IJDC | *Peer-Reviewed Paper*

Giving Datasets Context: A Comparison Study of Institutional Repositories that Apply Varying Degrees of Curation

Amy Koshoffer
University of Cincinnati

Amy E. Neeser
University of California Berkeley

Linda Newman
University of Cincinnati

Lisa R. Johnston
University of Minnesota

Abstract

This research study compared four academic libraries' approaches to curating the metadata of dataset submissions in their institutional repositories and classified them in one of four categories: no curation, pre-ingest curation, selective curation, and post-ingest curation. The goal is to understand the impact that curation may have on the quality of user-submitted metadata. The findings were 1) the metadata elements varied greatly between institutions, 2) repositories with more options for authors to contribute metadata did not result in more metadata contributed, 3) pre- or post-ingest curation process could have a measurable impact on the metadata but are difficult to separate from other factors, and 4) datasets submitted to a repository with pre- or post-ingest curation more often included documentation.

Received 23 October 2017 ~ *Accepted* 20 February 2018

Correspondence should be addressed to Amy Koshoffer, University of Cincinnati Libraries, GMP Library, 2825 Campus Way ML 0153, Cincinnati Ohio 45220-0153. Email: koshofae@ucmail.uc.edu

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v13i1.632

Introduction

Today's research community and data consumers increasingly recognize the value of data as an integral component of research output. It is no longer enough to publish a researcher's interpretation of a study. Publishers and funding agencies encourage, ask, and require researchers to share raw data upon which an interpretation is based (Briney, Goben and Zilinski, 2017; Holdren, 2013; Jones, 2007; Vasilevsky et al., 2017). Certain research communities have developed repository infrastructure to house and provide access to their data, but this is not available in all disciplines. Many libraries have invested resources and expertise to develop institutional repositories (IRs) to preserve and provide access to the scholarly output of their research communities. The IR is often a part of the library's mission and/or is supported by library staff (Heidorn, 2011), and increasingly these IR services are extended to support data. A recent survey of 80 American Research Libraries (ARL) institutions found that 80% had data curation and repository services in place or planned to provide them (Hudson-Vitale and ARL, 2017).

Shared research data that is easily found, accessed, combined with other data, analyzed with new methods and tools, and reused has the potential to expand its impact. To maximize this potential, research data needs context in order to be understood and used by others which can be added by the author or a data curator. "Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle [of usefulness]" (Digital Curation Centre, n.d.). Curators decide how to best describe what data is and how to use it. Because curation choices can vary, standards such as the FAIR data principles, which make data findable, accessible, interoperable and reusable, are increasingly more important (Wilkinson et al., 2016).

Despite these efforts, this study demonstrates the great variation in metadata contributed and documentation for datasets submitted to IRs. This research study compares four academic libraries' approaches to curating the metadata for dataset submissions in their IRs: those with no curation, pre-ingest curation, selective curation, and post-ingest curation. The goal is to understand the impact that curation may have on the quality of user-submitted metadata. The authors formulated the following research questions to understand this impact of curation on research data:

1. How do the metadata elements vary for each institution?
2. How complete is the metadata submission for datasets in each institution repository given the type of curation?
3. Are curated datasets more likely to have documentation associated with the work?
4. Does the number of datasets with DOIs vary given the type of curation?
5. What is the difference in number of keywords associated with each dataset?

These findings will help institutions understand the impact of curation on user-submitted metadata and how to best make use of an institution's limited resources. This study is unique in the comparison of metadata elements at four institutional repositories and the examination of documentation for datasets in those repositories. Future studies

can build on this work with the ultimate goal of determining if curation has the expected benefits of discoverability and reusability.

Literature Review

The majority of library literature focuses on why curation is important and how to best curate to ensure the data are accessible and reusable. Peer (2013) describes a set of curatorial practices, from maintaining, preserving, and adding value to digital research data throughout its lifecycle, that ensure data are accessible. Mannheimer, Sterman and Borda (2016) analyzed data citation counts and data download counts of datasets to determine that the following factors may facilitate reuse: robust data description, non-proprietary file types, and publication in open access repositories.

Others examine the quality of metadata. Rousidis, Garoufallou, Balatsoukas and Sicilia (2015) discuss the operational constraints related to financial resources and human factors that may "impede the effectiveness of several metadata elements" such as the dc.subject metadata element. Rousidis, Garoufallou, Balatsoukas, and Sicilia examined the Dryad research data repository and found quality problems related to the lack of controlled vocabulary and standardisation. Park (2009) determines that accuracy, completeness, and consistency are the most common criteria used in measuring metadata quality, and urged building a common data model that is interoperable across digital repositories. Gavrilis et al. (2015) proposed a robust metadata quality evaluation model that measured metadata quality based on five metrics: completeness, accuracy, consistency, appropriateness, and auditability. Furthermore, Park and Tosaka (2010) suggested mechanisms for building quality assurance into the metadata creation process itself and Walters (2009) proposed a model for using these types of criteria in order to develop a curation program. Margaritopoulos, Margaritopoulos, Mavridis and Manitsaris (2012), on the other hand, developed a metrics system used to measure completeness of metadata as a measure of quality.

In addition to curation and metadata quality, the literature shows that there are different models of deposit, as Koshoffer, Hansen, and Newman (2017) did in their examination of quality of metadata in a self-submission repository. Additionally, Johnston et al. (2017) in forming the Data Curation Network, a cross-institutional staffing model that compared six institutional models of curation, recommended a post-ingest curatorial review workflow to "alleviate any concern about gaining access to datasets that are not publicly available (e.g., behind password protection) or interacting with unfamiliar repository technologies." Finally, Lee and Stvilia (2017) conducted 13 interviews with 15 IR staff members from 13 large research universities in the United States to learn how IR staff members work with researchers to create metadata and readme files for their submissions. They describe the necessary roles played, skills needed, contractions and problems present, solutions sought, and workarounds needed in order to suggest curation best practices.

