Challenges with Quality of Data Set Metadata in a Self-Submission Repository Model

*Amy Koshoffer, Carolyn Hansen, and Linda Newman**

This use case will examine the challenges in acquiring quality metadata for the submission type *Data Set* in a self-submission institutional repository model and how the University of Cincinnati Libraries is responding to these challenges with modifications to the submission process and user education. Here we describe an effort to reduce barriers to data set submission by improving the repository's submission interface so that required and optional metadata fields provide adequate descriptive metadata for data set discoverability and reuse.

INTRODUCTION

As journals and funding agencies increasingly require authors to make the data behind their publications available through archiving and sharing, the creation of high-quality metadata will be essential in order to maximize the discoverability and reuse of archived data sets for future researchers. Data sets present unique challenges for description, especially within the context of a general purpose institutional repository that archives diverse work types. Data sets can be structurally complex because data sets may vary in file type or number of files. For example data can be in a single table formatted as a spreadsheet or can be many files functioning together as a unit. A text or photograph may be self-explanatory, but the meaning of raw data, as well as the relationships between files in a data set, is more difficult to decipher without adequate descriptive metadata provided by the repository submitter. A self-submission repository model places the responsibility on the submitter to provide structured and unstructured descriptive metadata as well as contextual information that make the data set discoverable in the repository and comprehensible by other researchers.

Structured descriptive metadata for a data set can be generated through a submission form that guides the submitter through the process of metadata gen-

* This study is licensed under a Creative Commons Attribution-ShareAlike 4.0 License, CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0/).

eration. Data set metadata, such as readme files containing protocols, data dictionaries, and explanations of file relationships, are equally important. However, the lack of uniform structure or standards for these documents may make the metadata difficult to collect through a structured mechanism such as a submission form. Researchers could formulate standard formats for machine-readable research reports, but individual researcher practices and local research group policies may generate variability in these documents.

THE UNIVERSITY OF CINCINNATI SELF-SUBMISSION REPOSITORY MODEL

The University of Cincinnati (UC) Libraries and University of Cincinnati Information Technology unit (IT@UC) collaborate on the continued development of UC's institutional repository, Scholar@UC (https://scholar.uc.edu/). The repository application is built on the Project Hydra framework (https://projecthydra.org/), which utilizes Fedora Commons (http://www.fedora-commons.org/) as the underlying repository. Project Hydra is a multi-institution collaboration to develop open-source repository solutions resulting in an "ecosystem of components" that allows partner institutions to customize their own repository solutions. Scholar@UC is a self-submission digital repository with the goal of preserving the scholarly output of UC researchers, such as preprints of articles, conference posters, images, and data sets. Scholar@UC went into production in September 2015, but software development and engagement with early adopters continues.

UC RESEARCH COMMUNITY ENGAGEMENT

To set policies and guide repository implementation, UC Libraries and IT@UC formed the Digital Repository Task Force (DRTF). In order to engage the greater UC research community, the DRTF formed the Early Adopter Working Group, composed of library faculty and staff who served as subject liaisons or whose responsibilities involved research data support, to recruit researchers as early adopters. Early adopters were research and teaching faculty who agreed to submit content (including data sets), provide feedback, and suggest additional functionality before the repository was open to all potential submitters campus-wide. The goal of early adopter sessions was to test more than just usability; it was to discover what researchers wanted and needed from a repository. Members of the Early Adopter Working Group selected early adopters based on discipline, types of digital content, and availability to collaborate with the working group.

EARLY ADOPTER FEEDBACK SESSIONS

As part of the pre-rollout development process, Early Adopter Working Group members engaged these early adopting research and teaching faculty in face-to-face sessions to obtain feedback on the usability of the system and to communicate what additional functionality was needed to describe and archive their scholarly output. Members of the Early Adopter Working Group documented the early adopters' comments as they contributed content such as data sets to Scholar@UC. When submitting content, the repository interface prompted early adopters to select a work type, upload their content, contribute metadata, select a license, request a DOI (digital object identifier), and set the level of access for their content. After submission, working group members asked early adopters to search for their content and browsed other contributed content. Scholar@UC currently supports the following work types: *Article*, *Image*, *Dataset*, *Document, Video* and *Generic Work*.

