A Destination for DRUM Dataset Deposits: Creating the UMD Data Collection

Allison Buser // UMD Libraries Research & Innovative Practice Forum // June 8, 2022

UNIVERSITY LIBRARIES

Introduction

As publisher policies and funding agencies increasingly require or encourage research data be made open for wider access and review, improving data collection and curation practices in institutional repositories has become commensurately necessary to support the needs of researchers and the goals of open scholarship. Since the Digital Repository at the University of Maryland's (DRUM) launch in 2005, it has been utilized in archiving research datasets produced by UMD researchers. However, as noted by Durden & Buser in the 2021 LRIPF poster, "Uncovering Hidden Datasets in DRUM," the repository's general self-submission workflow lacks mechanisms to consistently collect essential identifying metadata as well as other metadata necessary for best practices in research data archiving. The UMD Data Collection was created in the summer of 2021 to address such issues. This poster outlines the design and implementation of the collection and its customized workflow to better enable future curation, management, and discovery of research datasets archived in DRUM.

Collecting Data Metadata

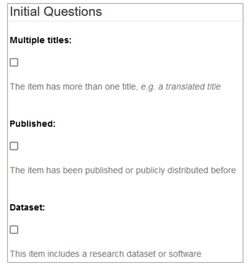
Workflow Design

In their 2016 article, Wilkinson et al. emphasize the necessity of improving infrastructure for increased distribution and reuse of scholarly data. To guide these efforts, they propose the FAIR Data Principles which state that all research data should be findable, accessible, interoperable, and reusable. In addition to considering such general best practices in data archiving, the UMD Data Collection project's following customizations to DRUM's self-submission workflow were informed by investigations of:

- The functionality and metadata collection of dedicated data repositories (ex. Dataverse)
- Comparable institutional repositories using DSpace to archive data
- Existing dataset deposits in DRUM

Cross-mapping Collections

Direct deposit to the UMD Data Collection is restricted for all users. Rather, researchers s select their associated department or program research collection to begin their submission. The Data Collection workflow is then prompted by a Dataset checkbox on the general workflow's "Initial Questions" page.
Checking the Dataset box subsequently cross-maps the submission to both the research collection and the UMD Data Collection. Unifying datasets in one collection facilitates comprehensive examination and curation of data in DRUM while additionally enabling greater discovery and impact of this specific research product archived on the platform.

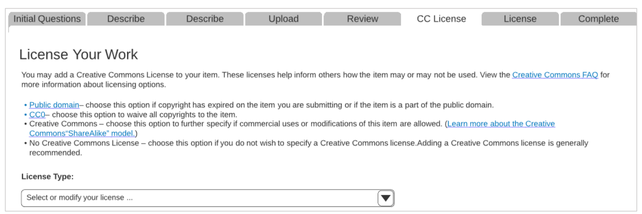


Adding Dc.description.methods

While researchers might include data collection methodology in their supporting readme documentation, this information was not being specifically captured in the DublinCore metadata. Adding an open field for methodology description in the workflow prompts submitters to include this essential descriptive metadata in the record which thus enables users to search and access contextual data collection information at the record level.

Adding Dc.rights

Prompting researchers to license their work and including that license in the record clarifies the extent of how research data may be shared and reused. The option to add various Creative Commons licenses was already utilized in the MD-SOAR instance of DSpace which is emulated in the Data Collection workflow.



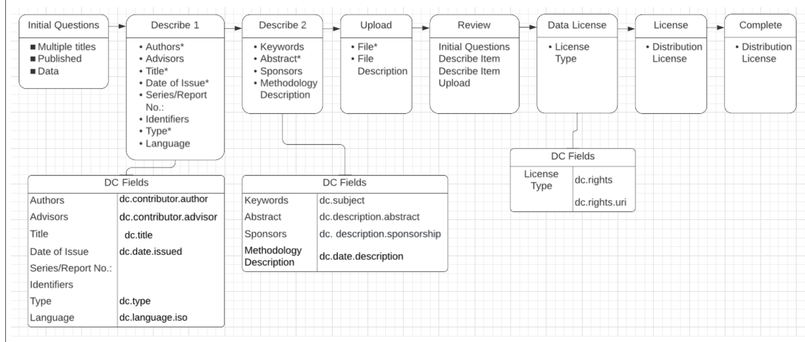
Controlling Dc.type

Dc.type is an existing controlled, non-repeatable field in the general DRUM workflow and is used to identify data records through the selection 'dataset.' However, there is an unknown quantity of datasets deposited in DRUM that are difficult to discover due to being misidentified under other material types. To prevent incorrect application of this controlled field, the Data Collection's workflow further limits the controlled vocabulary by only allowing users to select 'dataset' or 'software.' If a user seeks to deposit other material types alongside their data, this field limitation indicates that they will have to start a separate deposit for the non-data files.