A review of the literature on this topic shows the importance of curation and quality of metadata, along with suggestions for improving both. And though differing models of deposit have been examined, there is no literature to date that conducts an in-depth comparison of metadata elements and documentation across differing deposit and curation models for datasets. By examining the effectiveness of differing curation models, readers can better incorporate these related findings, such as building a

 doi:10.2218/ijdc.v13i1.632

common data model or metadata quality evaluation model, into their data repository services.

Participating Institutions

The authors represent the following three U.S. academic university libraries: University of Cincinnati (Cincinnati), University of Michigan (Michigan), and University of Minnesota (Minnesota); with data from a fourth institution contributed by a colleague at Oregon State University (Oregon State). Institutions were invited to participate in this study that represented various types of curation models used for their institutional or data-only repository: pre-ingest curation, post-ingest curation, selective curation, or no curation. The repository environments differed for each institution (see Table 1, which shows the type of repository software used; if the repository was a stand-alone data repository or integrated with an institutional repository; the age of the repository; and the total number of datasets in the repository). While each institution supports Digital Object Identifiers (DOIs) for datasets and none added additional keywords, the levels and intensity of curation processing differed for each institution:

- Oregon State supported pre-ingest curation, which required contributors to meet standard levels of description and documentation, and made datasets public only when they met curation standards. DOIs were automatically assigned to submissions.
- Minnesota provided post-ingest curation where a team of six domain-based data curators worked with researchers to bring datasets to suggested levels of standard description and documentation (a required component) before the submission was finalized. DOIs were manually assigned to submissions only after minimum curation standards were met. Curation steps at Minnesota involved appraisal/selection, check/run files (includes code review, review for sensitive information, licensing and rights management checks etc.), working with the author to collect missing files and to create custom documentation (e.g, readme.txt files), metadata augmentation (ie. curators supplement the author-supplied metadata), and file format transformations (Johnston, 2017).
- Michigan selectively curated datasets in cases where researchers willingly participated either before or after deposit. The Data Curation Librarian partnered with subject librarians and interested researchers to prepare their data for deposit into Deep Blue Data when the opportunity arose to do so. Staff also reviewed datasets post deposit by contacting researchers and making changes to the data deposit based on the responses received. Contributors chose to mint a DOI for a work as an optional step after the submission process.
- Cincinnati operated with no formal curation process and would handle issues as they arose. Contributors chose to mint a DOI for a work as a step in the submission process. Access to the work determines the DOI status. In order to mint a DOI, a work must be ‘open access’. Works submitted as ‘embargo’, ‘University of Cincinnati [only]’, or ‘private’ had a reserved DOI that resolved when the contributor made the work public.

Methodology

This study methodology captures a snapshot of the workflow for each repository. Each repository service is continually maturing, responding to its unique user and campus needs. These four institutions have practices that may compare to other institutions, but may not encompass all institutional practices.

Table 1. Institutional repository comparison

Institution	Repository Name	Repository Type	Curation Type	Repository Software	Start Date of Repo	Total datasets as of 10-17-2017
University of Cincinnati	Scholar@UC	General IR	No curation	Hydra Fedora	September 2015	48
University of Michigan	Deep Blue Data	Data-only IR	Selective curation	Hydra Fedora	September 2016	85
University of Minnesota	Data Repository for the University of Minnesota (DRUM)	Data-only IR	Post-ingest curation	DSpace	March 2015	148
Oregon State University	ScholarsArchive@OSU	General IR	Pre-ingest curation	DSpace	February 2005	70

Timeframe of Study

The four study partners selected the 20 most recent datasets submitted as of December 31, 2016. The authors chose not to select a fixed timeframe due to the variation in repository usage and maturity; for example, Minnesota received 50+ datasets in 2016 whereas Cincinnati received fewer than ten. Therefore, each repository analyzed the same number of datasets in each repository and the sample sizes were consistent and comparable.

Data Collection

The authors analyzed the metadata associated with 80 total datasets housed in the four IRs. For most self-deposit IRs, metadata are typically collected from end-users via a web-based submission form and then transformed into machine-actionable elements. Although metadata elements collected in the submission process by all four institutions used Simple Dublin Core, the application of the elements differed slightly. Before comparing the user-submitted metadata for the 80 datasets in the sample, the authors designed a comparison for the metadata schemas. The intent was not to create crosswalks between schemas, but rather to identify the common user-contributed metadata elements. The study analyzed the study analyzed the following information from each IR:

- the metadata element (noting if the field is required or optional);
- the name of the field as displayed on the submission form;

 doi:10.2218/ijdc.v13i1.632

- the order of how the fields display in the form; (results not included);
- the help text provided for each field. (results not included).

The authors compared the metadata elements used by each IR (e.g. 'Author(s)' field from one institution corresponded to the 'Creators' field from another). Next, the tabulated metadata from a sample of data records in each of the repositories was exported from each system, using either the built-in repository export feature (.csv file for DSpace) or queried directly from the database backend (Fedora). The authors then compared the metadata in each record to analyze:

- number of fields completed (taking into account the required fields);
- documentation types (if any);
- digital object identifiers;
- number of keywords.

Study Limitations

This study had several limitations. First, the four institutions from this study represented a small and self-selected non-random samples. In this study, it was difficult to separate institutional factors (such as number of curation staff, difference in minting DOIs, promotional efforts, researcher education) from curation factors (procedural steps) due to the small number of institutions participating in the study. Further, the differences in the repositories themselves made them difficult to compare. For example, each repository has been available for different lengths of time with different staffing models. Each also had a unique user interface (UI) and varying degrees of resources to devote to UI development; this likely also contributed to different levels of metadata submitted by users. Finally, this study is limited to IRs and does not examine domain repositories, which are likely to contain specific metadata elements and thus yield different results.

