# Metadata Capital in a Data Repository

Jane Greenberg
SILS/Metadata Research Center/University of North Carolina at Chapel Hill
Chapel Hill, USA
janeg@email.unc.edu

Shea Swauger
SILS/Metadata Research Center/University of North Carolina at Chapel Hill
Chapel Hill, USA
scswauger@gmail.com

Elena M. Feinstein
Dryad Repository/Metadata Research Center
Chapel Hill, USA
elenamf@email.unc.edu

## Abstract

This paper reports on a study exploring 'metadata capital' acquired via metadata reuse. Collaborative modeling and content analysis methods were used to study metadata capital in the Dryad data repository. A sample of 20 cases for two Dryad metadata workflows (Case A and Case B) consisting of 100 (60 metadata objects, 40 metadata activities) instantiations was analyzed. Results indicate that Dryad's overall workflow builds metadata capital, with the total metadata reuse at 50% or greater for 8 of 12 metadata properties, and 5 of these 8 properties showing reuse at 80% or higher. Metadata reuse is frequent for basic bibliographic properties (e.g., author, title, subject), although it is limited or absent for more complex scientific properties (e.g., taxonomic, spatial, and temporal information). This paper provides background context, reports the research approach and findings. Research implications and system design priorities that may contribute to metadata capital are also considered.

**Keywords:** metadata capital; metadata reuse; Dryad.

## 1. Introduction

Metadata is a necessary component of any digital information system. Metadata helps people and machines find, access, and use information. Positive aspects aside, there are known challenges related to quality, cost, and standards that limit metadata effectiveness (Doctorow, 2001; Dimitrova, 2004; Nunberg, 2009); and most metadata operations are far from perfect. Recognizing metadata limitations is important, although there is danger in presenting any challenge as a rationale for not pursuing metadata. A more productive approach is to consider how to capitalize on and promote valuable aspects of metadata. In line with this goal, it is important to identify metadata workflows that yield positive results. One approach is to reuse good quality metadata where it is practical and advantageous. In this context, metadata may be seen to have greater value than its *original net worth*. Reuse of good quality metadata promotes *metadata capital*—a concept introduced in this paper.

Metadata capital is an asset. Reuse of good *quality metadata* over time and across systems increases the value of this asset (metadata). Reuse of poor quality metadata may reduce metadata capital. It is important to identify aspects of metadata workflows that help build metadata capital, and reduce those that consume capital. These goals are crucial for the Dryad repository[1]—a general-purpose repository for data underlying scientific publications.

Dryad's metadata architecture supports automatic propagation and aims to reuse good quality metadata where it is feasible (Greenberg, 2009). Dryad is ideal for investigating metadata capital and served as the test environment for this study. The following section of this paper explores the concept of 'capital' and identifies a range of metadata reuse practices. Next, the study's research questions, methods, sample, and procedures are reviewed, followed by the data analysis and a discussion of the results. The discussion also considers potential steps for increasing Dryad's

[1] Dryad: http://datadryad.org/

metadata capital and notes research limitations. The conclusion summarizes this research, highlights key findings and identifies potential research directions.

## 2. Capital and Its Application to Metadata

The notion of capital is most commonly used as an economic concept. Capital is a topic of focus in business and operations literature that applies to impacts (net gains or losses) specific to finances, goods and services, and public needs. Max Weber's *The Protestant Ethic and the "Spirit" of Capitalism*, first published in 1905 (2002), and Adam Smith's *The Wealth of Nations* first published in 1776 (2000), are commonly referenced for theoretical context. Many notions of capital exist, such as intellectual capital, which refers to individual or organizational knowledge aiding profit (Marr, 2005), and social capital, which stems from productive beneficial social relationships. The commonality among these and additional renderings of capital is that some tangible result or phenomenon (e.g., a product, knowledge, a friendship, etc.) has a value, and that the value can increase over time.

Metadata capital fits the above framework in some respects; it is a product generated by human labor and/or machine-driven processes. Metadata in the library and related environments can be viewed as a public good (a service facilitator) because it supports resource discovery and access. There is a tremendous opportunity to increase metadata capital simply by aligning metadata reuse and digital resource life-cycle management. This opportunity has motivated this research and the ideas presented in this paper.