The first decision a submitter makes is which work type to select. We learned during these feedback sessions that this decision was not always intuitive, and that our assumptions about what constitutes a data set were challenged. Our work types are based on the function of a resource's content, as opposed to its physical or digital format. For example, the *Image* work type is designed for use with visual content such as art and has different descriptive focus than the *Dataset* work type, for which the content is viewed as research data. Our helper text on the work type selection page for *Image* is "Visual content: art, photographs, posters, graphics" and for *Dataset* is "Files containing collections of data, including: raw data, spreadsheets, logs, etc." In this case, a researcher has a choice to submit scanned index cards representing data as either an *Image* or as a *Dataset*. The submission process is designed to respect the researcher's decision, and example data sets where the data is represented within images exist in our repository.

By selecting a work type, the submitter prompts the system to populate the submission form with premapped metadata elements for that particular work type (appendix 1.0 D). For data sets, this included metadata fields to identify software that may be needed to view or open the data set as well as specifics about the format of the data set. Provided file format advice encouraged use of open over proprietary formats, and per our Terms of Use (TOU) agreement, library staff could create open formats at a later date. In the first version of Scholar@UC tested by early adopters, if a submitter chose the *Dataset* work type, every possible premapped metadata field for the *Dataset* work type would appear in the submission form, regardless of whether the field's input was required or optional. As a result, the submission form contained up to twenty-five metadata fields, which submitters found excessive and confusing. Submitters also found it difficult to tell if a field was required or optional; then

failure to complete required fields prevented the submitter from finishing the submission.

METADATA CHALLENGES

Librarians and developers attending feedback sessions documented early adopters' comments and identified many metadata issues. Early adopters did not routinely supply extensive metadata when submitting a data set to the repository. In some cases, these researchers thought that sufficient descriptive information was contained within basic, required repository metadata fields or summary findings published in a journal unaccompanied by supporting data.[20] Many early adopters did not consider how the quality of metadata impacted data set discoverability and reusability while others found the input forms unclear, thereby preventing the submission of quality metadata.[21]

Early adopters commented that some metadata elements were unclear to them or that they were unsure what information was needed. For example, some early adopters interpreted the metadata field *Title* as an individual's academic title, not the work's title (early adopter comment in feedback session). Other early adopters entered metadata that was irrelevant or useless such as "p.txt" for a data set title, while others left optional fields blank because they did not want to complete so many fields. We were not clear if this misunderstanding suggests that data sets are usually only given descriptions rather than specific titles. Some early adopters also had multiple files or related works and desired the functionality to reuse metadata from previous submissions to save time.

We used these feedback sessions to educate early adopters about data curation best practices and creating more complete metadata to improve the discoverability of their specific data set. However, we also learned what aspects of selection and metadata were unclear to our submitters and needed significant improvement in order to sustain the vision of a repository designed for faculty submitting data sets without librarian mediation.

USE CASES

Based on early adopter comments, library faculty developed use cases to describe the metadata changes that research and teaching faculty desired. Use cases were recorded and are available at https://github.com/uclibs/scholar_use_cases/blob/master/submission/submission_use_cases.md. Additionally, the source code for the UC repository can be accessed at https://github.com/uclibs/scholar_uc, and significant changes to the interface are summarized in a change log at https://github.com/uclibs/scholar_uc/blob/develop/CHANGELOG.md.

THE NEW SUBMISSION FORM

The new submission process presents submitters with a streamlined form (for a comparison see original form in figure 1.6). Researchers upload their data set first and then are prompted to complete only the core or required metadata elements (see figure 1.7). Submitters have the ability to show additional optional elements by clicking the Show Additional Description link (see figure 1.8). Metadata element descriptions and examples were added to the submission form to clarify the desired input. For example, to clarify the metadata desired for the field *Title,* the example text says "Enter the title of your dataset. If the Dataset doesn't have a title, please enter a brief descriptive label."

