Greenberg, J., White, H., C, Carrier, C. and Scherle, R. (in press). A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata. [24 manuscript pages.] [Special issue on metadata best practices]
Journal homepage: http://www.informaworld.com/smpp/title~content=t792306902~db=all

Title

A Metadata Best Practice for a Scientific Data Repository

Authors

Jane Greenberg*, Hollie C. White, Sarah Carrier, and Ryan Scherle

*Corresponding author email: janeg@email.unc.edu; telephone: 919-962-8066

Abstract

Digital data repositories ought to support immediate operational needs and long term project goals. This paper reports the Dryad Repository's *metadata best practice* balancing these two needs. The paper reviews background work exploring the meaning of science, characterizing data, and highlighting data curation metadata challenges. The Dryad repository is introduced, and the initiative's metadata best practice and underlying rationales are described. Dryad's metadata approach includes *two-prongs*: one addressing the long-term goal to align with the Semantic Web via a metadata application profile; and another addressing the immediate need to make content available in DSpace via an Extensible Markup Language (XML) schema. The conclusion summarizes limitations and advantages of the two prongs underlying Dryad's metadata effort.

Keywords: Metadata; Scientific data; Dublin Core Application Profile; Singapore Framework; Semantic Web

Acknowledgement

This work is supported by National Science Foundation Grant # EF-0423641. We would like to acknowledge contributions by the Dryad team members Hilmar Lapp and Todd Vision of NESCent; and Michael Whitlock, University of British Columbia. We would also like to thank Stuart Weibel, OCLC, for his thoughtful comments and support of with this work.

Introduction

The digital revolution is transforming the way scientific research is conducted. Proposal management and project reporting are increasingly free of physical media. Even data collection is frequently conducted using digital technology, and manually collected data is nearly always converted to a digital form for analysis. A significant consequence of this change is the burgeoning development of repositories providing access to data descriptions and often the data objects themselves. Examples include the DOE (Department of Energy) Data Explorer (DDE)[1]; the Knowledge Network for Biocomplexity Data (KNB)[2]; and institutional and consortial repositories storing members' research data (e.g., Rice, 2008; Witt, 2008).

The growth in data repositories is further motivated by opportunities for sharing research data across communities, domains, and time. To this end, scientists and sponsors of scientific research are calling for greater interoperability among data stores. Metadata figures prominently in these calls; and many rich data-driven metadata standards embed properties from the Dublin Core[3]—a base level set of metadata properties that can facilitate interoperability among digital information systems.

Metadata developments like the Dublin Core have, indisputably, improved access to the world's store of digital information. However, the unprecedented growth of digital data, combined with pressure to attain some form of immediate interoperability among repositories, has generally forced a continuation of legacy descriptive practices. Only more recently have there been efforts to *take-a-step-back* and consider what may be achieved in our digital information environment by rethinking traditional resource oriented information models. The Dublin Core Abstract Model (DCAM)[4] and the Singapore Framework for Dublin Core Application Profiles[5] represent efforts in this area, moving from the resource-driven legacy approach representing an information package, to focusing on the component parts of a resource description. These relatively new information models (in a loose sense) align with the Semantic Web, and are intended to better support automatic data synthesis and reuse of metadata; and they are guiding decisions underlying the Dryad Repository's metadata practice.

This paper reports the Dryad Repository's incorporation of these models into a *metadata best practice*. The paper begins by presenting the Dryad development team's background work exploring the meaning of science, characterizing data, and highlighting data curation metadata challenges. Next, the paper gives an overview of Dryad and presents the initiative's metadata

best practice and underlying rationales. Dryad's metadata approach includes *two-prongs*: one addressing the long-term goal to align with the Semantic Web via a metadata application profile; and another addressing the immediate need to make content available in DSpace via an Extensible Markup Language (XML) schema. The final part of this paper summarizes limitations and advantages of the two prongs and concludes by noting next steps for Dryad's metadata effort.

2. Understanding Science

Dryad operates in the world of Science. The most basic understanding of science can be defined through those things that we can see, touch, and hear. Science is not based on opinion, but grounded in observation and fact (Chalmers, 1999). This structured view generally encompasses the scientific method—a framework supporting research verification and enabling reliable knowledge advancement (Whewell, 1989; Gower, 1997). Science is conducted in laboratories, natural habitats, and environments known as "living labs" engineered for sociological study (Latour and Woolgar, 1986). Diversity among study environments helps explain why some scholars claim that modern natural science, humanities, and social sciences are all a part of "science" (Szostak, 2004; McKeon, 1994), a view reflecting Aristotle's broad interpretation that science (*dianoia*), in all respects, is practical, poetical or theoretical (Aristotle's *Metaphysics*). Regardless of where the science is conducted, or what phenomenon is studied, the fundamental goal is to *attain knowledge*.

According to Carlson (2006), "science is experiencing revolutionary changes thanks to digital technology, with computers generating a flood of valuable data for scientists to interpret." Cyberinfrastructure, wiki-type web pages, and new large storage possibilities are helping to more rapidly advance scientific endeavors (Borgman, 2007; Doctorow, 2008; Waldrop, 2008). These technological developments have encouraged scholars to probe what science actually is in our digital age, and support the emergence of *eScience*—a concept denoting the use of digital technology for solving scientific problems. In brief, eScience has been integrated into our understanding of science today, and metadata is a fundamental component of any eScience enterprise.