## 3. Metadata Reuse: A Growth Indicator

Metadata reuse is prevalent in many information workflows simply because system generated metadata automatically travels with a digital object during the object's life cycle. Formal metadata reuse has, historically, been a part of library workflows at least since 1901, when the Library of Congress (LC) launched its card distribution service. Today, there is an array of infrastructure technologies and practices promoting metadata reuse. A list of library/scholarly communication conventions promoting metadata reuse includes the following:

- Cooperative cataloging programs implemented on local, national and international levels.
- Library of Congress' Cataloging in Publication (CIP) program[2] for registering bibliographic data prior to resource publication.
- Infrastructure protocols and models, such as the International Standard Bibliographic Description (ISBD), Machine Readable Cataloging (MARC), and Functional Requirements for Bibliographic Records (FRBR).
- Shared classificatory, terminological, and authorized naming systems (e.g., the Library of Congress Classification System (LCC), General Multilingual Environmental Thesaurus (GEMET), and the Virtual International Authority File (VIAF)).
- Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH)[3] -initiated in 1999 (Van de Sompel and Lagoze, 2000).
- CrossRef[4] and PubMed,[5] as well as tools and standards, such as Zotero,[6] BibTex,[7] Mendeley,[8] and DataCite.[9]

[2] CIP: http://www.loc.gov/publish/cip/.
[3] OAI-MHP: http://www.openarchives.org/OAI/openarchivesprotocol.html.
[4] CrossRef: http://www.crossref.org/.
[5] PubMed: http://www.ncbi.nlm.nih.gov/pubmed.
[6] Zotero: http://www.zotero.org/.
[7] BibTex: http://www.bibtex.org/.
[8] Mendeley: http://www.mendeley.com/.
[9] DataCite: http://www.datacite.org/

- Linked data/Semantic Web developments, supporting persistent identifiers for concepts and names.

Metadata reuse, facilitated via these developments, epitomizes metadata capital by harvesting, sharing, and reusing metadata. Here, it is imperative to clarify that metadata capital increases only via reuse of 'good quality' metadata. Reuse of good quality metadata enables time, labor, and financial resource savings; and the value of metadata increases from this savings and growth. Market capital is calculated differently, with duplication typically leading to either the inflation or decrease in product value.

Well-formed workflows should include lucrative metadata reuse. Despite this observation, metadata workflow and reuse studies appear limited. Research needs to examine metadata workflows, particularly if we are to improve approaches to building metadata capital in our digital systems. The study reported on in this paper begins to address this need by designing a method for exploring metadata reuse in the Dryad repository.

## 4. Research Questions

Dryad's metadata architecture has been designed to support metadata reuse and consequently build metadata capital. Achieving and maintaining this goal requires an assessment of current metadata practices—ideally during timely intervals. The research presented in this paper is an updated and extended investigation of previous work assessing Dryad's metadata workflow (Greenberg 2009, 2010, 2011). The following research questions guided this investigation:

1. Does metadata reuse within Dryad's curation workflow build metadata capital?
2. Where is metadata reuse most common or lacking?
3. How might Dryad's curation workflow and/or repository design be modified to enable greater metadata capital and efficiency?

## 5. Methodology

The research questions posited above were examined using collaborative modeling and content analysis methods. Collaborative workflow modeling is primarily conducted to aid project implementation, although the technique has implications for understanding instantiations within a system (Reijers, Song, and Jeong, 2009). The collaborative workflow was conducted to help gather an appropriate sample. Content analysis is an established method for examining metadata record quality (e.g., Moen, et al, 1997; Greenberg, et al, 2002). The combined methods support the scientific study of metadata reuse.

## 6. Dryad Context

Dryad, as noted in the introduction, is a general purpose repository for data underlying scientific publications. Dryad operates as a nonprofit organization; its core mission to make data accessible for research and educational reuse. Dryad content is submitted by authors, and each submission is associated with a scholarly publication—generally a peer-reviewed article.

Dryad's architecture centers around data packages. A data package consists of a metadata object that describes the scientific publication and associated data files. A data package may contain one or more data files and readme files. At the close of the first week of July 2013, a Dryad tally recorded 3,572 data packages, 10,280 data files, and 242 distinct journal titles associated with the full population of Dryad's holdings.

A range of relationships with journals and professional societies has spawned several Dryad submission workflows. These different workflows impact when and how metadata is generated. Three main workflows, identified as Case A, Case B, and Case C, are summarized in Table 1.

