Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption

Jane Greenberg

ABSTRACT. The Dryad repository is for data supporting published research in the field of evolutionary biology and related disciplines. Dryad development team members seek a theoretical framework to aid communication about metadata issues and plans. This paper explores *lifecycle modeling as a theoretical framework for understanding metadata* in the repository environment. A background discussion reviews the importance of theory, the status of a metadata theory, and lifecycle concepts. An analysis draws examples from the Dryad repository demonstrating *automatic propagation*, *metadata inheritance*, and *value system adoption*, and reports results from a faceted term mapping experiment that included 12 vocabularies and approximately 600 terms. The paper also reports selected key findings from a recent survey on the data sharing attitudes and behaviors of nearly 400 evolutionary biologists. The results confirm the applicability of lifecycle modeling to Dryad's metadata infrastructure. The paper concludes that lifecycle modeling provides a theoretical framework that can enhance our understanding of metadata, aid communication about the topic of metadata in the repository environment, and potentially help sustain robust repository development.

KEYWORDS. Metadata, Metadata theory, Repositories, Dryad, Lifecycle modeling, Automatic metadata propagation, Metadata inheritance, Value system adoption

Jane Greenberg, MLIS, PhD, is a Francis Carroll McColl Term Professor, School of Information and Library Science, University of North Carolina at Chapel Hill, 205ManningHall, CB#3360, Chapel Hill, NC (E-mail: janeg@email.unc.edu).

Cataloging & Classification Quarterly, Vol. 47(3/4)
2009 Available online at http://ccq.haworthpress.com
© 2009 by The Haworth Press. All rights reserved.
doi:*xx*

INTRODUCTION

Metadata activities underlying the evolving digital repository environment reflect challenges that have been part of library cataloging operations for years. This is particularly evident via tensions associated with cataloging as a public good and the exploration of automatic techniques for cataloging/creating metadata. [1] Generating metadata for repository objects, similar to library cataloging/metadata, is motivated by the notion of a public good (1), (2), although administrators are generally pressed to demonstrate a positive return-on-investment (ROI) for production costs (3), (4). Increased collection access and use are positive outcomes, although it is truly difficult to measure long-term returns, such as a *significant* information discovery attributed to cataloging produced years or even decades earlier.

Repositories, like libraries, use automated techniques for mundane, repetitive, time-consuming metadata-related tasks, such as global updates for revised subject headings. Research employing automatic metadata generation in the repository environment has been promising (5), (6). However, as with libraries, automatic techniques can extend metadata quality and resource allocation challenges (7), (8), (9).

Given that libraries and repositories face similar challenges, both environments may benefit from similar solutions. Indeed, library approaches, such as the use of controlled vocabulary systems and name authority files, have been incorporated into repositories. There are, however, fundamental differences in how libraries and archives are conceived and operate, and these differences affect the metadata infrastructures in both environments.

Chief goal. *Repositories* are constructed with the chief goal of storage and preservation, and emphasize use/re-use. The *library's* overriding goal is to help users with their information needs via resource discovery, supported by a series of linked services (collection development, cataloging, reference, and circulation/access). Clearly, storing and preserving collection materials are part of library operations, although these activities are intended chiefly to serve users, rather than as ends in themselves. Additionally, libraries de-accession and replenish collections to keep services current, unlike repositories, which intend to keep resources indefinitely or may dispose of resources, following archival or recordkeeping practices.

Metadata types and function. *Repository* metadata is generated to support curation—a process requiring multiple types of metadata (e.g., descriptive, structural, provenance, and preservation metadata) and packaging metadata together in a wrapper, such as METS.[2] *Libraries* emphasize descriptive metadata to support resource discovery and access. Acquisition, preservation, circulation, and other types of management metadata are also generated, although, historically, not with the same deliberate attention as descriptive metadata/cataloging.

Who generates the metadata? Many *digital repositories* support or require metadata creation by the individual completing the deposition, as seen with arXiv.org, the Open Archives Initiative repository for off-prints in physics, mathematics, computer science, quantitative biology and statistics.[3] Professionals may perform some degree metadata curation after authors initially create metadata, as viewed with the Virtual Observatory metadata repository (10). *Libraries* more tightly control metadata generation, and trained information professionals generally perform and oversee this task.

Admittedly, the distinction between libraries and repositories is fuzzy at times; and, indeed, the terms are often used interchangeably. The above noted differences do, however, provide

context for why library approaches are not fully suited for addressing repository metadata challenges. This predicament has become obvious during the development of the Dryad repository,[4] which focuses on data underlying published works in the field of evolutionary biology and related fields.

Dryad, named after tree nymphs in Greek mythology, has a metadata plan emphasizing the use of automatic techniques as much as possible without compromising metadata quality. Administratively, there is a need to demonstrate a positive ROI for metadata creation costs and to keep the barriers low so as to not discourage author metadata contributions. Although Dryad's high-level metadata goals are clear, the theoretical context and language aiding communication among team members about metadata issues and potential solutions are limited.

This paper considers Dryad's challenges and the need for a theoretical context to aid communication about metadata. The paper reviews theory in general, metadata theory, and lifecycle modeling concepts. An analysis illustrating automatic propagation, metadata inheritance, and value system adoption is presented; this work if further supported by initial data from a faceted term mapping experiment involving multiple vocabularies, and results from a recent survey on data sharing attitudes and behaviors of evolutionary biologists. The goal is to better understand metadata generation and replication for Dryad, and aid the development of a robust and functional repository.

THEORY AND FRAMEWORKS

Theory is often defined as a set of arguments or axioms supported by evidence (11, 33-36). Scientific theories are supported by research using scientific methods; that is, gathering and analyzing data, and presenting empirical proof (12). The term theory is also used more loosely for speculation, observation, and sense making. We often turn to this latter use of theory to help us understand phenomena, guide our actions, or lay groundwork for ideas to be

validated.

Speculation may be presented in the context of a *theoretical framework*—a collection of interrelated themes or concepts often involving metaphor (13) or more specifically metaphoric frames (14). Popular examples are found in describing the World Wide Web (web) through *transportation* and *nautical* themes, particularly during its emergence: The former includes the concepts "information highway" and "infrastructure," and the latter defines the web as a "vast ocean" and uses "surfing" to connote searching. Even the word "web" provides a metaphorical framework, prompting one to envisage a spider web with computer nodes connected via a wired (and now wireless) network of threads.

Metaphors, as illustrated with the web, help to make sense of phenomena. On a research level, they can guide in identifying data requiring analysis for advancing knowledge of the targeted phenomena. Given the need to advance our understanding of metadata in general, and within the repository environment in particular, it seems prudent to consider themes and metaphors, and even to speculate on a theoretical level. This paper takes a step in this direction for the Dryad repository by first reviewing temporal factors and advances toward a metadata theory.

METADATA: TEMPORAL FACTORS AND MOVING TOWARDS A THEORY

There does not seem to be a unified, cohesive theory of metadata, let alone a well-defined theoretical proposition to guide study of this concept. A key factor here is *time*—both the limited expanse of time that has passed since the emergence of metadata as a concept; and the limited daily time individuals have to study metadata. For measure, today's modern library is tied to the Renaissance time-period (15), and scholars and librarians have had more than half a millennium to theorize about cataloging and the modern notion of bibliographic control (16), (17). In comparison, the launching of the web during the mid-1990s has presented less

than a quarter of a century (less than a life-time) for developing a theory for metadata in the digital world. Additionally, the web's limited time-span has been fraught with an extreme need for developing metadata schemes and generating metadata for sustaining day-to-day operations relating to digital resources, leaving little time to develop discussion and debate leading to a cohesive theory of metadata. One only needs to view metadata topics in the professional and scholarly literature to see that scheme development and metadata generation issues are prevalent, and limited attention is directed toward a metadata theory.

Time factors aside, there has been tremendous progress in our understanding of metadata on many levels. Authors exploring the theoretical underpinnings of metadata emphasize a link to cataloging foundations (18), (19), including Cutter's objects of the library catalog (20). Closely related are metadata principles, generally recorded in introductory documentation for metadata schemes (EAD[5] and TEI[6]), or tied to a scheme in some way (21). Additionally, advancements have been made defining metadata as structured data supporting functions (22), (23), rather than simply and abstractly as data about data; and a significant amount of attention has been directed to defining different types or classes of metadata (24), (25). Another important area of growth is metadata quality (26), which is tied to notions of best practices (27). Some of most important advancements have been made via the articulation of models that shape our conception of metadata, such the Resource Description Framework (RDF)[7] together with the emerging idea of a metadata ecosystem in the context of the Semantic Web (28); the Dublin Core Abstract Model (DCAM),[8] which incorporates FRBR (29); and models, such as such as METS, for wrapping metadata to document the digital object's *lifecycle*. All of these developments are important to advancing our understanding of metadata in general and within the repository environment. Most important to the work presented in this paper is modeling involving the notion of *lifecycle* because it provides a way to view metadata as mutable, sharable, and reusable.