Statistical Analysis

The authors consulted with the Center for Open Science as to appropriate statistical tests for data analysis. The current study design does not enable the authors to separate institutional impact and curation process impact completely. Given the study design, the small number of institutions involved in the study, the small number of datasets from each institution, and the limitations of the study described above, statistical analysis tools for normal distributions are not applicable. Instead, non-parametric descriptive statistics and use of the Mann-Whitney U test (Social Science Statistics, n.d.) were most appropriate.

Results and Discussion

Question #1: How do the metadata elements vary for each institution?

Submitters to a data repository make decisions about how to describe their content (e.g., which metadata fields to complete, how much detail to include in each field, etc.). Each institution represented in the study used an online submission form that guided contributors through the set of metadata options.

The authors identified how the metadata elements varied for each institution and limited comparisons to descriptive metadata. Shown in Table 2, the metadata elements are listed as either required or optional fields for each repository. Table 3 describes these elements in more detail and indicates if there are fields that are 'auto generated,' in other words, this metadata is system supplied and cannot be overridden by contributors. Non-public, internal metadata, or hidden fields are omitted. Examples of this are in Scholar@UC and Deep Blue Data, contributors can add an additional person who can edit the metadata for the work but this name will not be displayed on the record.

There was a high amount of variability between metadata elements collected in the four institutional repositories. Each institution's submission form varied in the number of total fields in the submission form, number of required fields for a submission to be submitted and which metadata element fields are required. The four institutions had only six elements in common and whether or not the element was required for submission varied (see Table 3). The six common elements are: title, creator / author, description, subject terms or keywords that describe the topic of the dataset, persistent identifiers (i.e. DOI's and PURL's) and licenses. Also, even if an element was common across institutions, the definition or usage of the element varied slightly in meaning. For example, the Related Materials field for Scholar@UC was intended only for other content within the repository whereas the same element was used by the other three repositories for citations to publications or links in external locations. Oregon State was an outlier with no required fields for submission, rather the emphasis was on reusability through data documentation (i.e. readme files). Each repository had one or more fields in the submission form that was unique to their submission form. For example, Michigan's Deep Blue Data is the only repository that required a metadata describing the Method used to collect the data.

Table 2. Repository submission form required and optional metadata elements.

	Cincinnati 20 Fields (7 required)	Michigan 10 Fields (6 required)	Minnesota 19 Fields (3 required)	Oregon State 20 Fields (0 required)
Required Fields	Title Creator(s) College Department or Program Description License Access rights	Title Creator Method Description CC License Discipline	Title Contact Contact Email	

 doi:10.2218/ijdc.v13i1.632

	Cincinnati 20 Fields (7 required)	Michigan 10 Fields (6 required)	Minnesota 19 Fields (3 required)	Oregon State 20 Fields (0 required)
Optional Fields	Publisher Required software DOI Date created Alternate title Subject Geographic subject Time period Language Citation Note External link Related Works	Date Coverage Keyword Language Citation to Related Work(s)	Author(s) Group Author Subject Keywords Abstract Description DOI Funder Information Date of Collection - start Date of Collection - end Date Completed Citation to Related Paper(s) Time Period Geographic Area/Coordinates Source Information Source Data URL License Type	Title License Authors ORCID Abstract Subject(s) or Keyword(s) Contributor(s) Date(s) Sponsorship Related materials Format of data Version Geolocation Affiliations Contact name Contact email address Username Embargo

Table 3. Detailed comparison of metadata elements for each institution.

Metadata	Dublin Core Element	Institution	Submission Form Display Name	Req?
Metadata Elements Used by One Institution				
Title of the Dataset	dc.title	Cincinnati	Title	✓
		Michigan	Title	✓
		Minnesota	Title	✓
		Oregon State	Title	
Author or Creator of the Dataset	dc.creator dc.contributor.author (MN)	Cincinnati	Creator(s)	✓
		Michigan	Creator	✓
		Minnesota	Author(s)	
		Oregon State	Lead Investigator(s) / co-author(s)	
License applied to the dataset	dc.rights dc.rights.uri	Cincinnati	License	✓
		Michigan	CC License	✓
		Minnesota	License Type	
		Oregon State	License	
Related works or publications that use or	dc.is referencedby(MI) dc.relation.isreferenced	Cincinnati	External Link (unmapped)	

Metadata	Dublin Core Element	Institution	Submission Form Display Name	Req?
Metadata Elements Used by One Institution				
cite the dataset	by (MN) dc.description (OSU)	Michigan	Citation to Related Work(s)	
		Minnesota	Citation to Related Paper(s)	
		Oregon State	Related materials	
Subject Terms or Keywords that describe the topic of the dataset	dc.subject dc.relation (MI)	Cincinnati	Subject	
		Michigan	Keyword	
		Minnesota	Subject Keywords	
		Oregon State	Subject(s) or Keyword(s)	
DOI	dc.identifier.doi RDF.doi (MI)	Cincinnati	DOI	
		Michigan	DOI (assigned outside of the submission process)	
		Minnesota	Persistent Identifier*	
		Oregon State	DOI	
Metadata Elements Used by Three Institutions				
Description of the dataset	dc.description	Cincinnati	Description	✓
		Michigan	Description	✓
		Minnesota	Description	
		Oregon State	--	
Date of Publication (ie. the date of issue from the standpoint of the IR)	dc.date.issued	Cincinnati	--	
		Michigan	Date Uploaded*	
		Minnesota	Date Published*	
		Oregon State	Date*	
Dates or time span covered by the dataset	dc.coverage.temporal dc.temporal (MI)	Cincinnati	Time period	
		Michigan	Date Coverage	
		Minnesota	Time Period	
		Oregon State	--	
Geographic location covered by the dataset	dc.coverage.spatial	Cincinnati	Geographic subject	
		Michigan	--	
		Minnesota	Geographic Area/Coordinates	
		Oregon State	Geolocation	
Metadata Elements Used by Two Institutions				
Abstract describing the dataset	dc.description.abstract	Cincinnati	--	
		Michigan	--	
		Minnesota	Abstract	