Table 1: Dryad Submission Workflows

| |
|---|
| **Case A: Integrated Submission (at manuscript acceptance):** Dryad submission occurs during the manuscript acceptance stage and prior to publication allowing DOI minting and inclusion in the publication. This is the most frequent submission workflow. |
| **Case B: Non-integrated Submission:** Dryad submission occurs sometime after the associated literature has been accepted for publication or after it has been published (this study only treats the post-publication case). In this case authors take the initiative to share their data, often after having experience with Case A. |
| **Case C: Integrated Review Submission:** Dryad submission occurs as part of the editorial or peer review process for the associated manuscript. In this case Dryad holds these submissions in a private passkey-protected workspace. |

## 7. Sample and Procedures

A convenient sample of 20 full cases (10 each for Case A and Case B) was gathered from January through March 2012. For Case B, only submissions associated with published articles were used. Case A and Case B were selected for sampling because they have analogous metadata workflows and comprise the majority of Dryad's deposits. Case A and Case B combined represented approximately 85% of Dryad's deposits received during the three-month data collection period. Figure 1 illustrates the workflows for Case A and Case B. The diagram is the result of a collaborative workflow modeling activity to conceptualize instantiations that comprise each case (Reijers, Song, and Jeong, 2009). There are two types of instantiations: 1) green rectangles represent metadata objects, and 2) purple arrows represent metadata editing phases or activities.

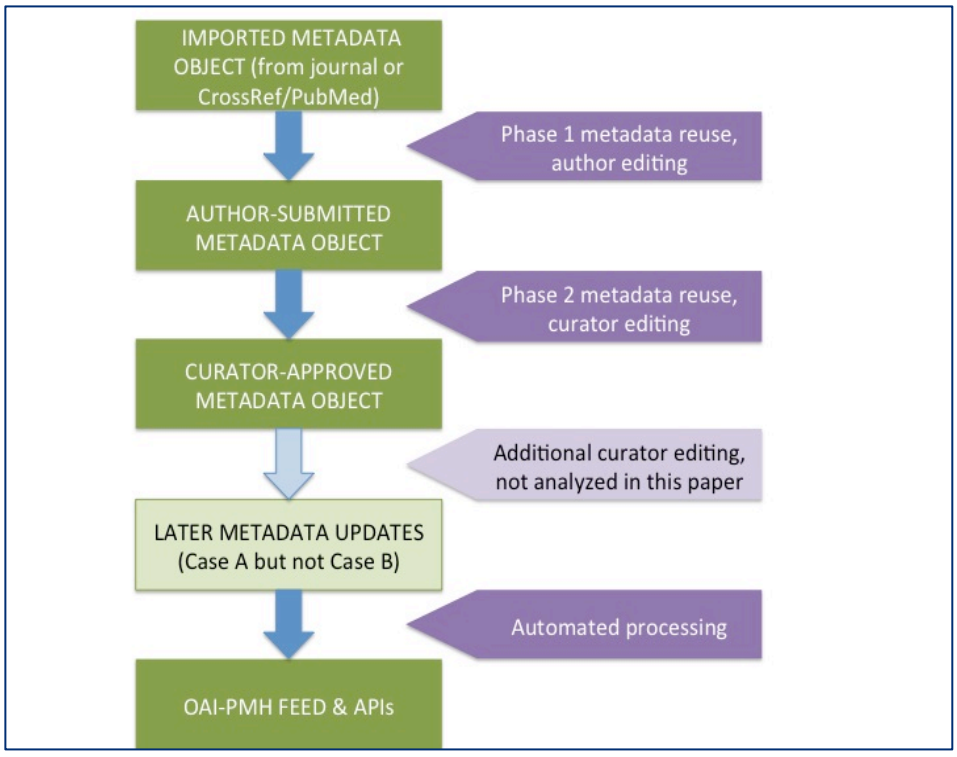



FIG. 1: Dryad Metadata Workflow: Objects and Editing Phases

Data collection involved the following tasks:

1. Capturing snapshots of three metadata objects:
   - Imported metadata object, before submitter editing
   - Author-submitted metadata object
   - Curator-approved metadata object
2. Archiving and converting the metadata snapshots into plain text.
3. Tracking content changes with Microsoft Word's Track Change feature during the transitions between Phase 1 and Phase 2:
   - Phase 1: imported metadata is reused and submitting author may make additions, deletions, and modifications to this metadata
   - Phase 2: author-submitted metadata is reused and Dryad curatorial staff may make additions, deletions, and modifications to this metadata
4. Flagging all metadata changes to be reviewed by curatorial staff. Each metadata change was categorized as one of the following:
   - Addition (metadata field added)
   - Deletion (metadata field removed)
   - Modification (metadata field edited)
   - Reuse (if a field was not flagged as changed, it was recorded as reuse)
5. Calculating all changes and reuse frequencies separately and then aggregating these results to give an overall picture.