LIFECYCLE MODELING: A FRAMEWORK FOR METADATA RECORDS IN DIGITAL REPOSITORIES

Specific to repositories, Higgins (30) defines the lifecycle model for the Digital Curation Centre (DCC), University of Edinburgh, as an approach intending to support the "successful curation and preservation from initial conceptualization [of the material] to either disposal, or selection for reuse and long-term preservation. (31), (31). Lifecycle is, traditionally, a biological concept for referencing the stages through which living organisms progress—fertilization, maturity, and death. The concept of lifecycle has been co-opted for describing processes and activities that are not biological, such as product development and project management. More closely related to repositories, computer science has embraced lifecycle concepts by using words such as inheritance and propagation for explaining computational functions (32). Also germane is the presentation of a *metadata lifecycle* (33), (34), although existing work focuses primarily on metadata infrastructure and metadata layers within an information system, rather than the "life" of the metadata and the associated resource.

Repository literature increasingly emphasizes digital lifecycle management, and the notion of preservation management over the course of the resource's life (35), (36). A key reason for using lifecycle concepts for repositories is that digital resources are more mutable and sharable than their physical printed counterparts; and the mutable nature presents a seemingly *organic* object, despite the absence of genetic material.

This mutable and organic nature of the object is reflected in the resource's associated metadata as well, although the notion of metadata as an organic object is generally left out of current discussions. A metadata record is a document, a resource, or an object, and it too has a lifecycle. And, like the digital resources, metadata—in digital form—is more mutable and sharable than traditional cataloging records printed for library card catalogs, or maintained in closed databases.

Repositories like Dryad seek to facilitate data reuse by interoperating with and sharing resources with other repositories;

and part of this process includes sharing metadata that is coupled with the data object. Drawing from biological concepts, the reuse of a resource may result in the creation of an offspring or progeny, connoting a sense of propagation, stemming from an initial progenitor. Significant to these ideas is research advancing our understanding of the nature of a *work* (37), (38), (39); bibliographic relationships (40), (41); and Smiraglia's insightful work on *instantiation* as a "phenomenon of information objects"(42). Relating to science is Coleman's (43) analysis showing the application of *bibliographic families* for scientific models; her work emphasizes the centrality of a progenitor source, drawing heavily from Smiraglia's taxonomy of derivative relationships (44). The growing body of research on the nature of a work, along with the high-level mutability of any digital object—including *metadata representations*—encourages the exploration lifecycle modeling as a framework for understanding metadata representing digital data objects. This paper seeks to move these related ideas forward by exploring the questions outlined directly below.

CONTEXT AND QUESTIONS

As previously indicated, Dryad's high-level metadata goals are clear to the development team, although the language, and in many respects a theoretical framework, for communicating issues and potential solutions are limited. Dryad team members seek to communicate more deeply about the value and centrality of the metadata record, and issues related to metadata generation replication, and revision. One approach to moving forward is to ground discussion and ideas within a theoretical framework. This paper takes a step in this direction by first assessing Dryad's context more fully, and then by presenting several examples and initial analyses in response to the following questions:

1. Can a metadata record for a published research article serve as a source for propagating metadata for data objects represented in the published research?

2. Given propagation, do metadata records, as seemingly *organic* progenies, inherit characteristics from their original source and adopt values from outside systems?

THE DRYAD REPOSITORY

Dryad, as already stated, is a repository for scientific data objects supporting published research in the field of evolutionary biology and related disciplines. The repository was initiated via a collaboration involving the National Evolutionary Synthesis Center (NESCent)[9] and the School of Information and Library Science, Metadata Research Center (SILS/MRC), University of North Carolina at Chapel Hill (UNC/CH).[10] The partnership has recently expanded, via support from the National Science Foundation, to include North Carolina State University, University of New Mexico, and Yale University.[11] Additional partners include major societies and journals in the field of evolutionary biology.[12]

Dryad was initiated in response to a "crisis of data attrition[13] in the field of evolutionary biology. Evolutionary biology is an interdisciplinary field that includes ecology, genomics, paleontology, population genetics, physiology, systematics and other related disciplines. Evolutionary biologists rely on several specialized databases for preserving the most common data types (such as DNA sequences, character state matrices, and phylogenetic trees). Despite these developments, data associated with a published research is extremely difficult to find and reuse. Even supplementary data stored in journal-specific repositories is difficult to discover, due to sparsely populated metadata that lacks quality control. Journal-specific repositories index data files by author(s), although they generally omit subject/keywords, geographic descriptors, or species names—all of which are significant for discovering, examining, using and reusing existing data. Finally, specific to metadata and lifecycle issues, obvious problems include limited linking among data instantiations and costly reproduction of metadata that could easily be automatically replicated and revised.

Metadata representing the lifecycle of a phylogenetic tree deposited in TreeBASE[14] or a gene sequence in GenBank[15] is not sufficiently linked to the published research recording this information. As a result, metadata for certain entities is generated multiple times in different contexts, when it could be automatically replicated and updated in a more efficient and less costly manner. Most problematic is that current practices fail to include contextual metadata that is important for fully and accurately interpreting the data object's intrinsic value, and can potentially have a negative impact on the science.

Dryad is addressing the above problems via the provision of a one-stop data deposition and access service for data underlying published research in evolutionary biology and related disciplines. Repository development was initiated toward the end of 2007, following a year-long planning cycle and the formalization of functional requirements (45). Dryad has been implemented within the DSpace framework (46); the repository contains over 125 data objects from 33 publications, although it is still very much in a development cycle. As the initiative progresses, researchers publishing in major journals in the field of evolutionary biology will be required to deposit data in Dryad. The data underlying the published research will be deposited *as part* of the publication process. Dryad will also establish a "handshaking" mechanism linking data deposition with GenBank, TreeBASE, and other specialized databases over time, so that researchers can automatically port their data objects and associated metadata to the appropriate specialized databases during a single deposition activity. Dryad also aims to support one-stop shopping for scientists seeking access to data supporting published research, and access to data objects housed in specialized repositories like GenBank and TreeBASE, via sufficient metadata.

Dryad's development team has been making good progress, and the accomplishments thus far have been encouraging. The progress has motivated the development team to think more deeply about metadata and target where automatic metadata generation, replication, and revision can improve workflow, promote linking among data object instantiations, and preserve important contextual

information. At this stage in the Dryad development cycle, it is difficult to conduct an extensive empirical analysis; however, through focused analyses of selected metadata records, we can begin to consider and deliberate about the questions posited above and the applicability of a metadata theory grounded in lifecycle modeling for Dryad and other like repositories.

EXPLORING THE QUESTIONS

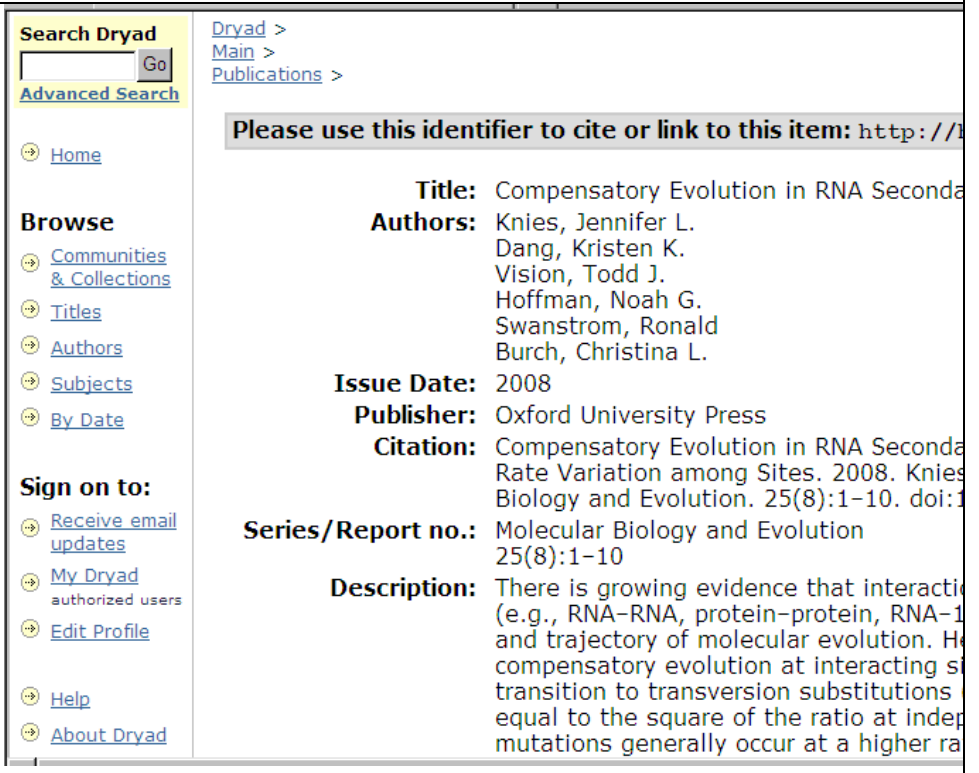
Research question one

Question 1: Can a metadata record for a published research article serve as a source for propagating metadata for data objects represented in the published research?