 doi:10.2218/ijdc.v13i1.632

Metadata	Dublin Core Element	Institution	Submission Form Display Name	Req?
Metadata Elements Used by One Institution				
		Oregon State	Abstract	
Contact Information	dc.contributor.contactname dc.contributor.contactemail	Cincinnati	--	
		Michigan	--	
		Minnesota	Contact	✓
			Contact Email	✓
			Contact name (unmapped)	
		Oregon State	Contact email address (unmapped)	
Discipline of the Dataset (controlled vocabulary)	dc.subject.department RDF.subject (MI)	Cincinnati	Department or Program	✓
		Michigan	Discipline	✓
		Minnesota	--	
		Oregon State	--	
Language of the dataset	dc.language	Cincinnati	Language	
		Michigan	Language	
		Minnesota	--	
		Oregon State	--	
Publisher	dc.publisher	Cincinnati	Publisher (required if DOI assigned)	
		Michigan	--	
		Minnesota	Publisher*	
		Oregon State	--	
Sponsorship or Funder of the dataset	dc.description .sponsorship	Cincinnati	--	
		Michigan	--	
		Minnesota	Funder Information	
		Oregon State	Sponsorship	
Type or Format of the dataset	dc.type	Cincinnati	--	
		Michigan	--	
		Minnesota	Type*	
		Oregon State	Format of data	
Metadata Elements Used by One Institution				
Other elements unique to each institution's submission form	Varies by institution	Cincinnati	Alternative Title	✓
			Citation	
			College	
			Date created	
			Related Work	
			Note	
			Required Software	
			Access Rights	

Metadata	Dublin Core Element	Institution	Submission Form Display Name	Req?
Metadata Elements Used by One Institution				
		Michigan	Method	✓
			Collection period - start Collection period - end	
			Dataset Type	
			Date Completed	
			Group Author	
			Source Information	
		Minnesota	Source Data URL	
			Affiliations	
			Embargo	
			ORCID	
			OSU Username	
		Oregon State	Readme	
			Version	

*Auto generated field that is not completed by the contributor

Question #2: How complete is the metadata submission for datasets in each institution repository given the type of curation?

Metadata is crucial to preserving research data provenance and for data discovery (FORCE11, 2017), and there are global initiatives such as the Research Data Alliance (RDA)[1] and International Council for Science: Committee on Data for Science and Technology (CODATA) to promote good data description standards and documentation practices (RDA, 2017; CODATA, 2017). Certain research communities have well-defined metadata standards for data, like the Sequence Read Archive and Expressed Sequence Tag Database Metadata schemas used in Genbank, a repository for Genomics Research. IRs often fill a special niche for data that does not have a discipline repository or provides a more economical solution to data preservation and therefore handle data that may not have community defined standards (Cragin et al., 2010).

Metadata completeness is defined as the required and optional completed metadata fields in the submission process for each dataset (Margaritopoulos, Margaritopoulos, Mavridis and Manitsaris, 2012). This comparison showed the impact of the model of submission and curation support on the metadata completeness for a given dataset. Margaritopoulos, Margaritopoulos, Mavridis and Manitsaris (2012) (represented as blue points in Figure 1) calculated percent completeness as follows:

$$\%\ completeness = \frac{(\#\ of\ required\ field + \#\ of\ optional\ completed)}{(\#\ of\ required + \#\ optional\ fields\ available)}$$

Additional formulas developed by the authors calculated the minimum percent completeness (represented as orange points in Figure 1) and the average percent completeness (represented as yellow points in Figure 1) for the 20 datasets as follows:

1 Research Data Alliance – RDA/WDS publishing data workflows working group: https://www.rd-alliance.org/groups/rdawds-publishing-data-workflows-wg.html

 doi:10.2218/ijdc.v13i1.632

$$Minimum\ \%\ completeness\ =\ \frac{(\#\ of\ required\ completed)}{(\#\ of\ required\ +\ \#\ optional\ fields\ available)}$$

$$Average\ \%\ completeness\ =\ \frac{(\%C1\ +\ \%C2\ +\ \%C3\ +\ldots+\ \%C20\ )}{20\ datasets\ for\ each\ institution}$$

The authors made a direct comparison for each data set regardless of discipline or type. The percent completeness for each of the 20 data sets from the four institutions are visualized in Figure 1. The percent completeness profile was unique for each institution, which was not too surprising since each dataset was unique. In all submission process types (no curation, selective curation, pre-ingest curation and post-ingest curation), researchers contributed information for more than the minimum elements required.