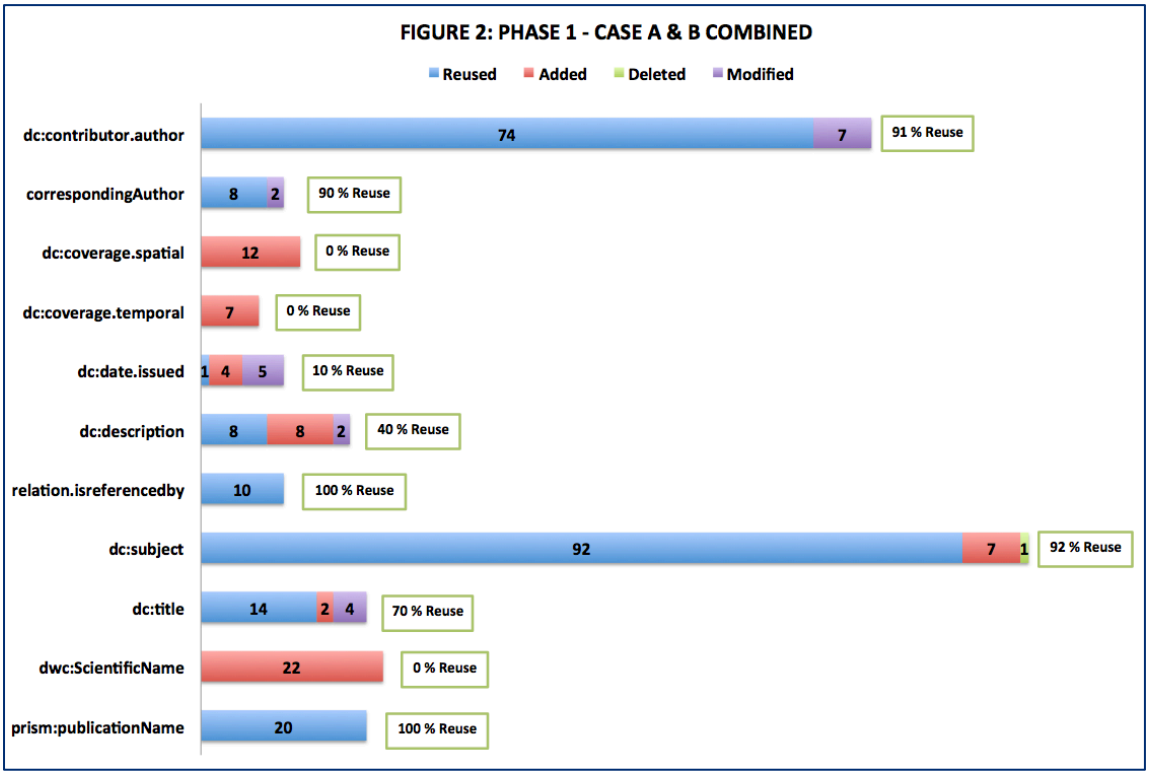
## 8. Data Analysis

Data analysis activities combined Case A and B—a total of 20 metadata workflow cycles—to give a more complete picture of metadata reuse and to gain a sense of Dryad's metadata capital. The analysis focused on package level metadata, and the full sample included 100 instantiations (60 metadata objects, 40 activities). The imported metadata objects (top green box, Figure 1) provided the baseline from which changes were measured. Table 2, column two, presents the total number of metadata entries per property for the imported objects (20 cases). Table 2, column three, reports new metadata entries per property, added by an author/submitter during Phase 1. The sequence of steps is illustrated as the second-level green box in Figure 1. These additional properties are also numerically presented and graphically shown in Figure 2, in red.