Search Engine Optimization

As well as adding or customizing metadata fields to the Data Collection workflow, key fields of DRUM's DublinCore metadata were mapped to schema.org's properties. By marking up these elements, DRUM provides structured data to search engines, optimizing dataset findability from external searches.

Self-Submission Workflow Overview



Ongoing Considerations

While the Data Collection workflow addresses several past issues with dataset submission and curation, certain ongoing considerations remain. Several functions available through dedicated data repository software that support best practices in data archiving are not supported by DRUM's Dspace software at this time. This notably includes file versioning which allows researchers to add new versions of data files to a record as they actively collect data without overriding or replacing previous versions, capturing an additional piece of the data lifecycle. However, while dedicated data repository software may be considered if resources become available in the future, DRUM's more general functionality currently provides a central location for research product deposit.
Ongoing considerations also include user ability to bypass the Dataset checkbox (and thus the Data Collection workflow) and submit their dataset amongst other research file types in one record, "hiding" their dataset from DRUM's current search capabilities. It is also possible for a user to bypass the data checkbox and still assign "dataset" as dc.type, which would exclude the record from the Data Collection but still capture essential identifying type metadata. While each of these scenarios are possible, there is not enough evidence yet to determine likelihood of occurrence. One instance of a researcher incorrectly defining files as data and submitting through the Data Collection workflow has occurred. However, close curation enabled by the creation of the Data Collection allowed DRUM administrators to detect this error and assign more accurate metadata to the researcher's record.
Lastly, Colavizza et al. demonstrate the importance of linking between publications and open, archived research data in increasing research citation and reuse. The Data Collection workflow does not currently include a dedicated field for links to associated research publications. The addition of such a field should be considered in any future changes to the collection.

Conclusion

The UMD Data Collection addresses several past issues of improper dataset deposits and curation difficulty in DRUM and better aligns the repository's infrastructure for archiving data with the FAIR Data Principles and other best practices. However,. well-prepared readme files deposited by researchers alongside data files in records remain essential to preserving more specific project-based or disciplinary metadata that cannot be captured in the current Data Collection self-submission workflow.
Continued consideration of DRUM's workflows and functionality for preserving these files is necessary as best practices in data archiving within institutional repositories are still developing. By creating a platform that demonstrates present institutional repository capabilities, enabling more effective metadata collection in the self-submission process, and increasing discoverability of its current dataset records, UMD Libraries are taking a productive step toward further development of its research dataset preservation services.
The full extent of the UMD Data Collection's impact on research data deposits and their preservation will be a future topic of exploration as more data becomes available.

Getting to Know the UMD Data Collection

110 dataset records

new records since implementation **11**

44.5% of dataset records deposited since 2020

dataset records support the work and publications of over **375** researchers worldwide

73.6% of dataset records directly support published, peer-reviewed works

Data collected 6/6/2022

References & Resources

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The Citation Advantage of Linking Publications to Research Data. *PLOS ONE*, *15*(4). https://doi.org/10.1371/journal.pone.0230416

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data Sharing, Small Science and Institutional Repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. https://doi.org/10.1098/rsta.2010.0165

Durden, D., & Buser, A. (2021, June 3). *Uncovering Hidden Datasets in DRUM*. Digital Repository at the University of Maryland. Retrieved from http://hdl.handle.net/1903/27104

Forero, D. A., Curioso, W. H., & Patrinos, G. P. (2021). The Importance of Adherence to International Standards for Depositing Open Data in Public Repositories. *BMC Research Notes*, *14*(1). https://doi.org/10.1186/s13104-021-05817-z

Lee, D. J., & Stvilia, B. (2017). Practices of Research Data Curation in Institutional Repositories: A Qualitative View from Repository Staff. *PLOS ONE*, *12*(3). https://doi.org/10.1371/journal.pone.0173987

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The Fair Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18