A metadata record for a published research article can serve as a source for propagating metadata for data objects represented in the article. Published research is an artifact generated from analyzing, abstracting, and contextualizing data. This is a simplified view, as data may need to be coded, re-coded, synthesized, merged with other data and so forth; but, fundamentally, the output from these efforts are presented in published research for consumption. It is blatantly obvious that published research—as *output*—is intimately linked to the underlying data—the *input*. It also seems credible that the "data objects" are progenitors of the published research, although discussion of this is beyond the scope of the metadata questions explored in this paper. Specific to question one, the scientist, as a research article author, is also the creator or a contributor to the data object from which published results have been derived. As a result, author metadata for the published research article, containing the published research results (e.g., data facts, a table, diagram, etc.), can be automatically propagated for representing the data objects. Examples 1A, 1B-1, and 1C-1 below are Dryad metadata records

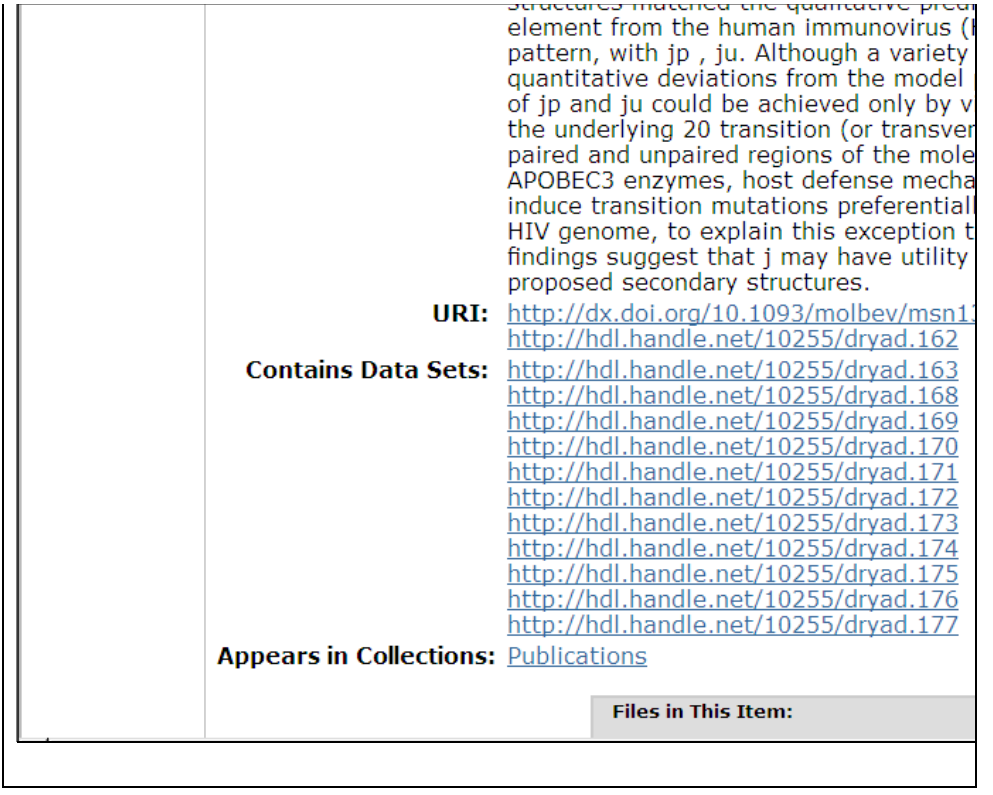
illustrating a metadata propagation sequence executed via automatic processes. Example 1A is a metadata record for an article authored by Knies, et al, (47) published in *Molecular Biology and Evolution.* This metadata record (1A) is used to propagate metadata for the eleven data objects, following the label "Contains Data Sets" and linked via Dryad-specific handles ending with the numbers 163-177. The author metadata in 1A (the names of six authors presented as: surname, forename; with five cases including a middle initial) is automatically propagated to the author field for the data object metadata records presented in 1B-1 and 1C-1; and this procedure follows for all eleven datasets, with the handles numbering 163-177. Examples 1B-2 and 1C-2 present a portion of the actual data objects respectively corresponding metadata records 1B-1 and 1C-1.

Example 1A: Metadata record for a journal article in *Molecular Biology and Evolution*. Note that there are 11 datasets associated with this publication, represented with handles ending with the number 163-177.



[Some abstract text has been omitted from the screen capture.]

Example 1A (continued)


element from the human immunovirus (H
pattern, with jp , ju. Although a variety
quantitative deviations from the model
of jp and ju could be achieved only by v
the underlying 20 transition (or transver
paired and unpaired regions of the mole
APOBEC3 enzymes, host defense mecha
induce transition mutations preferentiall
HIV genome, to explain this exception t
findings suggest that j may have utility
proposed secondary structures.

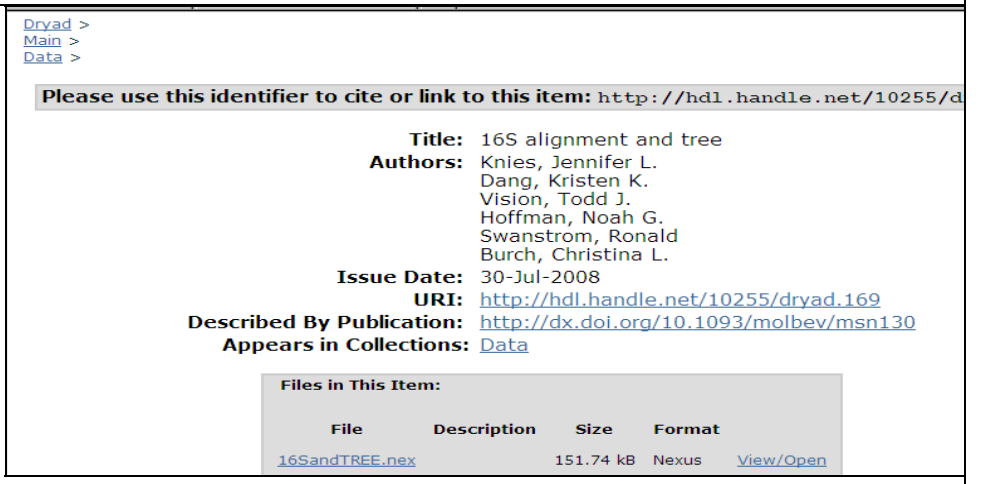
URI: http://dx.doi.org/10.1093/molbev/msn1
http://hdl.handle.net/10255/dryad.162

Contains Data Sets: http://hdl.handle.net/10255/dryad.163
http://hdl.handle.net/10255/dryad.168
http://hdl.handle.net/10255/dryad.169
http://hdl.handle.net/10255/dryad.170
http://hdl.handle.net/10255/dryad.171
http://hdl.handle.net/10255/dryad.172
http://hdl.handle.net/10255/dryad.173
http://hdl.handle.net/10255/dryad.174
http://hdl.handle.net/10255/dryad.175
http://hdl.handle.net/10255/dryad.176
http://hdl.handle.net/10255/dryad.177

Appears in Collections: Publications

Files in This Item:

Example 1B-1: Metadata record for the data set with the handle: http://hdl.handle.net/10255/dryad.169.