2. Data and Data Curation Metadata Challenges

A principal connection between eScience and our more general conception of science is *data*. Data is the essence of science—the holder of scientific truth when findings are reviewed. Davis (2007) claims that data sets can range "from geographic information systems or geospatial (GIS) data to genomic data to any data set supporting a scholarly publication, such as census data;" while White (2008) confirms a fuller scope of data that includes traditional observations, numbers and measures stored in spreadsheets and databases, as well as fossils, phylogenetic trees, and even herbarium samples.

Today, the tremendous growth in digital data presents what Lord, et al (2004) call the "data-deluge"—a predicament helping to shape an emerging field of data-curation. Motivating and interconnected with data curation work are a number of metadata challenges:

- Data sought for repository deposition is being generated at a much faster rate than metadata can be created for representing, organizing, and access data. In fact, Doctorow (2008) states that the size of the data being collected can no longer be managed by a single scientist alone.
- Legacy resource-oriented approaches are common: Likely reasons for this continued practice include the unprecedented growth in digital data, intense and growing demand for interoperability, and limited time for rethinking descriptive practices or trying new models (Greenberg, 2009).
- Biocuration (curation by professional scientists) is being practiced, but projects seeking to employ skilled curators are not being funded as readily as data collection initiatives (Howe, 2008).
- Researchers prefer rich descriptive metadata supporting discovery and reuse, although they are not necessarily dedicated to allocating time required for creating good quality metadata (Greenberg et al, 2001).
- Metadata generation is inefficient, with automatic applications not being fully employed, and often the same metadata being generated via humans in more than one setting (Greenberg, 2009).

The data curation/metadata related challenges outlined here highlight pressing questions about the processes and procedures by which digital data should be preserved, organized,

accessed, and used. In pursuing solutions, new partnerships have been formed between scientific disciplines and the field of information and library science (Carlson, 2006; Davis, 2007; Heidorn, 2008). This development is evident via cyberinfrastructure and repository initiatives seeking to develop user-friendly systems enabling scientists to deposit and share digital data. These developments are impacting repository design, and they are shaping the Dryad repository's metadata practice—the topic of focus for the remainder of this paper.

3. The Dryad Repository

The Dryad Repository[6] is designed for the preservation, access, and reuse of scientific data objects underlying published research in the field of evolutionary biology, ecology and related disciplines. Dryad is supported by the National Science Foundation, and being developed via a collaboration involving the National Evolutionary Synthesis Center (NESCent); the Metadata Research Center at the School of Information and Library Science, University of North Carolina at Chapel Hill (UNC/CH); North Carolina State University; University of New Mexico; and Yale University. Additional partners include major societies and journals in the field of evolutionary biology[7] .

Dryad has been implemented within DSpace (Scherle et al 2008; White et al 2008); and data deposition is currently voluntary. Dryad is being developed in anticipation of a future requirement of deposition upon publication, in compliance with a Joint Data Archiving Policy (JDAP)[8] being adopted by a consortium of ecology and evolution journals. A key focus of Dryad development work has been designing a metadata infrastructure supporting an user-friendly and burden-free data deposition process. As part of this effort, the Dryad development team has identified high-level metadata functional requirements; and they are guiding Dryad's two-pronged metadata best practice.

4. Dryad's Metadata Requirements: Supporting a Two-pronged Metadata Best Practice

A first step for Dryad's development was delineating the repository's functional requirements. These are reported on in Dube, et al (2007), and White, et al (2008). Briefly summarized, they include developing a repository supporting resource discovery and use; data interoperability; computer-aided metadata generation, augmentation, and quality control; linking publications and underlying datasets; and data security. These repository functionalities

are reflected, to a large degree, in Dryad's high-level metadata requirements, stipulating a metadata architecture that is as follows:

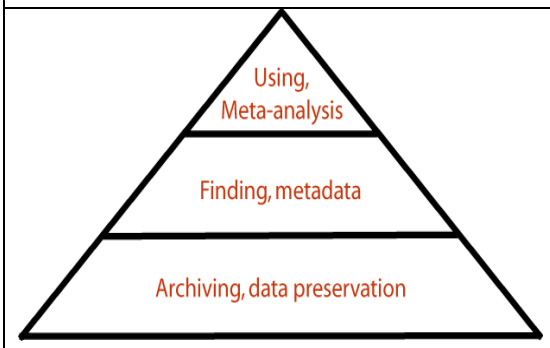
- *Simple*. Dryad's metadata properties (represented in an application profile or schema) are basic and understandable to the average Dryad depositor or user. This is imperative, given that depositors are required to create some metadata during deposition. A simple scheme is easier to manipulate for automatic metadata generation and quality control. Additionally, a simple, basic scheme can collectively support the description and discovery of heterogeneous resources, including data objects.
- *Interoperable*. Dryad's metadata system is based on the Dublin Core, a low-level scheme developed to facilitate interoperability that is often used, or at least reflected to some degree, in many systems. This design more easily supports metadata harvesting, cross-system searching, and metadata exchange with other formats. (A long term goal is to achieve greater interoperability corresponding to the recommended Interoperability Levels for Dublin Core Metadata [9].)
- *Semantic web compatible*. Dryad's metadata architecture conforms to the Dublin Core abstract model, which aligns with the Semantic Web. The long term goal is to create a sustainable and adaptable metadata architecture, supporting machine processing, including automatic data synthesis, and potentially other uses of Dryad metadata and data not yet envisioned.

These requirements form the core of Dryad's metadata best practice that is both *sustainable in the evolving digital information environment* and also *functional now*. Articulating these requirements has been important for developing Dryad's two-pronged metadata best practice, supporting these two integral needs.