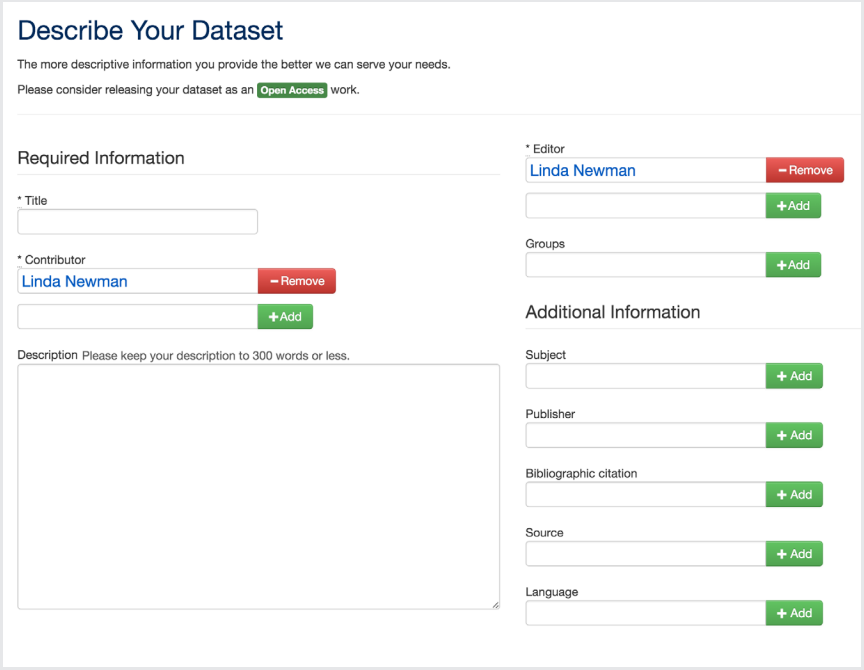

FIGURE 1.6
Our starting point with a "vanilla/generic" version of software inherited from other Hydra participants. Before we redesigned the form as shown in figures 1.7 and 1.8, we added fields for Required Software, Alternate Title, Geographic Subject, Time Period, Date Created, Note, and others, making this a very long form without any help information for metadata, and with permissions fields such as Editor Assignment intermixed with description.