Dryad >
Main >
Data >

Please use this identifier to cite or link to this item: http://hdl.handle.net/10255/d

Title: 16S alignment and tree
Authors: Knies, Jennifer L.
Dang, Kristen K.
Vision, Todd J.
Hoffman, Noah G.
Swanstrom, Ronald
Burch, Christina L.
Issue Date: 30-Jul-2008
URI: http://hdl.handle.net/10255/dryad.169
Described By Publication: http://dx.doi.org/10.1093/molbev/msn130
Appears in Collections: Data

Files in This Item:

File	Description	Size	Format	
16SandTREE.nex		151.74 kB	Nexus	View/Open

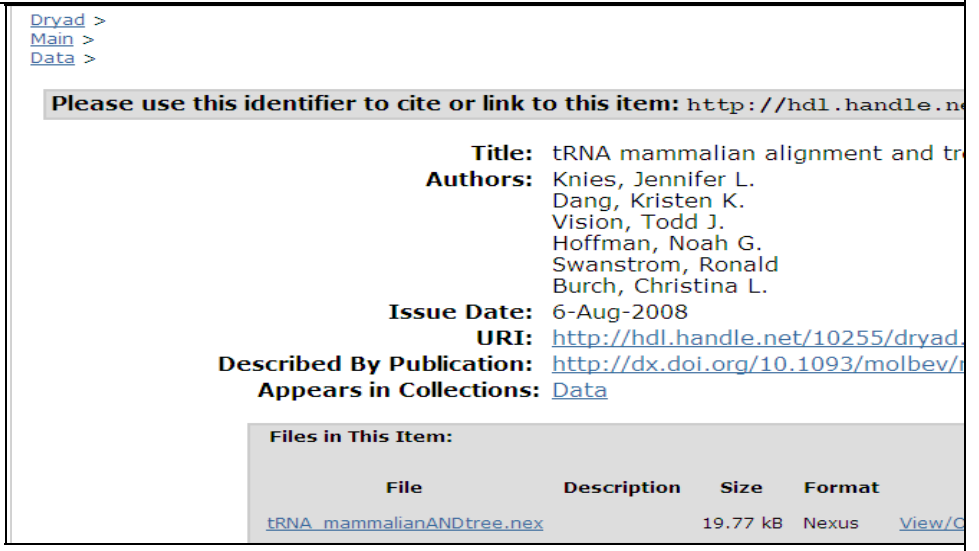
Example 1B-2: A segment of the data set with the handle http://hdl.handle.net/10255/dryad.169.

Scientific description: Aligned RNA sequence data, in NEXUS format, used as input for evolutionary analysis of bacterial 16S ribosomal RNA molecules. Data were derived from the Comparative Analysis of RNA website (http://www.rna.ccbb.utexas.edu/) and aligned relative to the *E. coli* reference sequence.

```
#NEXUS
BEGIN DATA;
  DIMENSIONS NTAX=94 NCHAR=1542;
  FORMAT interleave=no DATATYPE=DNA missing=N gap=-;
  MATRIX

00001Escherichia_coliREFERENCE                      AAAUUGAAGAGUUUGAUCAUGGCUCA
00024Deinococcus_radiodurans_R1AE000513             UUUAUGGAGAGUUUGAUCCUGGCUCA
00039Bacteroides_fragilisAP006841                   ACAAUGAAGAGUUUGAUCCUGGCUCA
00061Planctomyces_limnophilusX62911                 --------GAGUUUGAUCCUGGCUCA
00085Chlamydia_trachomatisDQ019310                  ----CUGAGAAUUUGAUCUUGGUUCA
00105Chlamydophila_pneumoniae_AR39AE002161          UUUUCUGAGAAUUUGAUCUUAGUUCA
00121Spirulina_platensisDQ279771                    -------AGAGUUUGAUCCUGGCUCA
00143Anabaena_flos_aquaeDQ234825                    -------AGAGUUUGAUCCUGGCUCA
00156Synechococcus_spX03538                         AAAAUGGAGAGUUUGAUCCUGGCUCA
00169Borrelia___SCGT_10AF467970                     -------AGAGUUUGAUCCUGGCUUA
00183Borrelia_garinii_PBiCP000013                   AUAACGAAGAGUUUGAUCCUGGCUUA
00203Acetobacter_intermediusAJ012697                --------GAGUUUGAUCNUGGCUCA
00213Gluconacetobacter_azotocaptansAY958232         ------UAGAGUUUGAUCCUGGCUCA
00224Rickettsia_prowazekiiAJ235272                  AAACUUGAGAGUUUGAUCCUGGCUCA
00235Rhodobacter_sphaeroidesX53855                  CAACUUGAGAGUUUGAUCCUGGCUCA
00251Bradyrhizobium_japonicumD11345g464202          AACUUGAAGAGUUUGAUCCUGGCUCA
00261Bradyrhizobium_japonicumDQ133342               -------AGAGUUUGAUCCUGGCUCA
00273Nitrobacter_hamburgensis_X14CP000319           CAACUUGAGAGUUUGAUCCUGGCUCA
00284Rhodopseudomonas_palustrisAF184625             CAACUUGAGAGUUUGAUCCUGGCUCA
00299Bradyrhizobium___ISLU207AJ558028               -------AGAGUUUGAUCCUGGCUCA
00311Rhizobium_leguminosarumD14513g757511           -AACUUGAGAGUUUGAUCCUGGCUCA
00322Rhizobium_leguminosarum_bv_trifoliiDQ196416    ------UAGAGUUUGAUCCUGGCUCA
00339Agrobacterium_vitisU28505g1143912__bases_800_  CAACUUGAGAGUUUGAUCCUGGCUCA
00353Bartonella_quintanaM11927                      CAAUAUGAGAGUUUGAUCCUGGCUCA
00365Sinorhizobium___ORS3178AY875975                ------UAGAGUUUGAUCCUGGCUCA
```

Example 1C-1: Metadata record for the data set with the handle: http://hdl.handle.net/10255/dryad.177.



Dryad >
Main >
Data >

Please use this identifier to cite or link to this item: http://hdl.handle.n

Title: tRNA mammalian alignment and tr
Authors: Knies, Jennifer L.
Dang, Kristen K.
Vision, Todd J.
Hoffman, Noah G.
Swanstrom, Ronald
Burch, Christina L.
Issue Date: 6-Aug-2008
URI: http://hdl.handle.net/10255/dryad.
Described By Publication: http://dx.doi.org/10.1093/molbev/
Appears in Collections: Data

Files in This Item:

File	Description	Size	Format	
tRNA_mammalianANDtree.nex		19.77 kB	Nexus	View/O

Example 1C-2: A segment of the data set with the handle http://hdl.handle.net/10255/dryad.177.

Scientific description: Aligned DNA sequence data, in NEXUS format, used as input for evolutionary analysis of mammalian mitochondrial tRNA genes (a concatenated alignment of tRNAs for Ala, Cys, Glu, Asn, Gln, and Tyr). The phylogenetic tree for these sequences is also reported later in this file. Unaligned sequences were obtained from the Organellar Genome Retrieval System (http://drake.physics.mcmaster.ca/ogre/).

```
#NEXUS

[
Generated by HYPHY 0.9920060327beta(MP) for MacOS(Carbon) on Fri Aug  3 11:11:55

]

BEGIN TAXA;
        DIMENSIONS NTAX = 40;
        TAXLABELS
                'LAMPACMIT' 'BALMUSMIT' 'BOSGRUMIT' 'EUBAUSMIT' 'MUNMUNMIT' 'MUN
END;

BEGIN CHARACTERS;
        DIMENSIONS NCHAR = 419;
        FORMAT
                DATATYPE = DNA

                GAP=-
                MISSING=?
        ;

MATRIX
        'LAMPACMIT'  AAGGGCTTAGCTTAATTAAAGTAGTTGATTTGCATTCAATTGATGTAGGATAGAGTCTT
        'BALMUSMIT'  GAGGATTTAGCTTAATTAAAGTGTTTGATTTGCATTCAATTGATGTAAGATATAGTCTT
        'BOSGRUMIT'  GAGGATTTAGCTTAATTAAAGTGGTTGATTTGCATTCAATTGATGTAAGGTGTAGTCTT
        'EUBAUSMIT'  GAGGATTTAGCTTAATTAAAGTGTTTGATTTGCATTCAATTGATGTAAGATATAGTCTT
        'MUNMUNMIT'  GAGGATTTAGCTTAATTAAAGTGGTTGATTTGCATTCAATTGATGTAAGATATGGTCTT
        'MUNREEMIT'  GAGGATTTAGCTTAATTAAAGTGGTTGATTTGCATTCAATTGATGTAAGATATGGTCTT
        'SUSSCRMIT'  GAGGACTTAGCTTAATTAAAGTGTTTGATTTGCATTCAATTGATGTAGGATA-AGTCCT
        'CAPHIRMIT'  AAGGATTTAGCTTAATTAAAGTGGTTGATTTGCATTCAATTGATGTAAGATATGGTCTT
```

In a recent Dryad development team meeting, a question was raised about a hypothetical scenario in which a research article presents data or results produced by a scientist who is not credited as an article co-author. The proposed scenario was not about referencing results *published elsewhere*, but actually presenting data or results of an analysis as *key content.* The response from scientists on the Dryad development team is that such practice is not ethical scholarly behavior. If a group of researchers desire to publish data generated by someone not in their research group, they need to first consult with the data owner, and the protocol is that the data owner is included as one of the authors of the published research. In other words, published research needs to credit the authors of all the data presented in the publication, simply because without the data, that is obviously integral to the publication, the research could not be published.