Dryad's metadata requirements have further been developed inline with Dryad's data hierarchy of needs (Figure 1A), conceived by Dryad management board member, Michael Whitlock, Professor, Department of Zoology, University of British Columbia, and former Editor-in-Chief of *The American Naturalist* (2006-2008); these needs were first introduced at the 2006 Dryad stakeholders' workshop[10]. Inspired by Maslow's hierarchy (see Figure 1B), Whitlock considered Dryad scenarios requiring immediate attention, such as basic archiving

and data preservation, as well as long-term, high-level (top-of-they-pyramid) goals of *meta-analysis*. Whitlock's model targets the sheer value of data, and underscores that metadata without access to data is useless.

Figure 1A-B: Parallels between Dryad's Data Hierarchy and Maslow's Hierarchy of Needs

1A. Dryad's data hierarchy of needs✹	1B. Maslow's hierarchy of needs★
	

✹ https://www.nescent.org/wg_dryad/Dec_5_Workshop_Minutes
★ Maslow, http://en.wikipedia.org/wiki/Maslow_hierarchy_of_needs [Wikimedia Commons is a media file repository/public domain]

In parallel with Maslow's hierarchy, a data object at the top of Dryad's data hierarchy (Figure 1A) may operate as an autonomous agent and be a part of meta-analyses. Theoretically, a data object at this top level has reached its *full potential*; the top level data object is parallel, in an atypical way, to Maslow's self actualization level (Greenberg, 2008). Not to lessen the impact of Maslow's theory, but his work provides a useful framework for communicating significance of coupling metadata and data for data to be used in new ways. The desire to construct a metadata architecture supporting Dryad's data hierarchy's top level, along with more immediate base-level functions, are reflected in Dryad's two-pronged metadata best-practice. In brief, Dryad's development team has steered clear of *solely* continuing legacy information management practices; and, instead, has made a commitment to exploring new approaches meeting long term project goals, with an eye toward enabling data to meet its full potential. The result is a *two-prong* metadata best practice—with one prong addressing the long-term goal to align with the Semantic Web via a metadata application profile, and another prong addressing the immediate needs that entails continuing proven legacy metadata curatorial practices.

5. Dryad's Application Profile

The Dryad's development team pursued creating a Dublin Core Application Profile (DCAP), complying with the Singapore Framework, to be sustainable in the evolving digital information environment. Application profiles increase interoperability by incorporating registered properties used in other systems. Drawing properties from multiple formal schemes circumvents limitations arising from using only a single scheme. Furthermore, it makes no sense to create an entirely new scheme when other initiatives have taken the time deliberating and formalizing sufficient metadata properties (Greenberg and Severiens, 2007). Dryad pursued developing an application profile for these very reasons, following recommendation by Heery and Patel (2000); the initial application profile work is documented in Carrier, et al, (2007) under the project's earlier name DRIADE (Digital Repository of Information and Data for Evolution).

The Singapore Framework is a loose standard (or model) containing guidelines for establishing DCAPs—Dublin Core endorsed application profiles. The guidelines define the following DCAP components: 1. functional requirements; 2. domain model; 3. description set profile; 4. usage guidelines; and 5. encoding syntax guidelines. Complying with the Singapore Framework includes the benefits of consistency, long-term quality control, and interoperability with other metadata structures (Nilsson, et al 2008). Perhaps most important is the framework's guidance for creating standard machine processable metadata suited for Semantic Web applications. This point is significant for Dryad's top-level data hierarchy goals (Figure 1A), and the capacity for data to reach full potential (Greenberg, 2008).

Although the Dryad team's initial application profile work pre-dates the release of Singapore Framework, complying with this standard has been a fairly straightforward and useful process. A chief reason is that the Dryad development team's application profile work had naturally included developing component parts of the Singapore Framework, prior to the realization and publication of these guidelines. For example, articulating project functional requirements and illustrating a domain model were seen as initial necessary steps and undertaken by the Dryad development, although the concepts spelled out in the Singapore Framework were not used to label this initial. The availability of the Singapore Framework actually helped the team to better understand work underway to create and maintain a metadata

application profile. Progress made addressing all five Singapore Framework components is summarized in Table 1.

Table 1: The Dryad Repository: Summary of Singapore Framework Components

Singapore Framework component	Status and indication of progress
Functional requirements	Dryad repository's functional requirements are fairly well documented (e.g. Dube, et al, 2007; Carrier, et al, 2007). The Dryad team has also identified high level metadata functionalities reported on above (see: "Dryad's Metadata Requirements…").
Domain model	A copy of the first domain model is found in (Carrier, 2008; White et al, 2008). The model is currently being revised to reflect more relationships, resulting from handshaking with TreeBASE and other systems.
Description set profile	The Dryad Description Set profile is maintained on the Dryad Development wiki at: http://www.unc.edu/~scarrier/dryad/DSPLevelOneAppProfDraft.xml
Usage guidelines	The Dryad Usage guidelines are maintained on the Dryad development at: https://www.nescent.org/wg_dryad/Dryad_Version_1.0_Cataloging_Guidelines.
Encoding syntax guidelines	Dryad's XML schema for current project operation is found at: https://www.nescent.org/wg_dryad/Metadata_Profile. RDF draft encoding is found in Carrier (2008), and further work is underway following: http://dublincore.org/documents/dc-rdf/.

The Dryad development team is continuing work to comply with the Singapore Framework as a metadata best practice for being sustainable in the future of evolving digital information. Dryad's application profile meets the level one interoperability requirements, and displays elements of level two—both following the Interoperability Levels for Dublin Core Metadata[9] recommendation, noted above under Dryad's metadata requirements. Details on Dryad's Singapore Profile and DCAM work, and metadata interoperability developments, are being maintained on the Dryad Development wiki. Defining the overall metadata structure and identifying appropriate properties (semantics) have been key steps in developing Dryad's application profile; and the following subsections provide further details on these outcomes.