TABLE 2: Baseline Package-level Metadata

| METADATA PROPERTY | IMPORTED METADATA OBJECT | PHASE 1 METADATA ADDITIONS/ MODIFICATIONS |
|---|---|---|
| dc:contributor.author | 74 | 7 |
| dc:contributor.correspondingAuthor | 8 | 2 |
| dc:coverage.spatial | 0 | 12 |
| dc:coverage.temporal | 0 | 7 |
| dc:date.issued | 1 | 9 |
| dc:description | 8 | 10 |
| dc:identifier.citation | 0 | 0 |
| dc:relation.isreferencedby | 10 | 0 |
| dc:subject | 92 | 7 |
| dc:title | 14 | 6 |
| dwc:ScientificName | 0 | 22 |
| prism:publicationName | 20 | 0 |

The highest number of initial metadata entries was found for subject metadata (92 entries) and author name/contributor (74 entries). Figure 2 provides a graphical representation of Phase 1 author editing. The blue bar chart sections represent metadata reuse without any human intervention. High reuse for author name makes sense. Simply, accurate authorship information should appear in an article draft submitted for peer review and in a published article. The former case impacts the editor email sent to Dryad, and the latter impacts author name data harvested from CrossRef or PubMed. The red bar chart sections (Figure 2) show new metadata for spatial, temporal, date issued, description, subject, title, and scientific name (these results are also reported above in Table 2, column 3).