Currently, exceptions to this protocol seem rare. However, the passing-away of scientists is real, as is interest in using and publishing data from someone deceased. In such cases, it is possible that the deceased scientist may not be credited as a publication author, given that the researcher is unable to review and approve the presentation of the data; however, the deceased would be appropriately cited and acknowledged. A potential scenario is that another scientist(s), or an institutional entity, has access and ownership rights to the data of the deceased, and is invited to join in the research publication as a co-author, with the appropriate acknowledgments still given. An interesting case is the Dryad dataset containing finch measurements collected by Darwin. Although Darwin is mentioned in the description, he is not listed as an author in Dryad's current metadata record, presenting a situation requiring curatorial attention and development team discussion. There field of evolutionary biology and other sciences have not had to grapple with this circumstance much, simply because data from deceased researchers has not been made easily discoverable or accessible for data reuse. It seems likely that the use and reuse of data from scientists no longer living will become more common practice, over time, along with scientists' increasing expectation

with respect to long-term data access; and new rights management protocols will need to be established to govern this situation.

Not shown in these examples (1A, 1B-1, and 1C-1) is subject/keyword metadata, although Dryad is working on means for automatically propagating subject metadata by harvesting keywords designated by the author(s) and extracting subject content from the abstract. The Dryad metadata application profile includes a subject/keyword property for two modules: module one for journal article metadata and module two for data object metadata (datasets) (48), (49), (see also Table 1 below). In addition to high-level subject/keyword metadata, the data object module also includes three more granular subject metadata properties: spatial coverage, temporal coverage, and species. Dryad's current subject index contains 110 keywords, generated from manual cataloging. Dryad is launching the Helping Interdisciplinary Vocabulary Engineering (HIVE) initiative for automatically generating metadata; mapping the subject metadata to existing controlled vocabularies, such as the *National Biological Information Infrastructure Thesaurus*, and the *Library of Congress Subject Headings*; and populating the subject metadata field, including the data object's more granular subject metadata fields (e.g., species), with controlled vocabulary to aid information retrieval.[16]

Research question two

Question 2: Given propagation, do metadata records, as seemingly organic progenies, inherit characteristics from their original source and adopt values from outside systems?

Given propagation, metadata records, as seemingly organic progenies, do inherit characteristics from their original source and adopt values from outside systems. As mapped out above, the metadata relationships between 1A and 1B-1 and between 1A and 1B-2 show that metadata is inherited from the journal article metadata for each of the data objects. The metadata found in 1A is sequentially inherited as the metadata records for each data set is

generated. Table 1 provides top-level summary of Dryad's application profile, which includes two modules (the bibliographic citation and the data object). Specifically property 17, column 3, "dcterms:isPartOf/DOI for Published Article," directly inherits the metadata value from the published article's metadata property 13 in column 1, "dcterms:identifier/DOI (document object identifier) from Published Article." Another instance of inheritance is visible via author metadata propagation, reviewed with question one. However, while *all author names* for the article metadata are first, automatically propagated for representing each data object, it is possible that not every article author is an author for each data object underlying the publication. A contributor (or a curator) can then verify the property values that are definitely inherited, and delete metadata that is not quite be an innate characteristic. A mechanism is being put in place to accurately control when inheritance does not occur. This will also allow for the addition of data creator names when they may possibly not be an author, as discussed above in the context of a deceased scientist. Archival and collection/item oriented metadata developments can further inform Dryad's work in this area, particularly Renear, et al's (50) work exploring modal notions and first-order logic formulations focusing on *attribute/value-propagation*, *value-propagation*, and *value-constraints.*

Table 1: Dryad Application Profile, Version 1.1 (October 2008)

KEY/Namespaces: dc=Dublin Core DCMES; dcterms=Dublin Core terms; Darwin=Darwin Core; DDI=Data Document Initiative; PREMIS= PREservation Metadata Implementation Strategies

Obligation: R=required, O=Optional

Module 1: Bibliographic Citation Metadata		**Module 2: Data Object Metadata**	
Namespace: Name/label	**Generation method/ obligation**	**Namespace: Name/label**	**Generation method/ Obligation**
1. dc:type/Type	automaticDefault/R	1. dc:type/Type	automaticDefault/R
2. dc:creator/Author	automaticCitation/R	2. dc:creator/Author	automaticCitation/R
3. dc:contributor/Coauthor	automaticCitation/R	3. dc:contributor/Coauthor	automaticCitation/R
		4. DDI:depositr/Depositor	automaticDerived/R
		5. DDI:contact/Contact Information	automaticDerived/R
4. dc:title/Title of Article	automaticCitation/R		
5. dcterms:abstract/ Abstract	automaticContent/R	6. dc:description/Description	manual/O (semi-automatic in the future)
6. dc:subject/Keyword	semi-automaticContent/R	7. dc:subject/Keyword	semi-automaticContent/R
		8. dcterms:spatial/Locality	semi-automaticContent/O
		9. dcterms:temporal/Date Range	semi-automaticContent/O
		10. Darwin Core:SpecificEpithet/ Species/Scientific Name	semi-automaticContent/O
7. dc:publisher/Publisher	automaticDefault/R		
8. dcterms:issued/Date of Publication	automaticCitation/R	11. dcterms:issued/ Date of Issue	automaticDerived/R
		12. dcterms:modified /Date Modified	automaticDerived/R

9. dc:rights/Rights Statement	automaticDefault/R	13. dc:rights/Rights Statement	automaticDefault/R
		14. dcterms:available/Embargo Date	automaticDefault/O
10. dc:language/Language	automaticContent/R	15. dc:language/Language	automaticContent/R
11. dcterms:bibliographic Citation/Citation	Automatic[Publisher data]/R		
12. dcterms:hasPartOf/ Dataset Identifier	automaticLinking/R	16. dc:identifier/Data set Identifier	automaticLinking/R
13. dcterms:identifier/ DOI for Published Article	automaticLinking/R	17. dcterms:isPartOf/DOI for Published Article	automaticLinking/R
		18. EML:software/Software	Manual/O
		19. dcterms:format/File Format	automaticDerived/R
		20. PREMIS:fixity/(hidden)	automaticDerived/R

Another aspect of lifecycle is that of *value system adoption*. Living organisms are exposed to value systems and behavioral norms from birth. Humans adopt religious and cultural values, and they may reject values that were encouraged during their youth. It is probably safe to say that all living species adopt values or behave a certain way due to environmental conditions. Reebs (51) has written extensively about the different behaviors fish exhibit when placed in aquariums and in the wild. Although there is a fuzzy line between *nature* and *nurture*, Reebs contends that there are adopted behaviors, such as the *actualizing means* or *degree of aggressive behavior* based on the environs and the other fish present.

The analogy to consider is that *metadata records are further shaped by adopting values from standardized value systems*, such as subject-controlled vocabularies and name authority files. Librarians and information scientists construct and maintain vocabularies and authority files to provide standard, accurate means for representing a

topic or name; to allow for interoperability among systems; and to facilitate retrieval and collocation. Table 1 gives an idea of where value system adoption may be relevant for Dryad metadata. Most obvious is the *bibliographic citation module property* "dc:subject/Keyword" (property number 6, column 1), and the *data object properties* "dc:subject/Keyword," "dcterms:spatial/Locality," "dcterms:temporal/Date Range," and "Darwin Core:SpecificEpithet/ Species/Scientific Name," (properties 7-10, column 3). Values for these properties can be generated via automatic means, extracting content for the journal article abstract and by harvesting author assigned keywords. Additional sources for automatic metadata generation include textual descriptions, in a journal article, immediately preceding or following a table, diagrams, or other data presentations, and associated labels.

As briefly noted above, Dryad plans to harvest keywords assigned by authors, and use automatic means to assign standardized terms from topical, geographical, taxonomic (in the scientific sense) controlled vocabularies and ontologies. Authors (and curators at times) will then confirm the accuracy of the suggested controlled vocabulary terms—making them adopted values. Table 1 indicates a semi-automatic metadata generation approach for this group of metadata properties, given the initial automatic process, followed by a secondary human evaluation/modification activity.