Dryad's DCAP: structure and semantics

Dryad's application profile has two modules: 1. a *citation module* for representing the

published research associated with the underlying data; and 2. a *data object module* for representing individual data objects or clusters of data objects underlying the published research. The citation module (Table 2) includes 18 properties, drawing from the Dublin Core[11]; Darwin Core[12], Publishing Requirements for Industry Standard Metadata (PRISM)[13] overseen by IDEAlliance; and the Journal Publishing Tag Set Tag Library [14] developed for the National Library of Medicine. This module supports the automatic generation for a journal article citation. The properties not generally included in a citation, such as abstract, keywords, taxonomic information, further enrich the description of a published article. Additionally, selected citation module metadata is automatically propagated as data object metadata (Greenberg, 2009). For example, keywords (representing a journal article) are automatically reproduced for associated data object representations (data underlying the published research), and a depositor can modify these keywords.

Table 2: Dryad Citation Module, Version 2.0

Namespace: Name/label	Obligation R=Required O=Optional	Metadata generation method	Cardinality R=Repeatable NR=Non-rep.
1. dc:type/Type	R	Automatic	NR
2. dc:creator/Author	R	Automatic	R
3. dcterms:issued/Date of Publication	R	Automatic	NR
4. journalpublishing3:Article-title	R	Automatic	R
5. journalpublishing3:Journal-title	R	Automatic	NR
6. prism:Volume	R	Automatic	NR
7. prism:issueIdentifier	R	Automatic	NR
8. prism:startingPage	R	Automatic	NR
9. prism:endingPage	R	Automatic	NR
10. dc:publisher/Publisher	R	Automatic	NR
11. dcterms:identifier/DOI for Published Article	R	Automatic	NR
12. dcterms:hasPartOf/Dataset Identifier	R	Automatic	NR
13. dcterms:isPartOfSeries/ISSN of Journal (*if journal does not have ISSN, journal title is put here)	R	Automatic	NR
14. dcterms:abstract/Abstract	O	Automatic	NR
15. dc:subject/Keyword	R	Automatic	R
16. Darwin Core:Scientific Name/Taxonomic Name	O	Semi-automatic	NR
17. dcterms:spatial/Locality	O	Semi-automatic	NR
18. dc:rights/Rights Statement	R	Automatic	NR

The data object module (Table 3) has 21 properties; and property definitions and endorsed controlled vocabularies are found in Table 4. The majority of data object module properties are drawn from the Dublin Core (again, this includes both the DCMES and the DCMI Metadata Terms. Properties are also drawn from the Darwin Core, Data Documentation Initiative (DDI)[15], Ecological Metadata Language (EML)[16], and PREservation Metadata Implementation Strategies (PREMIS) [17]; and one property labeled "status"[18] is declared in the Dryad namespace[19].

While developing the Dryad curation workflow, the need to indicate a metadata record status became increasingly apparent. Status, for Dryad, indicates a metadata record's *life-cycle position*. *Minimal* and *full level* cataloging information recorded in the MARC bibliographic header, and *meta-metadata* indicating metadata revision in schemes such as the TEI (Text Encoding Initiative) header [20], give some indication of status. However, these properties are not fully suited for Dryad's needs. Dryad's status property is being targeted to specify when a metadata record is prime for review, or has been reviewed by a curator; and when the metadata record can be made publically accessible.

Dryad's status property is part of the data object metadata header and records meta-metadata referencing the collection of other data object properties. Status is only applicable to Dryad data object metadata. It is possible that "status" may be considered for the citation module at as the project progresses, although there does not seem to be an explicit need at the time.

Table 3: Dryad Data Object Module, Version 2.0

Namespace: Name/label	Obligation R=Required O=Optional	Metadata generation method	Cardinality R=Repeatable NR=Non-rep.
1. **HEADER:** dryad:status/Status	R	Manual	NR
2. dc:type/Type	R	Automatic	NR
3. dc:creator/Author	R	Semi-automatic	R
4. dc:contributor/Contributing Author	O	Semi-automatic	R
5. dc:title/Dataset Title	O	Manual	NR
6. dc:identifier/Dataset Handle	R	Automatic	NR
7. dcterms:isPartOf/DOI of Published Article	R	Automatic	NR
8. DDI:depositr/Depositor	R	Semi-automatic	NR
9. DDI:contact/ Primary Contact – "Primary Contact" (instead of corresponding author)	R	Semi-automatic	R
10. dc:rights/Rights Statement	R	Automatic	NR
11. dc:description/Description	O	Manual	NR
12. dc:subject/Keywords	R	Automatic	R
13. Darwin Core:Scientific Name/Taxonomic Name	O	Semi-automatic	R
14. dcterms:spatial/Locality	O	Semi-automatic	R
15. dcterms:temporal/Date Range	O	Semi-automatic	R
16. dcterms:issued/Date of Issue	R	Automatic	NR
17. dcterms:available/Embargo Date	O	Semi-automatic	NR
18. dcterms:modified/Date Modified	R	Automatic	NR
19. dcterms:format/File Format	R	Automatic	NR
20. dcterms:extent/File Size	R	Automatic	NR
21. PREMIS:fixity/(hidden)	R	Automatic	R

Table 4: Dryad Data Object Module, Version 2.0: Property Names, Definitions, and Standard Vocabularies