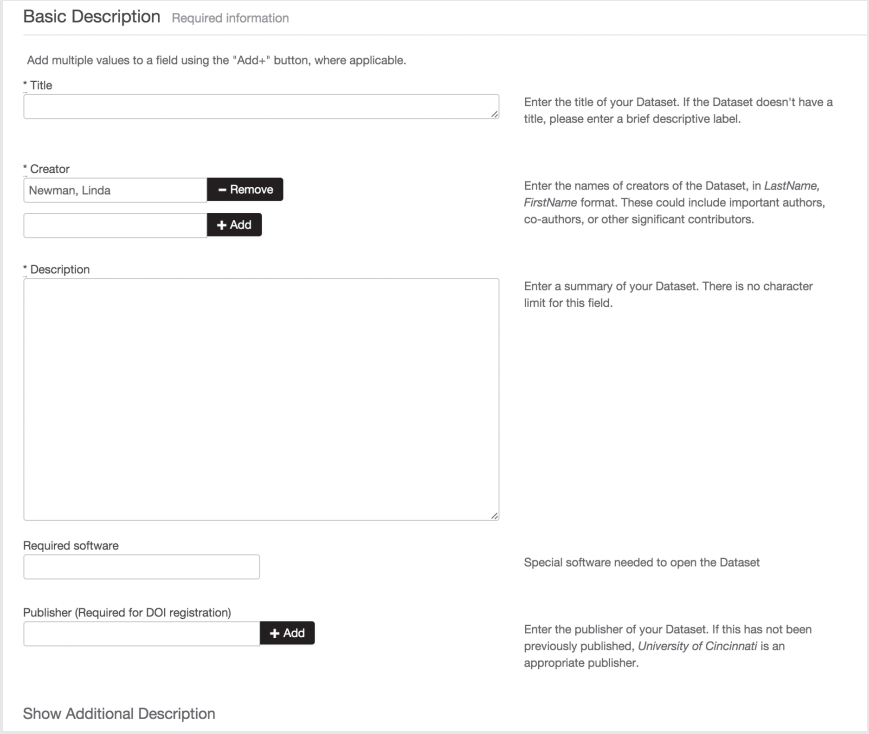

FIGURE 1.7
Screenshot showing only the required metadata that must be completed to submit a data set to the repository, with brief help information next to each field on the right. Fields for file upload, Creative Commons licenses, access rights, DOIs, permissions such as editor and group assignment, and the depositor agreement have been moved to be distinct from the Basic Description.

The early adopters gave positive feedback about the new streamlined form. Now that we have achieved our initial rollout ("Scholar@UC 1.0"), developers are continuing to add functionality designed to encourage submitters to input sufficient metadata. This includes recommending the inclusion of unstructured metadata (attaching documentation such as readme files and data dictionaries). We are also investigating going beyond currently offered metrics from Google Analytics to provide submitters with metrics about accesses and downloads of their content and to include altmetrics-style data about citations. Our untested theory is that if submitters see evidence that metadata quality impacts their work's discoverability through social media outlets such as Twitter or projects like ORCID (http://orcid.org), through search engines such as Google, and avenues other than publisher-indexed databases, submitters would provide more descriptive metadata

during submission. Future enhancements, such as a submission dashboard, will enable uploading groups of records with a common template for shared metadata. These ideas are based on suggestions by the use cases that focused more on submission than discovery, which is helpful at this stage in the repository development.