To address the vocabulary issue more thoroughly, a study was conducted assessing the applicability of selected vocabulary systems for Dryad. The overriding goal was to identify specific vocabularies for representing data objects stored in Dryad. A sample of approximately 600 keywords was collected from 104 articles appearing in several Dryad partner journals (*American Naturalist*, January 2007; *Molecular Biology and Evolution*; January 2007, *Systematics Biology*, October 2006; *Molecular Ecology*, January 2007; and *Evolution*, January 2007). The terms were categorized into the following nine facets: *topic*, *research method*, *geographic location*, *taxon*, *personal name*, *agency name*, *anatomical aspect*, *discipline*, and *habitat*. Terms from each facet were searched in the appropriate controlled vocabulary and ontological sources (e.g., *ERIC Thesaurus*, *NBII Thesaurus*, *Medical Subject Headings*

(MeSH), *LCSH*, *Getty Thesaurus of Geographic Names* (TGN), *Gene Ontology* (GO), *Integrated Taxonomic Information System* (ITIS)—to name a few). A total of 12 vocabularies were considered, although terms in each facet were not searched in each vocabulary. For example, terms in the "research methods" facets were searched in the *ERIC Thesaurus*, *NBII Thesaurus*, *MeSH*, and *LCSH*, but not in the *TGN*, *GO*, or *ITIS*, given that research methods terms are not contained in these latter vocabularies. A part of the analysis has measured:

Exact matches: An exact match between the *keyword* assigned by the author to either a *preferred term* (an authorized term) and *nonpreferred term* (a "see from" term preceded by UF (use for) reference or similar notation).

Partial and non-matches: A keyword that either partially matches a *preferred term* or *nonpreferred term*, or that does not match any thesaurus/controlled vocabulary term.

Initial results show that more keywords fall into the latter category of partial and non-matches. For example, of the 431 terms searched in the *NBII Thesaurus*, 25% of the terms were exact matches, while 75% were partial and non-matches. The same 431 terms were searched in *MeSH*, and only 18% of the terms were exact matches and 82% were partial and non-matches. A total of 531 terms were searched in *LCSH*, with 22% found to be exact matches and 78% partial and non-matches.

The terms in the first category "exact match," mapping directly to a preferred term, can be automatically adopted and placed in the data object's metadata record, given that they are already present in the standardized vocabulary; and if the exact match is for a nonpreferred term, an algorithm can select the referenced preferred term for the metadata record. With the latter category of "partial and non-matches," alternative terms will be selected via the use of more complex indexing algorithm (e.g., use of noun pairs and term proximity), demonstrating a less direct approach to *value system*

adoption. For example, the term "Phylogenetic inference," assigned by an author, is not in *LCSH*, although this concept may be represented via *LCSH* with the term "Cladistic analysis" and the precoordinated heading "Inference—Computer simulation." This is because "Phylogenetic systematics" is a *LCSH* nonpreferred term, referencing "Cladistic analysis" as a preferred term; and, additionally "Inference—Computer simulation" is a preferred *LCSH* term. The value system adoption example given also demonstrates the principle of coextensivity, (52); that is multiple controlled vocabulary terms may need to be adopted for accurately representing the intellectual content of the work—and in this case a single keywords originally assigned by an author. The use of controlled vocabularies will afford Dryad users with known discovery advantages, such as greater recall. Although, Dryad developers need to consider ROI and will consider rules, such as those developed by Losee (53), to determine the optimal number terms that should be adopted (assigned to a metadata record) in order to support robust searching and maximize retrieval results.

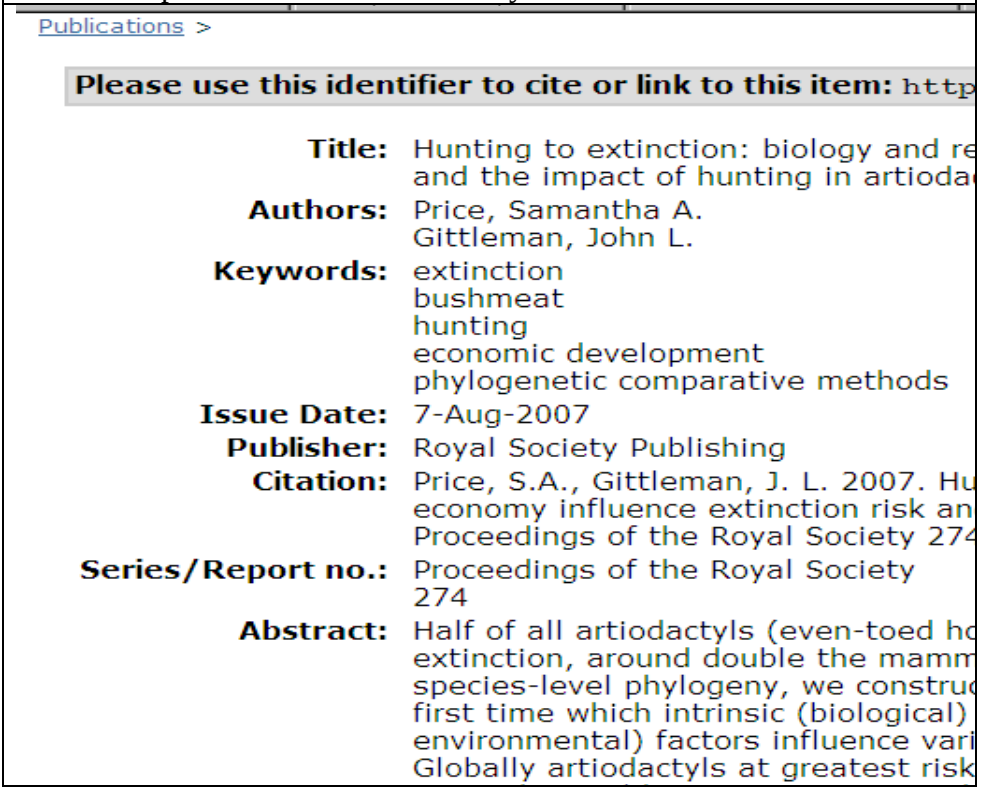
DISCUSSION: EXPLORING THE METADATA LIFECYCLE

Digital information allows for the flow and reuse of information in unprecedented ways. The above examples demonstrate Dryad metadata generation approaches via automatic propagation, metadata inheritance, and value system adoption from outside systems (standardized vocabularies). Although the Dryad development team is in an early stage of implementation, the examples and results presented encourage further exploration of a metadata theory that is grounded in lifecycle modeling.

In moving forward, it seems that exploring lifecycle modeling can extend beyond the immediate examples, and may aid in addressing other Dryad metadata challenges. As indicated in the overview of the Dryad repository, the same metadata is generated multiple times in different contexts, making for a costly process. For example, metadata is duplicated for a phylogenetic tree[17] published (or partially published) in a research journal, deposited as

supplementary data in a journal repository, and deposited in TreeBASE. Metadata duplication problems are further confounded by the failure to adequately record object reuse and modifications. Examples 2A and 2B help to demonstrate these points. Example 2A is for a journal article with the Dryad handle http://hdl.handle.net/10255/dryad.82, which has both an Excel data file with the handle http://hdl.handle.net/10255/dryad.83, and an associated phylogenetic tree in TreeBASE, represented by Example 2B, Accession #S1188. The journal article metadata (2A) does not link to the phylogenetic tree in TreeBASE, although the next version of TreeBASE will include Life Science Identifiers (LSID),[18] which will support persistent linking and allow Dryad metadata to be replicated and shared with TreeBASE, and vice/versa.

Example 2A: Metadata record for the journal article with the Dryad handle: http://hdl.handle.net/10255/dryad.82.

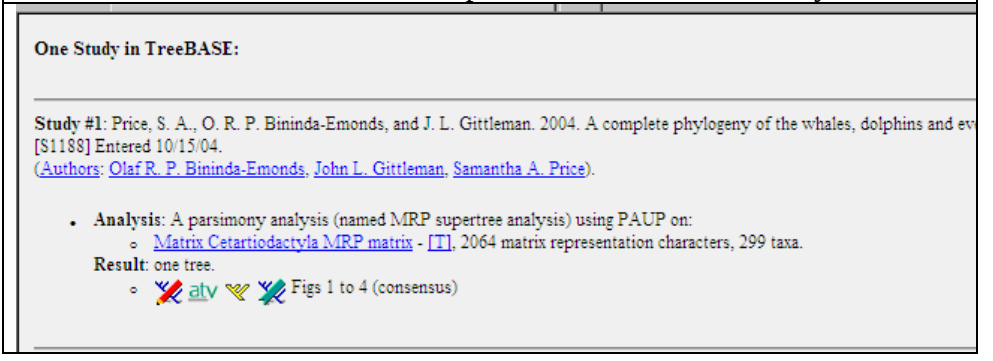


Publications >

Please use this identifier to cite or link to this item: http

Title:	Hunting to extinction: biology and re and the impact of hunting in artioda
Authors:	Price, Samantha A. Gittleman, John L.
Keywords:	extinction bushmeat hunting economic development phylogenetic comparative methods
Issue Date:	7-Aug-2007
Publisher:	Royal Society Publishing
Citation:	Price, S.A., Gittleman, J. L. 2007. Hu economy influence extinction risk an Proceedings of the Royal Society 274
Series/Report no.:	Proceedings of the Royal Society 274
Abstract:	Half of all artiodactyls (even-toed ho extinction, around double the mamm species-level phylogeny, we construc first time which intrinsic (biological) environmental) factors influence vari Globally artiodactyls at greatest risk

Example 2B: TreeBASE Metadata Record for accession #S1188 for a phylogenetic tree. The tree is associated with the article represented in 2A, via Dryad handle: http://hdl.handle.net/10255/dryad.82, and the Excel data file with the handle: http://hdl.handle.net/10255/dryad.83.