Property Name	**Definition**	**Vocabularies** [* = vocabulary key]
1. **HEADER:** dryad:status/ Status	Status of the metadata record [hidden from public view]	A code list is being developed.
2. Type	The nature or genre of the resource.	DSpace controlled vocabulary.
3. Creator/ Author	The author lead author, author chiefly responsible for the content.	LCNAF* if name is established.
4. Contributing Author	The entity or entities contributing to the creation and development of the data set.	LCNAF if name is established.
5. Dataset Title	Descriptive title of the dataset.	
6. Dataset Handle	Unique identifier of the dataset.	
7. Handle of Published Article	Handle of the published article with which dataset is associated.	
8. Depositor	The name of the person(s) and/or institution(s) who deposited the study with the archive.	A name authority will be developed as depositors register.
9. Primary Contact	This field indicates who to contact with queries about the data.	A name authority file will be developed as primary contacts register.
10. Rights Statement	Statement regarding rights held over the resource.	
11. Description	Human-readable description of the dataset; an abstract or summary.	
12. Keywords	Dataset keywords.	NBII*, MeSH*, LCSH*, and other HIVE* and related vocabularies.
13. Darwin Core: Scientific Name/ Taxonomic Name	The full name of the lowest level taxon to which the organism has been identified in the most recent accepted determination, specified as precisely as possible.	IT IS*, uBio*, etc.
14. Locality	The spatial description of the data set specified by a geographic description and geographic coordinates.	TGN*, etc.
15. Date Range	The temporal description of the data set including start date and end date of the collection/creation of the data set.	
16. Date of Issue	Date of formal issuance (e.g., publication) of the resource.	
17. Embargo Date	A date after which the dataset will be made public.	
18. Date Modified	Date on which the resource was changed.	
19. File Format	The format in which the data set is stored. Can also represent software format.	Handled by DSpace.
20. File Size	The size of the file storage.	
21. Fixity	[An algorithm will detect if a file has changed; property is hidden from public view.]	

Key for Table 4: Dryad Data Object Module, Version 2.0: Property Names, Definitions and Standard Vocabularies

Abbreviation	Full title and Web access address
▪ ITIS	Integrated Taxonomic Information System: http://www.itis.gov/.
▪ HIVE	Helping Integrated Vocabulary Engineering is an IMLS supported project that is developing a SKOS (Simple Knowledge Organization System) vocabulary server supporting Dryad metadata generation and user searching. (The HIVE project model is applicable beyond Dryad.) The HIVE development wiki is: https://www.nescent.org/sites/hive/Main_Page; information on SKOS can be found: http://www.w3.org/2004/02/skos/.
▪ LCNAF	LC Name Authorities File, accessible as part of LC Authorities: http://authorities.loc.gov/.
▪ LCSH	Library of Congress Subject headings, accessible as part LC Authorities: http://authorities.loc.gov/.
▪ MeSH	Medical Subject Headings: http://www.nlm.nih.gov/mesh/MBrowser.html.
▪ NBII	National Biological Information Infrastructure Biocomplexity Thesaurus: http://thesaurus.nbii.gov/portal/server.pt.
▪ TGN	Getty Thesaurus of Geographic Names: http://www.getty.edu/research/conducting_research/vocabularies/tgn/.
▪ UBio	Universal Biological Indexer and Organizer. http://www.ubio.org/.

[Insert Example 1 here:

Title of example: **Example 1: Dryad Data Module Metadata Example for Morphology**

Data Set: http://hdl.handle.net/10255/dryad.600]

Long-term goals supported by Dryad's application profile

The DCAM and the Singapore Framework have played a significant part in shaping Dryad's application profile and building a foundation for reaching Dryad's top level goals on Dryad's Data Hierarchy of Needs (Table 1A). The Singapore Framework encourages using the Resource Description Format (RDF) and XML for expressing metadata. Figures 2 and 3 present RDF graphs for "author" and "subject" properties associated with the Dryad data object having the handle: http://hdl.handle.net/10255/dryad.600 and file name Brown_morphology.nexus (Dryad data object metadata presented in Example 1). RDF-XML renderings of these statements have a tremendous potential for reuse in different contexts, beyond Dryad.

Figure 2: RDF Graph: For Dryad Morphology Data Set (Brown_morphology.nexus)