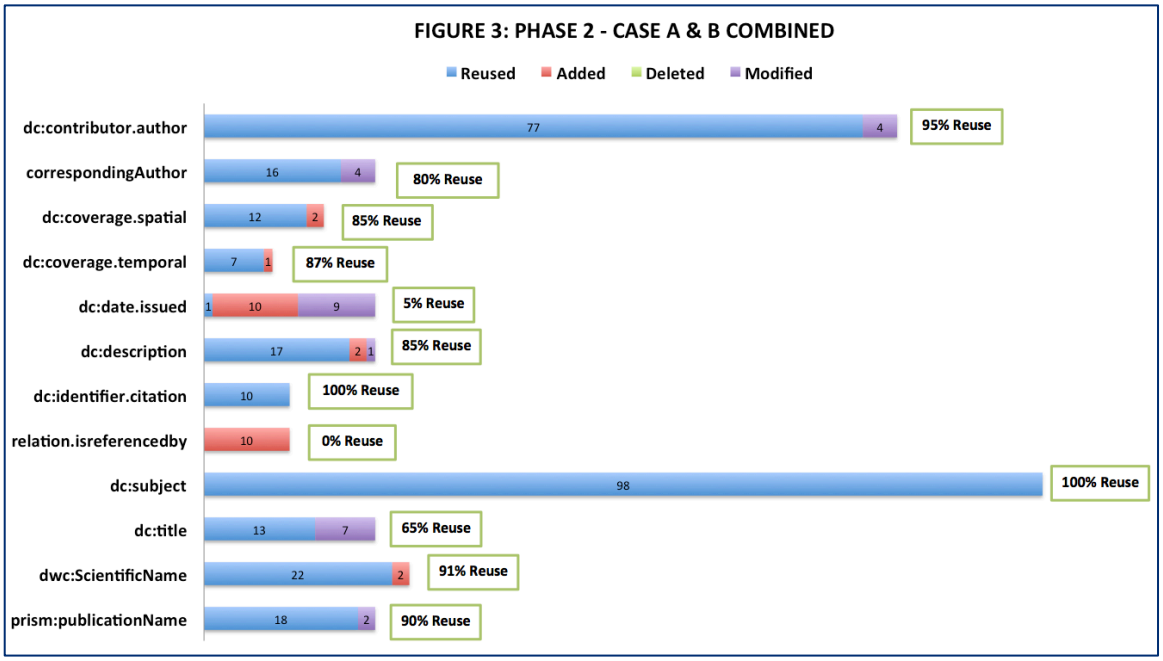


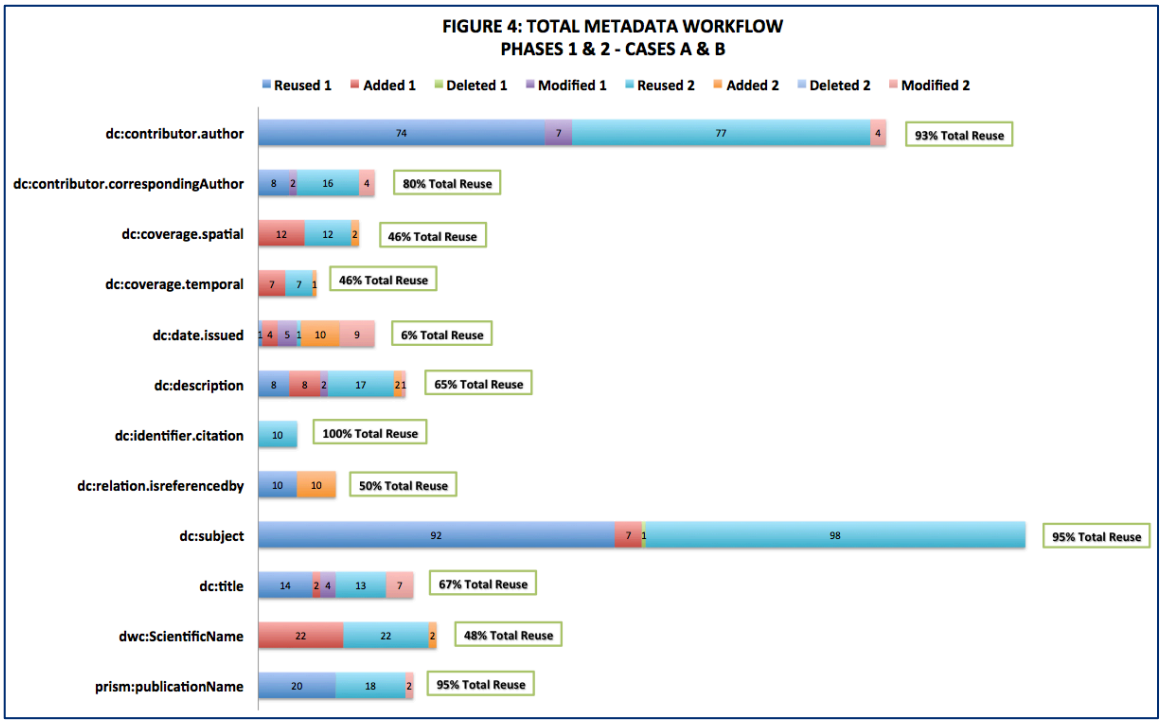

The purple bar chart sections in Figure 2 represent metadata reuse via human editing, specifically modification content. Essentially, the original metadata was deemed valuable, but required some attention. A modification is often as small as a capitalization correction, but it still requires human intervention. Only one subject was deleted during Phase 1.

Results for metadata reuse during Phase 2 are presented in Figure 3. The reuse of author-submitted metadata remains high, represented by the blue sections. Figure 3 indicates more original metadata generation during Phase 2. Curatorial staff oversees this phase and prepares the metadata for publication. New metadata is recorded for date.issued and identifier.citation; this work is currently executed manually, given system limitations.

A more complete picture of Dryad's metadata reuse combining Case A and Case B, Phase 1 and Phase 2 activities (initially imported metadata object to the curator-approved metadata object) is presented in Figure 4. The total reuse percentage is specifically for the metadata content that was reused without any human intervention and was deemed valuable at each stage. Results indicate that Dryad's overall workflow builds metadata capital, with the total metadata reuse at 50% or greater for 8 of 12 metadata properties, with 5 of these 8 properties showing reuse at 80% or higher. Metadata reuse is frequent for basic bibliographic properties (e.g., author, title, subject), although it is limited or absent for more complex, scientific properties (e.g.,

taxonomic, spatial, and temporal information). Reuse results vary for 'identfier.citation,' 'relation.isreferencedby,' 'date.issued' because of the different workflows presented by Case A and Case B, and it is challenging to give an overall assessment. However, each of these properties demonstrates reuse at selected junctures, and this reuse can be leveraged via system and workflow modifications.



## 9. Discussion: Toward a Capital Gain

Research questions guiding this study provide a framework for exploring the results and further study of metadata capital. This section discusses the findings, notes the study's limitations, and identifies several research contributions.

***Research question 1:*** The first question considered whether metadata reuse within Dryad's curation workflow builds metadata capital. The results of this research confirm a substantial amount of metadata reuse in Dryad. That is, Dryad's metadata workflow leverages the initial metadata object—a journal communication for Case A, and journal metadata from CrossRef or PubMed for Case B. Figures 2, 3, and 4 indicate metadata reuse for nearly all of the properties, with more than half demonstrating reuse above 50%. Dryad's workflow demonstrates a substantial amount of metadata reuse and likely builds capital.

Although Dryad's leaning toward a capital gain is likely the case, it is important to point out that this data does not encompass labor required when new metadata is being generated or when existing metadata requires modification. Dryad's curatorial staff members are positive about Dryad's overall reuse rate and believe capital is, ultimately, positive. Even so, items requiring human intervention for editing require resources and can interfere with workflow efficiency. Dryad's curation work is performed by both a full-time professional curator as well as several assistant curators. More data are needed on staff time required for manual curation tasks, in conjunction with staff pay level per task, in order to accurately calculate the capital investment and gain.