Enabling *new science* via data sharing/data reuse is a major if not the key incentive for digital data repositories. In fact, a recent survey involving nearly 400 evolutionary biologists, conducted by Dryad's development team, and focusing on data sharing attitudes and behaviors, confirms this view among scientists. Over 65% of the participants selected reasons, such as promoting new research, exploring new topics not envisioned, or creating new data sets, as top rationales for supporting data sharing, whereas validating results was indicated as a top rational by less than 5% of the participants. Evolutionary biologists enthusiastically engage in data sharing. Over half of the approximate 350 participants answering the data sharing behavior questions indicated that they have been asked to share data, and accommodated such requests 83% of the time. Additionally, 69% of these participants noted they have requested data from other researchers.

These initial survey results, and the fact that 70% of participants indicated their data was in digital format, bring metadata management needs to the forefront. This is an opportune time to push for a deeper understanding of metadata and to explore theoretical frameworks that may help address repository metadata challenges more effectively. In pursuing this agenda, it seems that lifecycle ideas and concepts not only have appeal, but a proven applicability. It also seems that such a framework can extend beyond the immediate confines of what has been presented in this article to include metadata replication and revision in other related venues. For example, what is the relationship between bibliographic citation tools, such as EndNote and RefWorks, and the lifecycle framework explored here? These applications allow one to download citations from bibliographic databases and automatically replicate citations for papers and other outputs. There are also social networking enterprises like Connotea[19] that scientists are using daily for the management of bookmarks and citations, and which support metadata enhancement with tags—a more *anarchistic form of value system adoption*. Metadata is being reproduced in many different contexts, and often associated with a distinct copy of the published

research, such a copy of an article that is downloaded on a personal computer. More research is needed to consider how these developments fit into the lifecycle framework, how they may support metadata and data sharing and reuse, and, ultimately, how they impact our conception of metadata within the digital repository environment.

CONCLUSION

The introduction of digital technology provides exciting new opportunities relating to information communication and the management of information via metadata, and with this comes new challenges that call for new solutions. This is evident with the development of digital repositories and the growing, critical role for metadata. Given similarities between the library and repository environment, it is natural to look to library cataloging practice and theoretical frameworks to help make sense of this new world and seek solutions for challenges. However, as articulated in the introduction of this article, there are fundamental differences between repositories and libraries that impact metadata activities.

This article considered lifecycle management as a framework for understanding metadata. The article explored the notion of theory in general, intellectual progress in our understanding of metadata, and the concept of lifecycle modeling. Dryad examples demonstrating lifecycle modeling via automatic propagation, metadata inheritance, and value system adoption were presented, along with results from a term mapping/controlled vocabulary experiment illustrating an approach for value system adoption. The discussion presented results from a recent Dryad led survey confirming that evolutionary biologists strongly support data sharing; the discussion also further speculates that the aspiration of data sharing is a key factor pushing metadata issues to the forefront.

The work presented here has been motivated by a need for a theoretical context to aid communication among members of the Dryad development team. It is likely that similar needs permeate

other repository initiatives, particularly those that link and share data with other repositories or databases. The analysis confirms the applicability of lifecycle modeling as theoretical framework for understanding metadata in the digital repository environment; and the ideas presented here are aiding Dryad develop team members in their ongoing discussions on metadata issues and plans Dryad is in an early stage of development, and as the repository grows, additional analyses will be conducted to further develop a more grounded metadata theory to enhance our understanding of metadata.

NOTES

1. Libraries catalog and create metadata for resource holdings. Cataloging is more often associated with print and physical resource, whereas metadata is more often associated with digital resources, although the terms *cataloging* and *metadata* are used interchangeably.
2. Metadata Encoding and Transmission Standard (METS) Official Website, http://www.loc.gov/standards/mets/.
3. arXiv.org e-Print archive, http://arxiv.org/.
4. Dryad, http://datadryad.org/.
5. Ann Arbor Accords: Principles and Criteria for an SGML Document Type Definition (DTD) for Finding Aids, http://sunsite.berkeley.edu/FindingAids/EAD/accords.html.
6. Design Principles for Text Encoding Guidelines: TEI ED P1. 1988 [rev. January 1990], http://www.w3.org/People/cmsmcq/1990/edp1.html.
7. Resource Description Framework (RDF), http://www.w3.org/RDF/.
8. DCMI Abstract Model, http://dublincore.org/documents/abstract-model/.
9. National Evolutionary Synthesis Center (NESCent), http://www.nescent.org/.
10. SILS Metadata Research Center, UNC/CH, http://ils.unc.edu/mrc/.

11. News of National Science Foundation/Division of Biological Infrastructure (NSF/DBI) recent award, http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0743720.
12. Dryad Partners, http://datadryad.org/partners.html.
13. Dryad project description, https://www.nescent.org/wg/digitaldata/images/9/96/Dryad.proj.descr.07.pdf.
14. TreeBASE, http://www.treebase.org/.
15. GenBank, http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html.
16. Helping Interdisciplinary Vocabulary Engineering (HIVE), http://ils.unc.edu/mrc/hive/.
17. A phylogenetic tree tracks the evolution of a biological species, by tracking relationships to among other species that are believed to, or shown to have had common ancestors.
18. LISD are "persistent, location-independent, resource identifiers for uniquely naming biologically significant resources including but not limited to individual genes or proteins, or data objects that encode information about them" (http://xml.coverpages.org/lsid.html).
19. Connotea: http://www.connotea.org/.

ACKNOWLEDGEMENTS

This work is supported by National Science Foundation (NSF), grants #EF-0423641 and NSF/BDI #0743720. I would thank Dryad development team members for their inspiration, input, and encouragement with this paper.

REFERENCES

1. Alan Danskin, "Tomorrow Never Knows: The End of

Cataloguing?" (paper presented at the World Library and Information Congress: 72nd IFLA General Conference and Council, Seoul, Korea, August 20-24, 2006). http://www.ifla.org/IV/ifla72/papers/102-Danskin-en.pdf (accessed November 2, 2008).

2. Thomas Mann, "Off the Record but Off the Track: A Review of the Report of The Library of Congress Working Group on The Future of Bibliographic Control, With a Further Examination of Library of Congress Cataloging Tendencies," 14 March 2008, http://www.guild2910.org/WorkingGrpResponse2008.pdf (accessed November 2, 2008).

3. Karen Coyle, "Technology and the Return on Investment," *Journal of Academic Librarianship* 32:5 (2006): 537-539.

4. Library of Congress Working Group on the Future of Bibliographic Control, "Report on the Future of Bibliographic Control," 30 November 2007, http://www.loc.gov/bibliographic-future/news/lcwg-report-draft-11-30-07-final.pdf (accessed November 2, 2008).

5. Jihie Kim, Yolanda Gil, and Varun Ratnakar, "Semantic Metadata Generation for Large Scientific Workflows." In *Proceedings of the 5th International Semantic Web Conference, ISWC-2006, Athens, GA, USA*, edited by I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, 357-370. Berlin: Springer, 2006.

6. Marko A. Rodriguez, Johan Bollen, and Herbert Van de Sompel, "Automatic Metadata Generation Using Associative Networks," *ACM Transactions on Information Systems* 27:2 (2008). http://arxiv.org/abs/0807.0023 (accessed November 2, 2008).

7. Jane Greenberg, "Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications," *Journal of Internet Cataloging* 6:4 (2004): 59-82.

8. Jane Greenberg, Kristina Spurgin, and Abe Crystal, "Functionalities for Automatic-Metadata Generation Applications: A Survey of Metadata Experts' Opinions," *International Journal of Metadata, Semantics, and Ontologies* 1:1 (2006): 3-20.

9. Jane Greenberg, Kristina Spurgin, and Abe Crystal, "Final Report for the AMeGA (Automatic Metadata Generation Applications) Project," 17 February 2005, http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf (accessed November 2, 2008).

10. M.C. Leoni, M. Dolensky, P. Padovani, P. Rosati, A. Wicenec, and A. Micol, "Multi-Purpose Metadata Repository for a Real and Virtual Observatory," *Astronomical Data Analysis Software and Systems XV* (2006): 414.