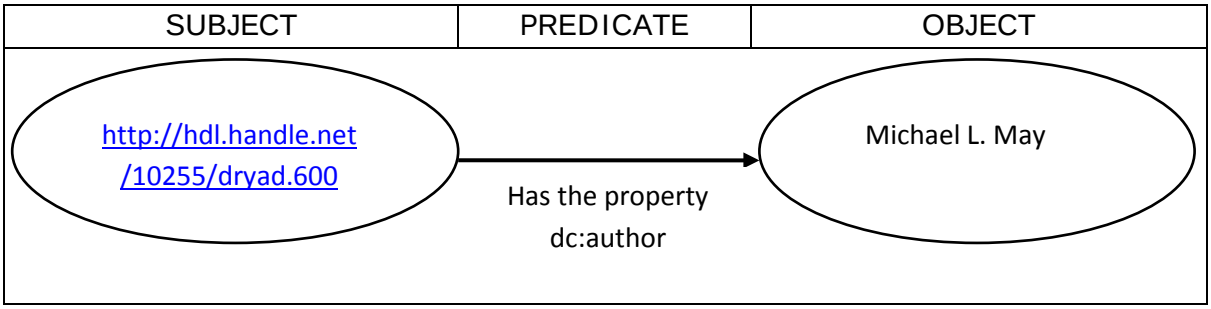
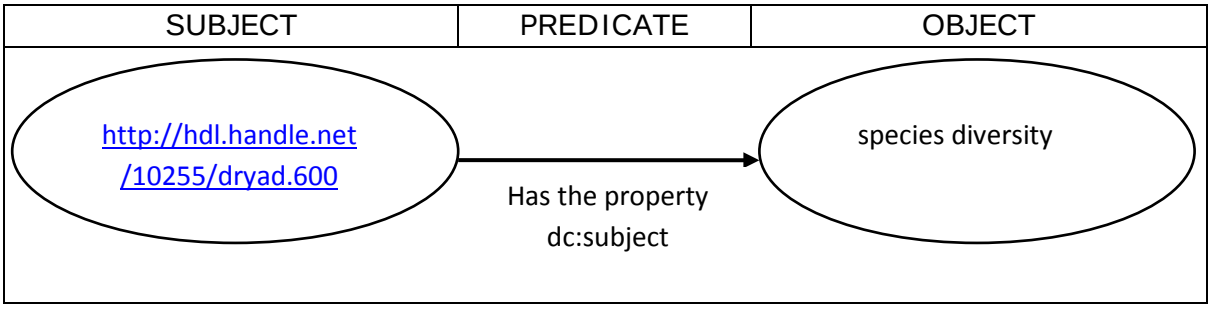


Figure 3: RDF Graph: For Dryad Morphology Data Set (Brown_morphology.nexus)



To further explain, consider how controlled vocabulary and standardized name headings are reused in many different library catalogs. Despite the benefits of these standardized systems, limitations stem from current efforts required for duplicating values stored in these systems. Headings may be automatically imported into a metadata record, but this task often

requires human oversight, verification, and sometimes re-coding. Consider also the benefits of copy cataloging and reusing core information in MARC (MAchine Readable Cataloging) bibliographic records. The reuse of metadata encoding, properties, and associated values is a major cost-savings, improve interoperability, and provides access to information more expediently.

Once a RDF statement (represented in XML) is created and made accessible for a property (such as a person's name), there is no need for recoding or editing; rather, the statement acts as an independent description and has the capability to become part of the global web of rich semantic linking. A RDF statement is, essentially, available for copy cataloging, but the statement represents a "property" not a complete information resource, such as a monograph, photograph, or archival collection. Metadata records representing information resources, in the envisioned future environment, are to be created by aggregating already existing property statements, such as those presented in the RDF graph examples (Figures 2 and 3). The reuse of RDF-XML statements intends to support robust linking beyond a metadata record, enabling both the discovery of new relationships beyond the confines of a single information system and inferencing—that is reasoning about information. This is reflective of Berners-Lee's notion of linked data[21], a development in which some sectors of the metadata community are quite engaged.

In moving toward the Semantic Web, a RDF statement about an author "Michael L. May" (from Figure 2), originating in one vicinity, can be reused and link work by this author/scientist in multiple systems. Furthermore, this can enable discovery of a much fuller spectrum of his work (publications, data sources, and other scientific contributions) potentially hosted in Dryad, GenBank[22], TreeBASE[23], and other scientific digital repositories. There are issues requiring further attention, as some properties, such as a "date," are easily represented by a URI—noted as a string, without any divergent representation; while other properties, such as a person's name, may be represented in multiple ways. Notwithstanding these challenges, the Dryad development team's efforts contribute to a robust framework supporting the generation of RDF statements for reuse in multiple contexts. Clearly, if this model is to be actualized more fully, other information sectors need to participate in this development as well. At present, the Dryad application profile moves forward a metadata best practice that, at the very least, aims to be sustainable—long term, and shows evidence of progress. And, finally,

the work completed has been important to developing Dryad's XML schema underlying the current repository operation.

6. Dryad's XML Schema

Indeed Dryad's metadata effort has emphasized the development of a DCAP complying with the Singapore Framework; and a considerable amount of time and discussion has been directed to the identification of appropriate properties for enriching publication citation information, and also describing data objects. In some respects, the Dryad development team has *reversed engineered* metadata development work by focusing on long-term goals as a first step. However, as soon as data exchange plans commenced, the development team's efforts diverged from DCAP-focused work to the construction of a basic XML schema. The initial undertaking in this area has included harvesting EML records from the Long Term Ecological Research (LTER) Network's Metacat data catalog[24], and a second effort underway will allow for data exchanged between Dryad and TreeBASE, a repository for published phylogenetic data. Dryad will regularly harvest metadata records from Metacat and TreeBASE, allowing users to extend searches to those records.

As the Dryad development team commenced work on the XML schema supporting these goals, it seemed like a "dumbing down" of the initial DCAP efforts, particularly given the tremendous amount of thought, time, and discussion supporting the development of Dryad's application profile. However, as Dryad's application profile was schematized, it became very clear that the two approaches were not necessarily at odds, and could instead be viewed as two renderings of the same intellectual work along a continuum. This was important to helping the Dryad development team understand the whole of their metadata effort, and make a commitment to the two pronged approach presented in this paper. The work on identifying appropriate properties from existing schemes has been instrumental in creating a sufficient XML schema. The approach taken show allow metadata produced following Dryad's XML schema to easily be rendered in RDF with enabling technologies like GRDDL (Gleaning Resource Descriptions from Dialects of Languages)[25], a mechanism for creating RDF data from XML documents. GRDDL is designed to convert metadata following Dryad's XML schema to metadata that is essentially equivalent with Dryad's application profile, and, thus, support long-term data sharing activities aligning with the Semantic Web. Although Dryad's

development team has not experimented with GRDDL, the development of this enabling technology is telling of efforts underway to enrich the web with linked data via RDF statements. Exploring GRDDL is an important consideration for Dryad's metadata planning as work continues to develop.