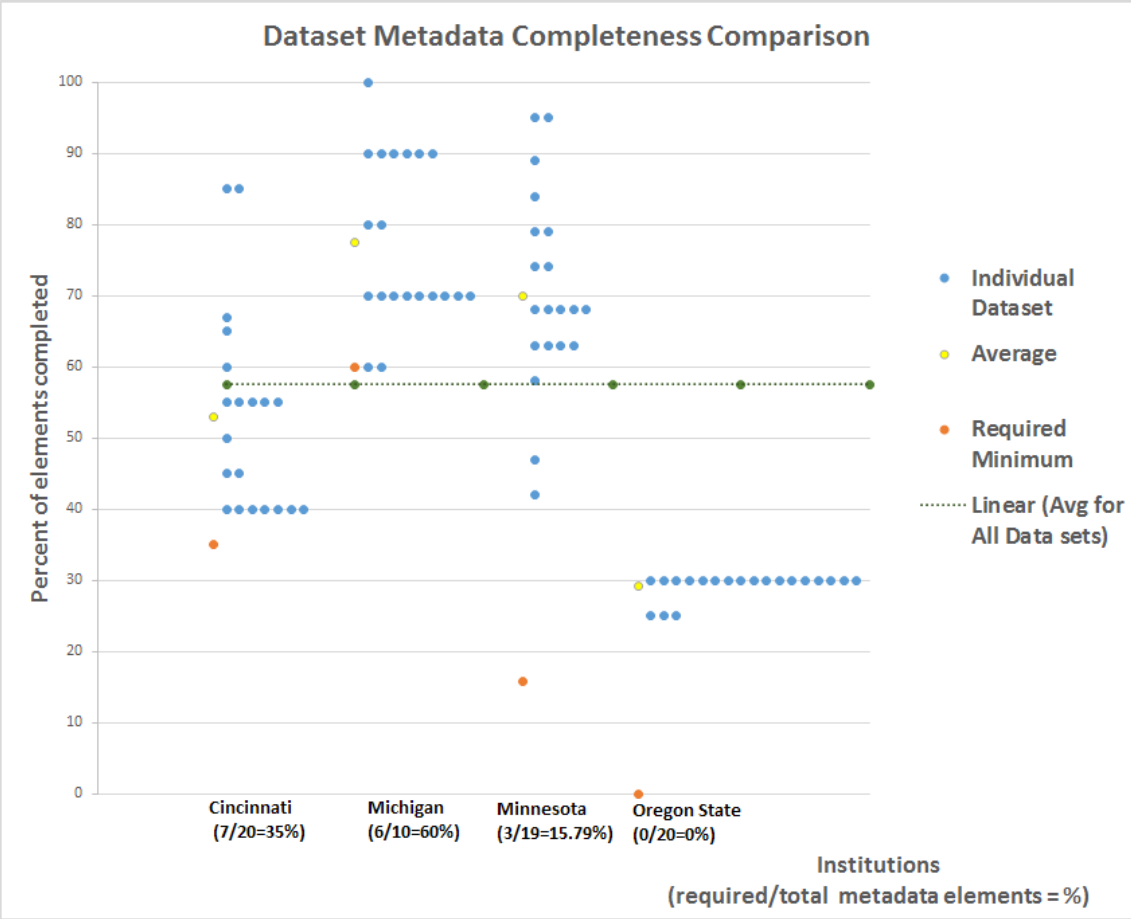

Figure 1. Comparison of metadata fields completed for 20 data sets from each repository where (X,Y) = (ordinal rank, percent completeness).

Figure 1 shows that datasets fell in a broad range of metadata completeness, well below and above average percent complete (58% for all datasets- green line on graph), for Cincinnati (40-85%), Michigan (60%-100%), and Minnesota (42%-95%). Only Oregon had consistent, but low, completeness, with all 20 datasets hovering near 25%-30% complete.

The difference between the average percent completeness and the minimum are shown in Table 4. The average percent completeness for Oregon State, Minnesota,

Michigan, and Cincinnati were 29%, 70%, 78% and 53% respectively. The minimum required for Oregon State, Minnesota, Michigan, and Cincinnati were 0%, 16%, 60% and 35% respectively (see Table 4). However, there was not a remarkable increase in optional metadata fields completed in the two models with curation support (Oregon, Minnesota) over the two repositories without consistent curation support (Cincinnati, Michigan).

Table 4. Descriptive Statistics for percent completeness of metadata fields per institution.

	Oregon (pre-ingest curation)	Minnesota (post-ingest curation)	Michigan (selective curation)	Cincinnati (no curation)
# Field required/Total # Fields	0/20	3/19	6/10	7/20
Minimum % Required	0%	16%	60%	35%
Minimum % Completed	25%	42%	60%	40%
Average % Completed	29%	70%	78%	53%
Median % Completed	30%	68%	70%	53%
Maximum % Completed	30%	95%	100%	85%
Range % Actual Completed (Max% - Min %)	5%	53%	40%	45%
Avg Percent of Metadata Completed Above Minimum (Avg% - Min%)	29%	54%	18%	18%
Kurtosis *x < ± 2	2.78	0.11	-1.11	0.85
Skewness *x < ± 0.5	-2.12	0.02	0.32	1.16

* indicates value range for normal distribution for comparison to results.

If all data sets should have at least their required fields completed (e.g., minimum completeness), then the fact that the average percent completeness are higher in all four cases demonstrates some effort, by users or curators, to give data greater context. There are several possible reasons for higher metadata percent completeness than required: users could be compelled to describe their data for greater discoverability, the user interface of the repository may lend itself to creating more complete records, the curator may be adding additional context on behalf of the user, or the number of required fields is simply too low for this complex type of work (e.g., data sets), or metadata fields may apply to some datasets and not others.

 doi:10.2218/ijdc.v13i1.632

However, these findings are inconclusive to directly link curation with metadata completeness. On the one hand, Minnesota (which employed post-ingest curation for all datasets) saw the greatest increase (54%) from the percent completeness of required fields (16%) to the average percent completeness (70%), which could be attributed, at least in part, to curation. On the other hand, Michigan (which did not routinely curate author-submitted metadata for the datasets) had the overall highest average percent completeness of 78% benefiting from its requiring 60% of its metadata fields and by having fewer metadata fields available. Finally, Oregon, which used a pre-ingest curation method and has no required fields, did not show a comparatively higher degree of completeness among optional metadata fields. Therefore, it is not possible to conclude that curator intervention will result in more completion of metadata beyond the minimum required fields.