***Research question 2:*** The second question framing this study served to identify where metadata reuse is most common or lacking. Metadata reuse was common for basic bibliographic properties such as author, title, and subject. Metadata reuse was lacking—in some cases non-existent—for more complex, scientific properties, such as taxonomic, spatial, and temporal information. These results may point to source of capital reduction. Although manual generation at this stage results in better metadata than what can be achieved via automatic processes.

One result of Dryad's metadata workflow is high percentage of reuse for 'subject' metadata. Curators perform minimal quality control, and Dryad relies a fair amount on author/researcher domain expertise for subject representation. This current study focused on package level metadata. An earlier study also showed a high percentage of subject metadata from package to data files although 12% of package metadata was not used for individual files (Greenberg 2009, 2010). The overall high percentage of reuse for subject metadata is encouraging *prima facie*—and indicates quality although it does not confirm quality; we recognize this as a limitation of this work.

***Research question 3:*** The third question articulated an overriding goal of this study, which was to determine how Dryad's curation workflow and/or repository design might be modified to enable greater metadata capital and efficiency. A strategic consideration is to invest in system functionalities that promote the generation of quality metadata earlier in Dryad's metadata workflow. This will allow for greater metadata reuse and can build capital. The results of this study point to the following three areas where metadata reuse can be improved:

- *Complex, scientific* properties (temporal, taxonomic, and geographic terms). Metadata for these aspects of data need to be gathered earlier in the workflow. One way is through publishers specifically asking authors to note terms for these properties during a publication submission or review, so that data can be shared with Dryad. Another way is to use automatic metadata extraction applications earlier in the metadata workflow and match results against selected vocabularies or ontologies.
- *Subject metadata, quality improvement.* As reported in the Data Analysis section subject metadata reuse is fairly high, although quality control is limited to author/researcher expertise. A prototype for the HIVE (Helping Interdisciplinary Vocabulary

Engineering)[10] module was integrated into the curatorial module; this system allows for the automatic assignment of terms from controlled vocabularies. A HIVE-like application could further leverage the expertise of authors/researchers and add quality by applying concepts from standard terminologies.

- *Identifier.citation, relation.isrefernecedby, and date.issued metadata*. It seems counterintuitive that metadata representing a standard number is not generated automatically and reused consistently. The inconsistent and, at times, limited metadata reuse for these properties largely relates to Dryad's different workflows; and new strategies and workflow paths may improve metadata reuse for these properties.

Dryad developers are aware of the target areas outlined above, and are considering how to improve system functionalities in future system releases.

***Limitations:*** The research presented in this paper is based on a sample collected during a three month period and was limited to Dryad workflows represented by Case A and Case B. Additional data gathered via the less common workflows (i.e., a Case C) may yield different results. Although Case A and Case B comprise 85% of the 'types' of deposits Dryad receives, the sample size analyzed is this study represents roughly 36% of Dryad's average weekly submissions in numbers. Dryad received approximately 55 new submissions per week. Collecting, archiving, and converting the records for studying metadata reuse is detailed work, and the sample size was deemed sufficient for this initial study.

Another limitation may be seen in the application of the concept of metadata capital. There is risk as well as innovation in co-opting a concept from one field of study and placing it in another context. The work presented in the paper is at an early stage, and it not supported by a full analysis of metadata quality or formulas noting cost. Even so, human judgment leading to metadata reuse connotes quality on some level. We also know reuse of metadata generated by automatic means is less expensive than having authors or curators repeatedly generate metadata. Future studies will target a larger sample, use more automatic techniques for analysis, and explore cost.

***Significance:*** The work presented in this paper makes several important research contributions.

- The data presented in this paper support the first articulation of the notion of metadata capital.
- The methodology and approach provides a framework for future studies examining metadata capital contextualized via metadata reuse or potentially via other means.
- The data collected provide a base for comparing results of future studies.
- The results provide insight into Dryad's workflow and help to identify priority areas where more accurate metadata might be generated earlier in the metadata workflow.

## Conclusion

This study introduces the concept of metadata capital in the context of reuse. The study examined metadata reuse within the Dryad repository, tracking two workflows (Case A and Case B), during two phases of work—from the initially imported metadata object to the curator-approved metadata object that was published in the repository. Questions guiding the study examined whether metadata capital was apparent in Dryad's workflow, where reuse was positive or lacking, and how Dryad's curation workflow and/or repository design might be modified to enable greater metadata capital and efficiency.