5. Conclusion

The metadata work reported on in this paper is valuable for any initiative supporting current operations and preparing for long term sustainability. The Dryad development team has discovered that a two prong hybrid philosophy works well in this environment. From the start, the Dryad team has been compelled to steer clear of legacy practices that may impede reaching future goals. In other words, placing an emphasis on long-term goals at the onset, and addressing immediate needs when absolutely necessary, has resulted in a flexible and adaptable metadata architecture.

The two prong approach reveals a number of additional obvious advantages. Pursuing the DCAP work has enabled Dryad's team to forward a metadata research agenda supporting long term goals presented in Whitlock's conception of Dryad's Hierarchy of Needs (Figure 1A). These goals support Dryad's metadata functional requirements of simplicity, interoperability, and Semantic Web alignment. Work completed thus far has been intellectually engaging for team members, provides machine processable metadata, and is advancing the field of metadata. The XML schema work has advantages as well. For example, it has easily accommodated the harvesting of LTER metadata, and the work accomplished should integrate with Dryad's application profile.

There have also been challenges in developing and pursuing the two prong metadata best practice reported on here. It is difficult to illustrate the benefits of the DCAP given the absence of metadata registries fully supporting Semantic Web agents. Furthermore, although new RDF ontology-driven efforts are frequent, the mass of RDF data is still limited. The DCAP effort has been time consuming, and Dryad's development team has, in a sense, had a luxury to deliberate and consider this approach from various points of view. Development work has been a consensus driven and team building activity—bringing together biologists, computer scientists, metadata researchers, and librarians for team building. The effort has been positive, although there have been obvious communication challenges, particularly given the

absence of documentation and examples illustrating the benefits of the DCAP. The XML schema work has required minimal resources, so it is difficult to note any real limitations at the moment. Even so, the Dryad development team anticipates unforeseen challenges may surface when modifying the XML schema to accommodate long term goals.

A key factor impacting Dryad's continuation of a two prong metadata best practice is Semantic Web development—a topic that is beyond the scope of this paper. However, what can be surmised here is that initiatives like Dryad *can* and *ought to* explore new information models as part of their best practice, particularly given the goals of information reuse—a goal that is being voiced by many in the digital data community.

The Singapore Framework and RDF-XML provide a potential opportunity advancing the use of metadata and associated information in this new context. Although *proof* of how DCAPs function for a complete Semantic Web activity is difficult, efforts are needed now to explore new models if we to advance our knowledge about and opportunities for data sharing. Considered from another vantage point—no exploration, no discovery; and even unsuccessful results can help narrow the scope of what needs to be accomplished to improve future outcomes. In conclusion, although Dryad's focus is the evolutionary biology domain, the two prong metadata best practice is relevant in many domains and digital initiatives preserving, facilitating discovery and access, and promoting the reuse of digital data and other resources; and Dryad's efforts are contributing a more robust, intelligent, and efficient web of linked data.

References

Aristotle (2007) *Metaphysics*. Translated by Ross, W. D. eBooks@Adelaide: http://ebooks.adelaide.edu.au/a/aristotle/metaphysics.

Borgman, C. L. (2007). *Scholarship in the digital age: information, infrastructure, and the internet.* Boston: MIT Press.

Carlson, S. (2006). Lost in a sea of science data. *Chronicle of Higher Education, 52*(42), A35.

Carrier, S. (2008). *The Dryad repository application profile: process, development, and refinement*. (Master's Paper, UNC Chapel Hill, 2008). DOI: http://hdl.handle.net/1901/534.

Carrier, S., Dube, J., & Greenberg, J. (2007). The DRIADE project: phased application profile development in support of open science. In *DC-2007: Application Profiles: Theory and Practice. International Conference on Dublin Core and Metadata Applications*, Singapore, August 27-31, 2007, pp. 35-42.

Chalmers, A. F. (1999). *What is this thing called science?* Indianapolis: Hackett.

Davis, H. & Vickery, J. (2007). Datasets, a shift in the currency of scholarly communication: Implications for library collections and acquisitions. *Serials Review, 33*: 26-32.

Doctorow, C. (2008). Big data: welcome to the petacentre. *Nature. 455* (7209): 16-21.

Dube, J., Carrier, S., & Greenberg, J. (2007). DRIADE: a data repository for evolutionary biology. In *Proceedings of the 2007 conference on Digital libraries*, Vancouver, BC, Canada, ACM Press, pp. 481.

Gower, B. (1997). *Scientific method: A historical and philosophical introduction*. New York: Routledge.

Heery, R., & Patel, M. (2000). Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne*, 25: www.ariadne.ac.uk/issue25/app-profiles/.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends, 57*(2): 280-299.

Herbermann, Charles G. et al., (Eds.) (1913). *The Catholic Encyclopedia: An International Work of Reference on the Constitution, Doctrine, Discipline, and History of the Catholic Church, vol. 1*. New York: The Universal Knowledge Foundation, Inc., pp. 714.

Howe, D., et al.. (2008). The future of biocuration: to thrive, the field that links biologists and their data urgently needs structure, recognition and support. *Nature, 455*(7209): 47-50.

Goldston, D. (2008). Data wrangling. *Nature, 455*(7209): 15.

Greenberg, J. (2008). The Dryad Repository: Metadata Research and Development [Presentation]. Library of Congress, December 19, 2008: https://www.nescent.org/wg/dryad/images/1/1b/DryadLC.pdf.

Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging and Classification Quarterly, 47*(3/4): 380-402.

Greenberg, J., Pattuelli, M. C., Parsia, B., & Robertson, W. D. (2001). Author-generated Dublin Core metadata for web resources: a baseline study in an organization. *Journal of Digital Information, 2*(2): http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/.

Greenberg, J. & Severiens, T. (2007). DCMI-tools: ontologies for digital application description. In *ELPUB2007. Openness in Digital Publishing: Awareness, Discovery and Access - Proceedings of the 11th International Conference on Electronic Publishing*, Vienna, Austria 13-15 June 2007, pp. 437-444.

Latour, B., & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton, N.J.: Princeton University Press.

Lord, P. and MacDonald, A., Lyon, L. & Giaretta, D. (2004). From data deluge to data curation. *3rd e-Science all hands meeting*, pp. 371-375.

McKeon, R. (1994). *On knowing :The natural sciences*. (D. Owen & Z.K. McKeon, Eds.). Chicago: University of Chicago.

Nilsson, M., Baker, T., & Johnston, P. (2008). The Singapore Framework for Dublin Core

application profiles: http://dublincore.org/documents/singapore-framework/.

Rice, R. (2008). Applying DC to institutional data repositories. In *International Conference on Dublin Core and Metadata Applications 2008*, September 22-26, 2008, Berlin, Germany, pp. 212.

Scherle, R., Carrier, S., Greenberg, J., Lapp, H., Thompson, A., Vision, T., & White, H. (2008). Building support for a discipline-based data repository. In *Proceedings of the 2008 International Conference on Open Repositories*: http://pubs.or08.ecs.soton.ac.uk/35/1/submission_177.pdf.

Szostak. R. (2004) *Classifying science: phenomena, data, theory, method, practice.* Norwell, MA: Springer.

Waldrop, M. (2008). Wikiomics. Big Data. *Nature, 455*(7209): 22-25.

Whewell, W. (1989). *Theory of scientific method*, (2nd ed.) . Ed. Robert E. Butts. Indianapolis: Hackett Publishing Company.

White, H.. (2008). Exploring evolutionary biologists' use and perceptions of semantic metadata for data curation. In *International Conference on Dublin Core and Metadata Applications 2008*, September 22-26, 2008, Berlin, Germany, pp. 202.

Witt, M. (2008). Institutional repositories and research data curation in a distributed environment. *Library Trends, 57*(2): 191-201.

Notes

1. DOE (Department of Energy) Data Explorer (DDE): http://www.osti.gov/dataexplorer/.
2. Knowledge Network for Biocomplexity Data (KNB): http://knb.ecoinformatics.org/.
3. The Dublin Core is comprised of both the 15 core properties from the DCMES Metadata Element Set (DCMES), Version 1.1: Reference Description:

http://dublincore.org/documents/2004/12/20/dces/, and a set of additional properties registered in the DCMI (Dublin Core Metadata Initiative) Metadata Terms namespace: http://dublincore.org/documents/dcmi-terms/.

4. Dublin Core Abstract Model (DCAM): http://dublincore.org/documents/abstract-model/.
5. Dublin Core Application Profile Guidelines: http://dublincore.org/usage/documents/profile-guidelines/.
6. Dryad Repository: http://www.datadryad.org/repo/.
7. Dryad Repository Partners: http://www.datadryad.org/repo/themes/Dryad/pages/partners.html.
8. Joint Data Archiving Policy: http://www.datadryad.org/repo/.
9. Interoperability Levels for Dublin Core Metadata: http://dublincore.org/documents/interoperability-levels/.
10. Dryad Workshop: https://www.nescent.org/wg_dryad/Dec_5_Workshop_Minutes.
11. Collectively the DCMES (http://dublincore.org/documents/2004/12/20/dces/) and DCMI Metadata Terms (http://dublincore.org/documents/dcmi-terms/), as explained in footnote 3.
12. Darwin Core (DwC), version 1.3: http://digir.sourceforge.net/schema/conceptual/darwin/core/2.0/darwincoreWithDiGIRv1.3.xsd; version 1.4 being reviewed, see: http://wiki.tdwg.org/twiki/bin/view/DarwinCore/DarwinCoreVersions.
13. Publishing Requirements for Industry Standard Metadata (PRISM): http://www.prismstandard.org/specifications/.
14. Journal Publishing Tag Set Tag Library, version 3.0, November 2008: http://dtd.nlm.nih.gov/publishing/tag-library/.
15. Data Document Initiative (DDI): http://webapp.icpsr.umich.edu/cocoon/DDI-LIBRARY/Version2-1.xsd?section=all.
16. Ecological Metadata Language (EML): http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html.
17. PREMIS Editorial Committee. PREMIS Data Dictionary for Preservation Metadata version 2.0, 2008: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.
18. Status Element—Dryad: http://www.purl.org/dryad/terms/status
19. Dryad Domain: http://www.purl.org/dryad

20. Text Encoding Initiative (TEI) Header, Chapter 2 (P5: Guidelines for Electronic Text Encoding and Interchange): http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html.

21. Tim Berners-Lee on the next Web (TED Conferences, LLC): http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html.

22. GenBank database: http://www.psc.edu/general/software/packages/genbank/genbank.php.

23. TreeBASE: http://www.treebase.org.

24. Long Term Ecological Research (LTER) Network's Metacat data catalog: http://metacat.lternet.edu/knb

25. Gleaning Resource Descriptions from Dialects of Languages (GRDDL): http://www.w3.org/TR/grddl-primer/