Skewness is the measure of the asymmetry of a probability distribution and kurtosis describes the shape of a probability distribution or its 'tailedness'. Skewness and kurtosis results indicate that the populations are non-normal in distribution. Skewness and kurtosis values were generated using the data analysis add-in for Excel 2013. Ideal results would be $x < \pm 0.5$ and $x < \pm 2$ for skewness and kurtosis respectively (See Table 4). The Mann-Whitney U test is designed for non-normal distribution populations and samples with small size ($n<20$). Analysis was done on the completeness profiles using a web-based statistics calculator (Social Science Statistics, n.d.). Criteria for the test were set for a two-tailed analysis at a p value/significance level of 0.05. Significance would have a U critical value less than 127 for $n=20$. Results of the Mann-Whitney U test indicated that there is significant difference in the numerical ranking of the completeness profiles in pairwise comparisons (i.e. UC to UM, UC to UMN, etc.)

Calculations for the analysis can be found in the reference dataset collection (Koshoffer et al., 2018).

Question #3: Are curated datasets more likely to have documentation associated with the work?

Each of the four participating institutions reported the number of documentation files associated with each dataset in the sample, as is shown in Figure 2. Documentation are necessary to ensure that datasets can be found and used in the future (Rolando, 2015). The authors hoped to understand if the type of curation had an impact on whether or not datasets included documentation and what types.

The sample showed that documentation is far less common in the repositories with selective or no curation. Minnesota reported documentation for every submission and the sample from Oregon State included documentation for 15 of the 20 submissions. Michigan and Cincinnati, on the other hand, showed very low numbers of documentation files associated with the data sets in their samples. Users are much less likely to submit documentation files unless they are required, either upon deposit or as part of the curation process.

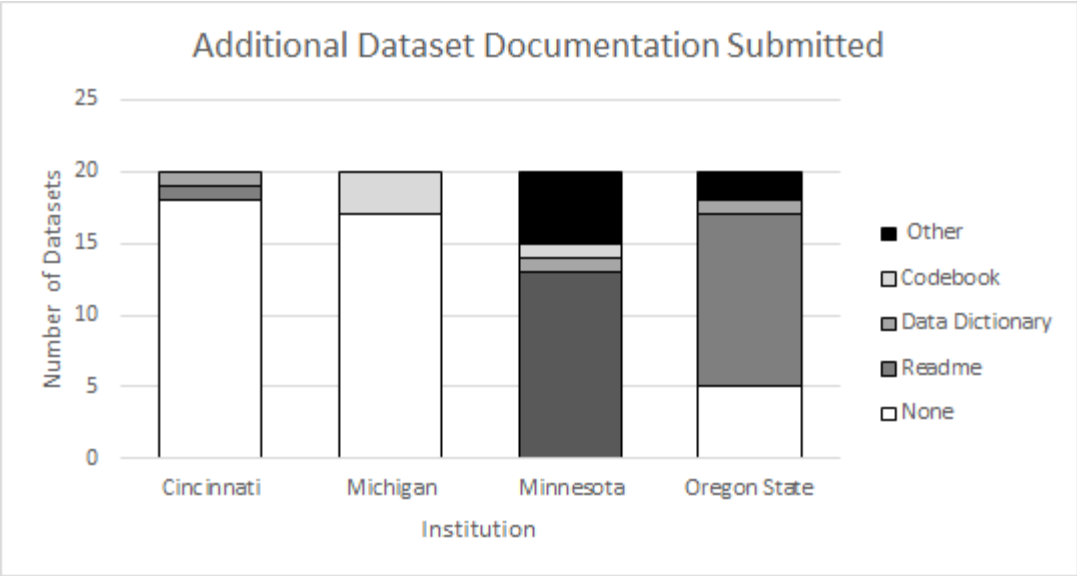

Figure 2. Datasets with documentation provided (by type).

Table 5. Datasets with documentation provided (by type).

	Cincinnati	Michigan	Minnesota	Oregon
None	18	17	0	5
Readme	1	0	13	12
Data Dictionary	1	0	1	1
Codebook	0	3	1	0
Other	0	0	5	2
Total	20	20	20	20

Rich metadata and documentation, such as protocols, data dictionaries, and readme files provide necessary context to research data (Peer, 2013). The majority of users included readme files, followed by other types, codebook, and data dictionary. Users submitted documentation types classified as other, including interview protocols, project summaries, schematics, and collection protocol. Twelve data sets had more than one documentation type, for example a schematic and a data dictionary.

Question #4: Do the number of data sets with DOIs vary in each repository?

Each institution supported Digital Object Identifiers (DOIs) and reported the number of dataset associated DOIs in their sample. 100% of the datasets from the two institutions with curation have DOIs. The fact that 90% of Michigan's datasets have DOIs may suggest that other factors (e.g. promotion) may also contribute.

