Science and Technology* 66(9): 1747-1762. https://doi.org/10.1002/asi.23358

Carlson, Jake, and Marianne Stowell-Bracke. 2013. “Data Management and Sharing from the Perspective of Graduate Students: An Examination of the Culture and Practice at the Water Quality Field Station.” *Portal: Libraries and the Academy* 13(4): 343-361. https://doi.org/10.1353/pla.2013.0034

Costello, Mark J., William K. Michener, Mark Gahegan, Zhi-Qiang Zhang, and Philip E. Bourne. 2013. “Biodiversity Data Should Be Published, Cited, and Peer Reviewed.” *Trends in Ecology & Evolution* 28(8): 454-461. https://doi.org/10.1016/j.tree.2013.05.002

Dearborn, Dylanne, Steve Marks, and Leanne Trimble. 2018. “The Changing Influence of Journal Data Sharing Policies on Local RDM Practices.” International Journal of Digital Curation 12(2): 376-389. https://doi.org/10.2218/ijdc.v12i2.583

Directorate-General for Research & Innovation. 2016. “H2020 Programme Guidelines on Fair Data Management in Horizon 2020.” *European Commission*. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Earth System Science Data. “Review criteria.” Accessed October 3, 2019. https://www.earth-system-science-data.net/peer_review/review_criteria.html

F1000Research. “Data Guidelines.” Accessed October 3, 2019. https://f1000research.com/for-authors/data-guidelines

Federer, Lisa M., Ya-Ling Lu, Douglas J. Joubert, Judith Welsh, and Barbara Brandys. 2015. “Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff.” Edited by Jyotshna Kanungo. *PLOS ONE* 10(6): e0129506. https://doi.org/10.1371/journal.pone.0129506

Geoscience Data Journal. “Guidelines for reviewers.” Accessed October 3, 2019. https://rmets.onlinelibrary.wiley.com/hub/journal/20496060/features/guidelines-for-reviewers

Griffin, Philippa C., Jyoti Khadake, Kate S. LeMay, Suzanna E. Lewis, Sandra Orchard, Andrew Pask, Bernard Pope, et al. 2018. “Best Practice Data Life Cycle Approaches for the Life Sciences.” *F1000Research* 6(June). https://doi.org/10.12688/f1000research.12344.2

Holdren, John. 2013. “Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research.” *United States Office of Science and Technology Policy*. https://petitions.obamawhitehouse.archives.gov/petition/require-free-access-over-internet-scientific-journal-articles-arising-taxpayer-funded

Johnston Lisa, and Jeffryes Jon. 2014. “Data Management Skills Needed by Structural Engineering Students: Case Study at the University of Minnesota.” *Journal of Professional Issues in Engineering Education and Practice* 140(2): 05013002. https://doi.org/10.1061/(ASCE)EI.1943-5541.0000154

Johnston, Lisa R., Jake R. Carlson, Patricia Hswe, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. 2017. “Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions’ Data Repository and Curation Services.” *Journal of eScience Librarianship* 6(1): e1102. https://doi.org/10.7191/jeslib.2017.1102

Journal of Open Archaeology. “Editorial Policy.” Accessed October 3, 2019. https://openarchaeologydata.metajnl.com/about/editorialpolicies

Koshoffer, Amy, Amy Neeser, Linda Newman, and Lisa R. Johnston. 2018. “Giving datasets context: a comparison study of institutional repositories that apply varying degrees of curation.” *IJDC* 13(1): 15-34. https://doi.org/10.2218/ijdc.v13i1.632

Kratz, John Ernest, and Carly Strasser. 2015. “Researcher Perspectives on Publication and Peer Review of Data.” *PLOS ONE* 10(2): e0117619. https://doi.org/10.1371/journal.pone.0117619

Lawrence, Bryan, Catherine Jones, Brian Matthews, Sam Pepler, and Sarah Callaghan. 2011. “Citation and Peer Review of Data: Moving Towards Formal Data Publication.” *International Journal of Digital Curation* 6(2): 4-37. https://doi.org/10.2218/ijdc.v6i2.205

Leberg, P. L., and J. E. Neigel. 1999. “Enhancing the retrdevability of population genetic survey data? An assessment of animal mitochondrial DNA studies.” *International Journal of Organic Evolution* 53(6): 1961-1965. https://doi.org/10.1111/j.1558-5646.1999.tb04576.x

Merson, Laura, Oumar Gaye, and Philippe J. Guerin. 2016. “Avoiding Data Dumpsters — Toward Equitable and Useful Data Sharing.” *New England Journal of Medicine* 374(25): 2414-2415. https://doi.org/10.1056/NEJMp1605148

Naudet, Florian, Charlotte Sakarovitch, Perrine Janiaud, Ioana Cristea, Daniele Fanelli, David Moher, and John P A Ioannidis. 2018. “Data Sharing and Reanalysis of Randomized Controlled Trials in Leading Biomedical Journals with a Full Data Sharing Policy: Survey of Studies Published in The BMJ and PLOS Medicine.” *BMJ* 2018; 360:k400. https://doi.org/10.1136/bmj.k400

Rolando, Lizzy, Chris Doty, Wendy Hagenmaier, Alison Valk, and Susan Wells Parham. 2013. “Institutional Readiness for Data Stewardship: Findings and Recommendations from the Research Data Assessment.” *Technical Report. Georgia Institute of Technology*. https://smartech.gatech.edu/handle/1853/48188

Savage, Caroline J., and Andrew J. Vickers. 2009. “Empirical Study of Data Sharing by Authors Publishing in PLoS Journals.” Edited by Chris Mavergames. *PLoS ONE* 4(9): e7078. https://doi.org/10.1371/journal.pone.0007078

Scientific Data. “Guide to referees.” Accessed October 3, 2019. https://www.nature.com/sdata/policies/for-referees

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. 2015. “Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.” Edited by Peter van den Besselaar. *PLOS ONE* 10(8): e0134826. https://doi.org/10.1371/journal.pone.0134826

Van Tuyl, Steve, and Gabrielle Michalek. 2015. “Assessing Research Data Management Practices of Faculty at Carnegie Mellon University.” Journal of Librarianship and Scholarly Communication 3(3): eP1258. https://doi.org/10.7710/2162-3309.1258

Van Tuyl, Steven, and Amanda L. Whitmire. 2016. “Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia.” *PLOS ONE* 11(2): e0147942. https://doi.org/10.1371/journal.pone.0147942

Vines, Timothy H., Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24(1): 94-97. https://doi.org/10.1016/j.cub.2013.11.014

White, Ethan P., Elita Baldridge, Zachary T. Brym, Kenneth J. Locey, Daniel J. McGlinn, and Sarah R. Supp. 2013. "Nine Simple Ways to Make It Easier to (Re)Use Your Data." *Ideas in Ecology and Evolution* 6(2). https://ojs.library.queensu.ca/index.php/IEE/article/view/4608

Whitmire, Amanda L., Michael Boock, and Shan C. Sutton. 2015. "Variability in Academic Research Data Management Practices: Implications for Data Services Development from a Faculty Survey." Edited by Dr Andrew Cox. *Program* 49(4): 382-407. https://doi.org/10.1108/PROG-02-2015-0017

Wiley, Christie. 2016. "Data Management Practices and Perspectives of Atmospheric Scientists and Engineering Faculty." https://doi.org/10.5062/f43x84nj

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Comments and Opinion. Scientific Data*. 3: 160018. https://doi.org/10.1038/sdata.2016.18