Key findings indicate that:

[10] HIVE wiki: https://www.nescent.org/sites/hive/Main_Page, demo: http://hive.nescent.org/home.html.

- Dryad's workflows (Case A and Case B) has a substantial amount of metadata reuse, leading to metadata capital, with 8 of the 12 metadata properties demonstrating metadata capital via reuse at 50% or greater, and 5 of the 8 showing reuse at 80% or higher.
- Metadata reuse was common for basic bibliographic properties such as author, title, and subject, and it was lacking—in some cases non-existent—for more complex, scientific properties, such as taxonomic, spatial, and temporal information.
- System design priority areas were identified to promote the generation of more accurate metadata earlier in the metadata workflow.

Generating quality metadata at time of initial creation is paramount. An implication of the work presented here is that there is a benefit to generating better quality metadata upstream in an object's life cycle. It should also be recognized that there is probably a sweet spot (ideal place in a workflow) to consider for certain metadata properties. For example, an article's proposed title may change between the time it is submitted for peer review and when a final draft is submitted for publication. In fact, it is not unreasonable for a journal editor or reviewer to request a title change based on a review, and this can impact the title of a Dryad data package. The conclusions presented here are not necessarily novel; however, the context of metadata capital, supported by data, may be, and may help spark more research in this area.

In closing, Dryad metadata reuse continues beyond the work reported in this paper, via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and an API that integrates with the specialized DataONE API. These developments allow Dryad data to be harvested and used across many repositories. For these reasons, and the rationales underlying the research presented in this paper, it seems imperative to continue the study of metadata reuse and how to build metadata capital across our systems.

## Acknowledgements

The research in this paper is supported in part by the National Science Foundation (NSF), Award number: 1147166/ABI Development: Dryad: scalable and sustainable infrastructure for the publication of data.

## Data Access

Data reported on in this paper has been deposited in the Dryad repository: doi:10.5061/dryad.8c1p6. Data citation given below in References.

## References

Doctorow, C. (2001). Metacrap: Putting the Torch to Seven Straw-men of the Meta-utopia: http://www.well.com/~doctorow/metacrap.htm.

Dimitrova, N. (2004). Is It Time for a Moratorium on Metadata? IEEE Multimedia, 11(4): 10-17.

Greenberg, J. (2010). Data Curation Research Summit: Panel on Current Directions in Research: Projects and Perspectives, December 2010, Chicago, IL.

Greenberg, J. (2009). Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption. Cataloging & Classification Quarterly, 47(3), 380-402.

Greenberg, J., Pattuelli, M., Parsia, B. & Robertson, W. (2002). Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information, North America: http://journals.tdl.org/jodi/index.php/jodi/article/view/42/45.

Greenberg, J., and Vision, T. (2011). The Dryad Repository: A New Path for Data Publication in Scholarly Communication", OCLC, April 25, 2011, Dublin, Ohio: http://www.oclc.org/research/news/2011/03-24.html.

Greenberg J., Swauger S, & Feinstein E.M. (2013). Data from: Metadata capital in a data repository. Dryad Digital Repository: doi:10.5061/dryad.8c1p6.

Nunberg, G. (2009). Google's Book Search: A Disaster for Scholars. The Chronicle of Higher Education: http://chronicle.com/article/Googles-Book-Search-A/48245/.

Smith, A. (2000). The Wealth of Nations. New York: Modern Library.

Weber, M. (c. 2002). The Protestant Ethic and the "Spirit" of Capitalism and Other Writings (edited, translated, and with an introduction by Peter Baehr and Gordon C. Wells. New York: Penguin Books.

Marr. B. [ed] (2005). Perspectives on Intellectual Capital Amsterdam: Elsevier Butterworth Heinemann.

Moen, W.E., Stewart, E.L., and McClure, C.R. (1997). The Role of Content Analysis in Evaluating Metadata for the US Government Information Locator Service (GILS): results from an exploratory study: http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm.

Reijers, H.A., Song, M. and Jeong, B. (2009). Analysis of a Collaborative Workflow Process with Distributed Actors, Information Systems Frontiers, 3 (11): 307-322.

Van de Sompel, H., and Lagoze, C. (2000) The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2): http://dx.doi.org/10.1045/february2000-vandesompel-oai.