 doi:10.2218/ijdc.v13i1.632

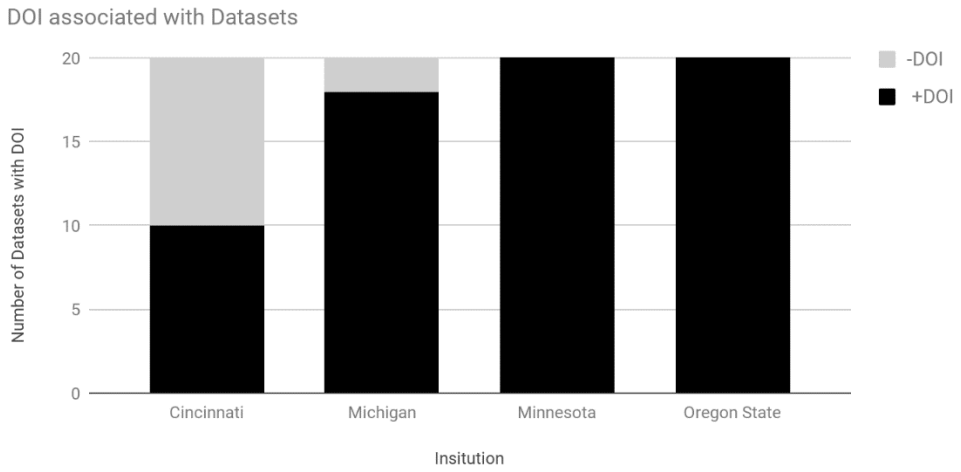

Figure 3. Comparison of number of data sets with a digital object identifier.

Table 6. Comparison of number of data sets with a digital object identifier.

	Cincinnati	Michigan	Minnesota	Oregon State
# of DOIs	10/20	18/20	20/20	20/20
Created automatically	No	No	Yes, manually after curation	Yes

Question #5: What is the difference in number of keywords associated with each dataset?

The authors examined the number of keywords that researchers submitted to describe their datasets. None of the institutions required keywords and none of them used a controlled vocabulary list, i.e. Library of Congress Subject Headings or Medical Subject Heading terms. The majority of the datasets had at least five keywords added per dataset in three of the four institutions. The overall average number of keywords was 4.35. Cincinnati was the outlier as most datasets had no keywords. A possible reason is that the Scholar@UC submission form did not display the option to add keywords on the first page of submission form. Instead the contributor needed to click on a link titled 'Add Additional Description' to open a second page of the submission form in order to add keywords.

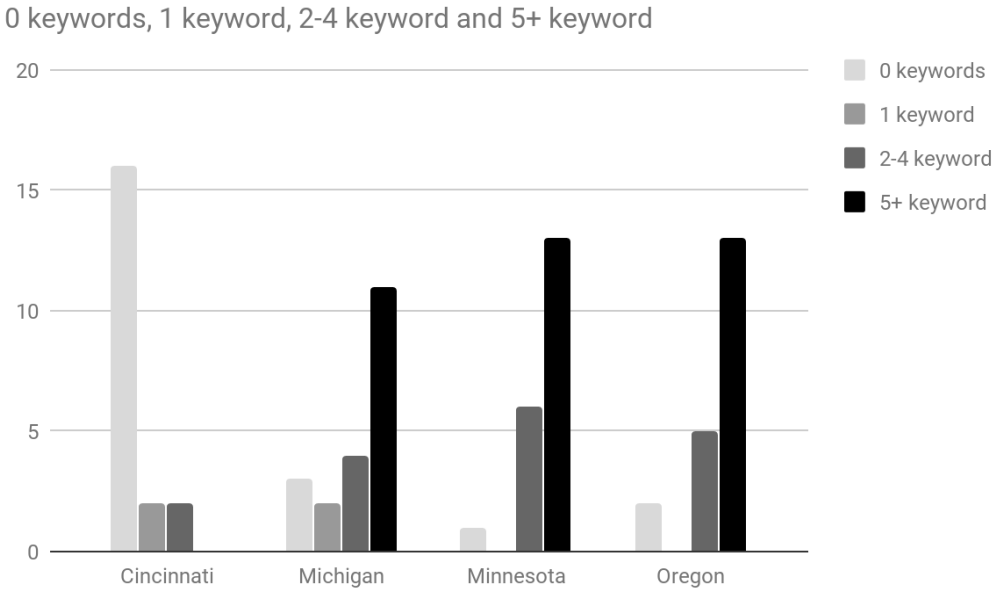

Figure 4. Number of keywords per data set.

Table 7. Number of keywords per data set.

	0 keywords	1 keyword	2-4 keyword	5+ keyword
Cincinnati	16	2	2	0
Michigan	3	2	4	11
Minnesota	1	0	6	13
Oregon State	2	0	5	13

Conclusion

The genesis of this project started with the idea to compare how the curation process contributed to the growth of datasets in an IR. The project quickly evolved into an examination of the metadata submission quality based on the type of curation process. The authors compared four institutions with curation processes that vary greatly and ranged from no-curation to submission acceptance based on post-ingest curation, to see how user-contributed metadata varied and what type of documentation resulted for each submission process. In the sample, the curation process may have had a measurable impact on the metadata captured and did result in more documentation, especially the inclusion of readme files, with a dataset submission.

Based on a review of the literature and the current research study, the authors recommend the following to the data repository community:

1. Institutional factors matter. When comparing samples across differing institutions it is important to keep in mind what factors make data repositories and their related services unique. An example is the number of staff; curation

practices will vary between a staff with a solo data librarian and a larger or more dedicated staff. Other factors may include promotion and training efforts around the repository. These factors are likely related to user behavior around depositing metadata, and should therefore be taken into account when designing data repository services.

2. Metadata schema should be standardized to promote interoperability between IRs. The authors did not anticipate the level of difficulty they encountered when trying to compare their metadata schema. Park (2009) underscores this recommendation by suggesting a common data model that could be interoperable across digital repositories.

3. The community should evaluate the differences between schemas and develop a minimum requirement for metadata for datasets in IRs.

4. Curation practices are important to consider. The purpose of the study was to compare differing curation practices to better understand the impact of curation on user-submitted metadata. Understanding the impact that curation has on metadata quality will allow institutions to make better informed decisions about how to spend their limited resources.