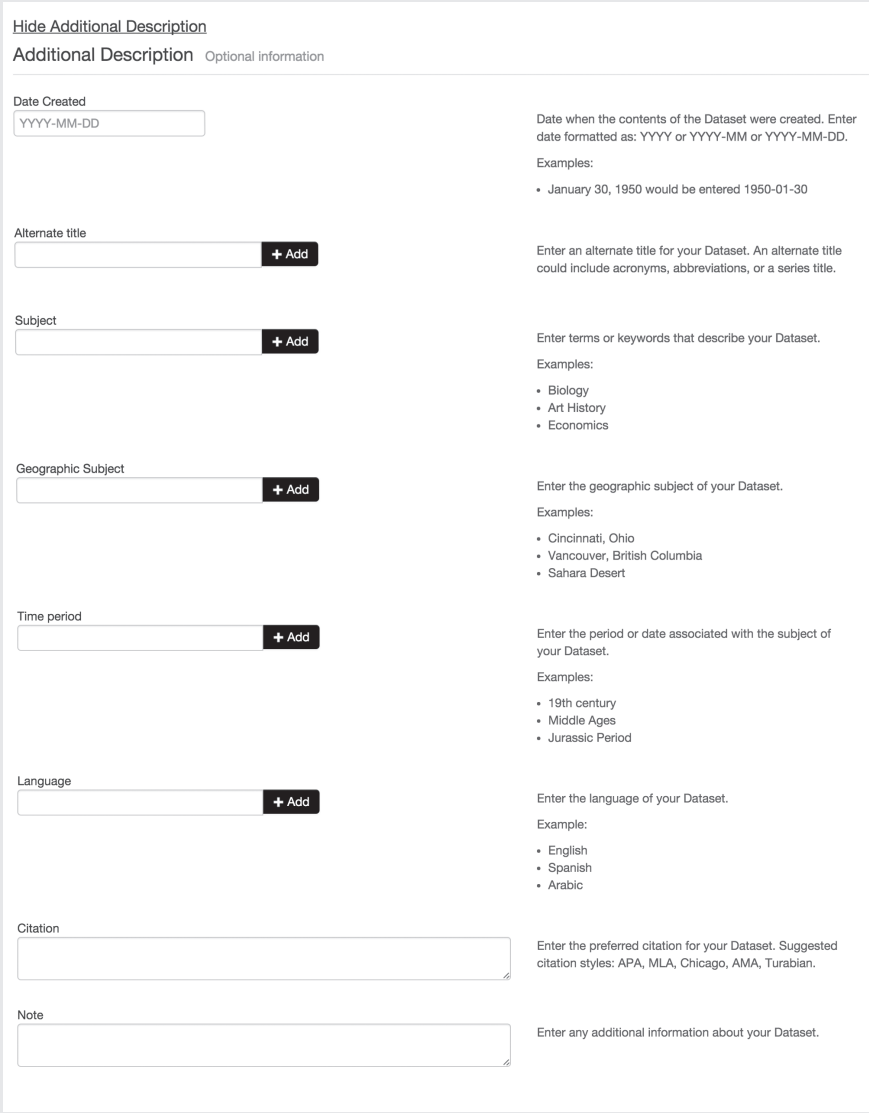

FIGURE 1.8
Screenshot showing Additional, Optional Metadata, with brief help information on the right.

NEXT STEPS

Primary responsibility for inputting complete and accurate metadata resides with researchers now that the repository has been released to the UC community. What steps can the library take to ensure that submitters provide complete metadata? And for data sets in particular, how do we convey the importance of including documentation about research protocols and data dictionaries? How can the library help researchers understand the value of this description and documentation? UC Libraries is working hard to find a balance between what metadata a repository submitter will incorporate into their records and what support the library can feasibly provide. For example, the nonexclusive distribution license that users must agree to when submitting content to Scholar@UC states that UC Libraries reserves the right to preserve, transform, or enhance metadata. However, the goal is to respect the autonomy of the content submitter and the metadata created.

UC Libraries faculty and staff will continue to educate repository users on the value of metadata, especially as it is related to increasing citations to their data.[22] This involves educating researchers about the process of data curation and archiving data in the repository and the impact of metadata on data dissemination and discovery. Education can take the form of workshops on data curation and preservation, particularly using Scholar@UC, and providing best practices for both structured descriptive metadata and unstructured metadata for reuse such as a readme file with file relationship explanation, file-naming conventions, protocols, variable explanations, and researcher contact information. Since researchers may not submit data documentation in structured format, encouraging the use of templates for protocols and data formats could also aid in discovery and reuse. In the data management workshops we offer, we teach the principle of file-naming conventions. We suggest researchers implement a convention that explains relationship between associated files where titles would be assigned to projects and experiments and by extension to the data set that would be generated by the projects and experiments. Education opportunities can include stand-alone educational tools, one-on-one consultations, and partnering with research group principal investigators. Consultations during our early adopter sessions suggest that we can persuade researchers that quality metadata is the conduit to discovery of their content by demonstrating how a data set submitted to the repository with extensive and complete metadata could result in increased data citations and data reuse.[23] We also learned from these sessions that we would get poor or no metadata without a simpler and self-explanatory input interface. By both streamlining the submission process and educating our researchers on the need for complete and quality metadata, we will provide a better service that is easy to use and benefits researchers in the long run.

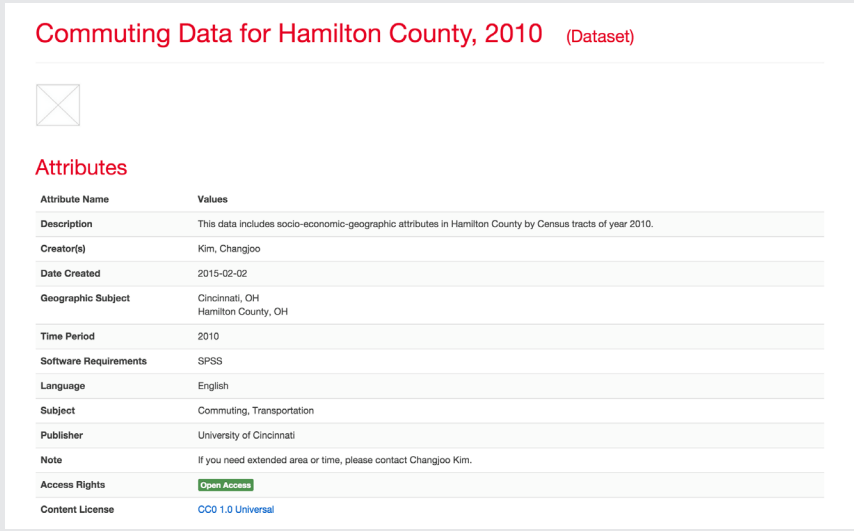

FIGURE 1.9
Example of metadata generated by system for submitter content.