11. William Whewell, *Theory of Scientific Method*, 2nd edition, ed. Robert E. Butts (Indianapolis: Hackett Publishing Company, 1989).

12. Ibid.

13. George Lakoff and Mark Johnson, *Metaphors We Live By* (Chicago: The University of Chicago Press, 1980).

14. Michael J. Reddy, "The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language," in *Metaphor and Thought*, 2nd edition, ed. Andrew Ortony (Cambridge: Cambridge University Press, 1993), 284-324.

15. Francis Miksa, "The Cultural Legacy of the 'Modern Library' for the Future," *Journal of Education for Library and Information Science* 37:2 (1996): 100-119.

16. Doralynn J. Hickey, "Theory of Bibliographic Control in Libraries," *Library Quarterly* 47:3 (1977): 253-273.

17. Doralynn J. Hickey, "Bibliographic Control in Theory," *IFLA Journal* 6:3 (1980): 234-241.

18. Lynne C. Howarth, "Metadata and Bibliographic Control: Soul-Mates or Two Solitudes?," *Cataloging and Classification Quarterly* 40:3 (2005): 37-56.

19. Jane Greenberg, "Understanding Metadata and Metadata Schemes," *Cataloging and Classification Quarterly* 40:3/4 (2005): 17-36.

20. Charles A. Cutter, *Rules for a Dictionary Catalog*, 4th ed. (Washington, DC: Government Printing Office, 1904).

21. Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L. Weibel, "Metadata Principles and Practicalities," *D-Lib Magazine* 8:4 (2002). http://www.dlib.org/dlib/april02/weibel/04weibel.html (accessed November 2, 2008).

22. Marieke Guy, Andy Powell and Michael Day, "Improving the Quality of Metadata in Eprint Archives," *Ariadne* 28 (2004). http://www.ariadne.ac.uk/issue38/guy/ (accessed November 2, 2008).

23. Jane Greenberg, "Metadata and the World Wide Web," in *Encyclopedia of Library and Information Science*, ed. Marcia J. Bates, Mary Niles Maack, and Miriam Drake (New York: Marcel Dekker, Inc., 2003), 1876-1888.

24. Anne J. Gilliland, "Setting the Stage," in *Introduction to Metadata, Version 3.0*, ed. Murtha Baca (Los Angeles, CA: Getty Information Institute, 2008). http://www.getty.edu/research/conducting_research/standards/intrometadata/setting.html (accessed November 2, 2008).

25. Greenberg, "Understanding Metadata and Metadata Schemes,"

17-36.

26. Thomas R. Bruce, and Diane I. Hillmann, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice*, ed. Diane I. Hillmann and E.L. Westbrooks (Chicago: American Library Association, 2004), 238-256.

27. Kat Hagedorn and Sarah Shreeves, eds., "Best Practices for OAI Data Provider Implementations and Shareable Metadata," 25 June 2007, http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/BestPracticesIntroduction (accessed November 2, 2008).

28. Mikael Nilsson, Matthias Palmér, and Ambjörn Naeve, "Semantic Web Metadata for E-Learning–Some Architectural Guidelines," in *Proceedings of the 11th World Wide Web Conference*, 2002, http://kmr.nada.kth.se/papers/SemanticWeb/p744-nilsson.pdf .

29. IFLA Study Group on the Functional Requirements for Bibliographic Records, "Functional Requirements for Bibliographic Records: Final Report," February 2008, http://www.ifla.org/VII/s13/frbr/frbr_2008.pdf (accessed November 2, 2008).

30. Sarah Higgins, "The DCC Curation Lifecycle Model," *The International Journal of Digital Curation* 1:3 (2008): 134-140.

31. Ibid., 135.

32. Steve Duplessie, Nancy Marrone, and Steve Kenniston, "The New Buzzwords: Information Lifecycle Management," 31 March 2003, http://www.computerworld.com/hardwaretopics/storage/story/0,108

01,79885,00.html (accessed November 2, 2008).

33. Ya-ning Chen and Shu-jiun Chen, "Metadata Lifecycle Model and Metadata Interoperability" (paper presented at the 5th International Conference on Conception of Library and Information Science (CoLIS 5), June 4-8, 2005, University of Strathclyde, Glasgow, UK). http://pl11.sinica.edu.tw:8080/dspace/handle/1868/2273 (accessed November 2, 2008).

34. Ann Green and Jean-Pierre Kent, "The Metadata Life Cycle," in *MetaNetWork Package 1: Methodology and Tools*, ed. Jean-Pierre Kent (The MetaNet Project, 2002), 29-34. http://www.epros.ed.ac.uk/metanet/deliverables/D4/IST_1999_29093_D4.pdf (accessed November 2, 2008).

35. Research Libraries Group, *Trusted Digital Repositories: Attributes and Responsibilities, an RLG-OCLC Report* (Mountain View, California: RLG, Inc., 2002). http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (accessed November 2, 2008).

36. Ronald Jantz and Michael J. Giarlo, "Digital Preservation: Architecture and Technology for Trusted Digital Repositories," *D-Lib Magazine* 11:6 (2005).

37. Richard P. Smiraglia, *The Nature of "A Work": Implications for the Organization of Knowledge* (Lanham, MD: Scarecrow Press, 2001), 88-119, 165.

38. Richard P. Smiraglia, ed., *Works as Entities for Information Retrieval* (Binghamton, New York: Haworth Press, 2002).

39. Martha M. Yee, "What is a work?" (paper presented at the International Conference on the Principles and Future Development of AACR, Toronto, Ontario, Canada, October 23-25, 1997). http://repositories.cdlib.org/postprints/3085/ (accessed November 2,

2008).

40. Barbara Tillet, "A Taxonomy of Bibliographic Relationships," *Library Resources and Technical Services* 35:2 (1991b): 150-158.

41. Gregory H. Leazer and Richard P. Smiraglia, "Bibliographic Families in the Library Catalog: A Qualitative Analysis and Grounded Theory," *Library Resources and Technical Services* 43:4 (1999): 191-212.

42. Richard P. Smiraglia. "A Meta-Analysis of Instantiation as a Phenomenon of
Information Objects," *Culture del testo e del documento* 9: 25 (2008): 5-25).

43. Anita Sundaram Coleman. (2002) "Scientific Models as Works," *Cataloging and Classification Quarterly* 33:3/4 (2002): 129-159.

44. Richard P. Smiraglia, "Authority Control and the Extent of Derivative Bibliographic Relationships" (Ph.D. diss., University of Chicago, 1992).

45. Jed Dube, Sarah Carrier, and Jane Greenberg, "DRIADE: A Data Repository or Evolutionary Biology," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, edited by Ray Larson, Edie Rasmussen, Shigeo Sugimoto and Elaine Toms, 481. New York, New York: ACM Press, 2007.

46. Ryan Scherle, Sarah Carrier, Jane Greenberg, Hilmar Lapp, Abbey Thompson, Todd Vision, and Hollie White, "Building Support for a Discipline-Based Data Repository" (poster presented at the Third International Conference on Open Repositories 2008, April 1-4 2008, Southampton, United Kingdom).

47. Jennifer L. Knies, Kristen K. Dang, Todd J. Vision, Noah G. Hoffman, Ronald Swanstrom, and Christina L. Burch,

"Compensatory Evolution in RNA Secondary Structures Increases Substitution Rate Variation among Sites," *Molecular Biology and Evolution* 25:8 (2008): 1778-1787.

48. Sarah Carrier, Jed Dube, and Jane Greenberg, "The DRIADE Project: Phased Application Profile Development in Support of Open Science," in *Proceedings of the International Conference on Dublin Core and Metadata Applications 2007*, edited by S.A. Sutton, A.S. Chaudhry, and C. Khoo, 35-42. Dublin Core Metadata Initiative, Singapore, 2007.

49. Sarah Carrier, "The Dryad Repository Application Profile: Process, Development, and Refinement" (Master's Paper, University of North Carolina at Chapel Hill, 2008).

50. Renear, Allen H., Karen M. Wickett, Richard J. Urban, Dave Dubin, and Sarah L. Shreeves, "Collection/Item Metadata Relationships," in *Proceedings of the International Conference on Dublin Core and Metadata Applications 2008*, edited by J. Greenberg and W. Klaus, 80-89. Dublin Core Metadata Initiative, Berlin, Germany, September 22-26, 2008.

51. Stephan Reebs, *Fish Behavior in the Aquarium and in the Wild* (Ithaca, New York: Cornell University Press, 2001).

42. Richard P. Smiraglia, "Subject Access to Archival Materials Using LCSH," *Cataloging and Classification Quarterly* 11:3/4 (1990): 63-90.

53. Robert Losee, "A Performance Model of the Length and Number of Subject Headings and Index Phrases," *Knowledge Organization* 31:4 (2004): 245-251.