Each institution in the study strives for a robust curation workflow. IRs can advocate for datasets to be discoverable and reusable and take curation steps to improve submission metadata and documentation above the levels provided by contributors. Indeed this is happening at each institution. Since the study concluded, Michigan implemented a post-deposit curation model similar to Minnesota's program and added several additional metadata elements, including funding agency name and grant number. Cincinnati is evaluating possible new staff positions with some dataset curation tasks (i.e. confirm addition of readme files) in the job responsibilities for these posts as well as implementing outreach and educational programs on long term data preservation that include data curation best practices. Oregon State's IR, ScholarArchive@OSU migrated to the new front end user interface Hyrax 2 in November 2017. Their new platform provides an easier and clearer user interface which helps contributors contribute metadata. It will be interesting to revisit datasets collected by these institutions in the future to see how they compare in light of such positive changes.

Acknowledgements

The authors thank Steve Van Tuyl, Digital Repository Librarian at Oregon State University, for his contributions to the dataset underlying this article and Courtney Soderberg of the Center for Open Science for discussions on statistical analysis.

Data

The data underlying this study can be accessed through the UC institutional repository Scholar@UC (Koshoffer et al., 2018).

References

Briney, K., Goben, A. & Zilinski, L. (2017). *Institutional, funder, and journal data policies.* In L.R. Johnston (Ed.), Curating research data V.1. (pp. 61). Chicago: Association of College and Research Libraries. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf

Committee on Data of the International Council for Science. (n.d.). *Coordinating data standards amongst scientific unions.* Retrieved from http://www.codata.org/task-groups/coordinating-data-standards

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions, Series A, Mathematical, Physical, and Engineering Sciences, 368*(1926), 4023-4038.

Digital Curation Centre. (n.d.). *What is digital curation.* Retrieved from http://www.dcc.ac.uk/digital-curation/what-digital-curation

FORCE11. (2016). The FAIR Principle. Retrieved 10/22, 2017, from https://www.force11.org/group/fairgroup/fairprinciples

Gavrilis, D., Makri, D., Papachristopoulos, L., Angelis, S., Kravvaritis, K., Papatheodorou, C., et al. (2015). Measuring quality in metadata repositories. *International Conference on Theory and Practice of Digital Libraries*, pp. 56-67.

Heidorn, P.B. (2011). The emerging role of libraries in data curation and e-science. *Journal of Library Administration, 51*(7-8), 662-672.

Holdren, J.P. (2013). Increasing access to the results of federally funded scientific research. Office of Science and Technology Policy, Executive Office of the President.

Hudson-Vitale, C. & Association of Research Libraries. (2017). *Data curation.* Washington, D.C.: Association of Research Libraries.

Johnston, L.R. (2017). *Curating research data volume two: A handbook of current practice.* Chicago: Association of College and Research Libraries, a division of the American Library Association. Retrieved from http://hdl.handle.net/11299/185335

Johnston, L.R., Carlson J., Hudson-Vitale C., Imker H., Kozlowski W., Olendorf R., & Stewart, C. (2017). *Data curation network: A cross-institutional staffing model for curating research data.* University of Minnesota Digital Conservancy. Retrieved from http://hdl.handle.net/11299/188654

Jones, S. (2007). A report on the range of policies required for and related to digital curation – Version 1.2. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.pdf

Koshoffer, A., Hansen, C., & Newman, L. (2017). *Challenges with quality of data set metadata in a self-submission repository model*. In L. R. Johnston (Ed.), Curating research data V.2. (pp. 32). Chicago: Association of College and Research Libraries, a division of the American Library Association. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988633_crd_v2_OA.pdf

Koshoffer, A., Neeser, A., Johnston, L.R., & Newman, L. (2018). Metadata_Repositories_IDCCsubmission [Data set]. University of Cincinnati, Cincinnati, Ohio, USA: Scholar@UC. Retrieved from http://scholar.uc.edu/collections/9w0323021

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

Mannheimer, S., Sterman, L., Borda, S., & Montana State University-Bozeman. (2016). Discovery and reuse of open datasets: An exploratory study. *Journal of eScience Librarianship, 5*(1), e1091. doi:10.7191/jeslib.2016.1091

Margaritopoulos, M., Margaritopoulos, T., Mavridis, I., & Manitsaris, A. (2012). Quantifying and measuring metadata completeness. *Journal of the Association for Information Science and Technology, 63*(4), 724-737. doi:10.1002/asi.21706

Park, J. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging and Classification Quarterly, 47*(3-4), 213-228. doi:10.1080/01639370902737240

Park, J. & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging and Classification Quarterly, 48*(8), 696-715. doi:10.1080/01639374.2010.508711

Peer, L. (2013). The repository as data (re) user: Hand curating for replication. Message posted to http://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research

Rolando, L. (2015). Beyond metadata: Leveraging the 'README' to support disciplinary documentation needs Georgia Institute of Technology. Retrieved from http://hdl.handle.net/1853/53322

Rousidis, D., Garoufallou, E., Balatsoukas, P., & Sicilia, M.A. (2015). Evaluation of metadata in research data repositories: The case of the DC.subject element. In: Garoufallou E., Hartley R., Gaitanou P. (eds) Metadata and Semantics Research. MTSR 2015. *Communications in Computer and Information Science, 544*. doi:10.1007/978-3-319-24129-6_18

Social Science Statistics. (n.d.). Mann Whitney U Test. Retrieved from http://www.socscistatistics.com/tests/mannwhitney/

Vasilevsky, N.A., Minnier, J., Haendel, M.A., Champieux, R.E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ 5*:e3208. doi:10.7717/peerj.3208

Walters, T.O. (2009). Data curation program development in US universities: The Georgia Institute of Technology example. *International Journal of Digital Curation, 4*(3), 83-92. doi:10.2218/ijdc.v4i3.116

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. doi:10.1038/sdata.2016.18