1.6 Receive Notification of Data Arrival

Repository staff should get an alert once the data has arrived at the repository. This can be in the form of an alert or notification of new objects in the submission staging area. Note that if one staff member is in charge of receiving these notifications that there are mechanisms in place to forward alerts due to absence or other busy times. For example, use a general e-mail account that several staff members are able to log into or receive forwarded e-mails to maintain consistent service during busy or understaffed periods.

A notification should be also given to the data author that their data was received (see figure 1.10). This notification may happen automatically by the system. The submission tool may also indicate what stage the data submission is in to keep the author up-to-date on the data curation progress. The systems used by publishers for article submissions (ie. bePress software's editorial submission workflow) are a good example for monitoring how your submission is moving through the process: received and pending review, under review, accepted pending recommended changes, author changes, finalized for publication, and so on.

Appendix 1.0 D: Descriptive Metadata for a Data Set Using Resource Description Framework (RDF) from "Challenges with Quality of Data Set Metadata in a Self-Submission Repository Model"

Amy Koshoffer, Carolyn Hansen, and Linda Newman

The model and standards at UC allow us to assign namespaces from different metadata standards for different types of work. For example, our *Image* work type includes some metadata elements with a Dublin Core namespace and some with a Visual Resource Association (VRA) namespace.[24] For the *Dataset* work type, however, all elements are from Dublin Core.

The Friend of a friend (FOAF) namespace is used for information about a submitter that is stored in their user profile.[25] In Scholar@UC, the only required metadata fields on the *Dataset* work type input form are Title, Creator, Description and Rights (Creative Commons Content License, [http://creativecommons.org] on the submission form. The Content License defaults to "All rights reserved.") The system will supply the name of the submitter as the Creator, but the submitter can remove their name and insert another creator name if applicable. In addition, the system supplies two date fields, dateSubmitted and Modified. When the submitter selects the Dataset input form, the system supplies the type field value as Dataset. The list below shows all metadata field content that could be generated in the submission process for an example data set.

The example below is descriptive metadata based on the record available at https://scholar.uc.edu/works/datasets/dr26xx804 and shown in figure 1.9.

```
    <info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/type>
"Dataset" .

    <info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/date-
Submitted>    "2015-08-05Z"^^<http://www.w3.org/2001/XMLSchema#-
date> .

    <info:fedora/sufia:dr26xx804>   <http://purl.org/dc/terms/ti-
tle> "Commuting Data for Hamilton County, 2010" .

    <info:fedora/sufia:dr26xx804>  <http://purl.org/dc/terms/cre-
ator> " Kim, Changjoo" .
```

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/description> "This data includes socio-economic-geographic attributes in Hamilton County by Census tracts of year 2010." .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/identifier> "doi:10.7945/C29G68" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/requires> "SPSS" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/publisher> "University of Cincinnati" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/date#created> "2015-02-02" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/title#alternate> "Commuting Data" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/subject> "Commuting" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/subject> "Transportation" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/coverage#spatial> "Cincinnati, OH" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/coverage#spatial> "Hamilton County, OH" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/coverage#temporal> "2010" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/language> "English" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/bibliographicCitation> "Kim, C. (n.d.). Commuting Data for Hamilton County, 2010 // Dataset [dr26xx804] // Scholar@UC. http://doi.org/10.7945/C29G68".

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/description#note> "If you need extended area or time, please contact Changjoo Kim." .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/rights> "CCO 1.0 Universal" .

<info:fedora/sufia:dr26xx804> <http://purl.org/dc/terms/modified> "2015-08-05Z"^^<http://www.w3.org/2001/XMLSchema#date> .