

# Linking and Hiving Data in the Dryad Repository

## *The Semantic Web: Fact or Myth*

CENDI, FLIICC, and NFAIS Workshop  
National Archives, Washington, DC

Tuesday, November 17, 2009

JANE GREENBERG, PROFESSOR,  
DIRECTOR OF THE METADATA RESEARCH CENTER  
SCHOOL OF INFORMATION & LIBRARY SCIENCE  
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

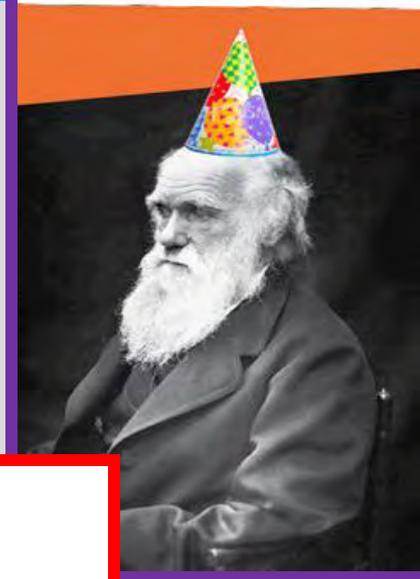
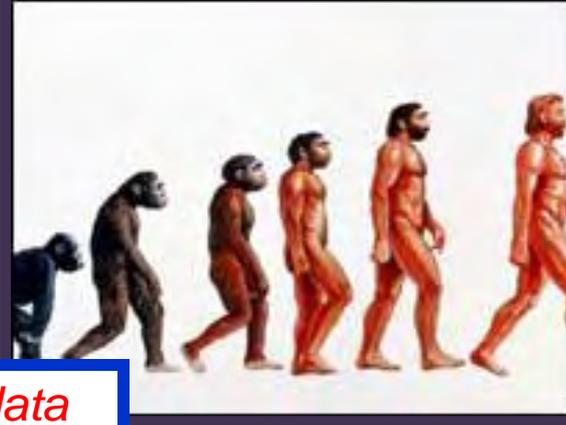
[janeg@email.unc.edu](mailto:janeg@email.unc.edu)



# Overview

1. Dryad repository project
  - Motivation and goals
  - Dryad's metadata framework
  - Some conclusions
2. Helping Interdisciplinary Vocabulary Engineering (HIVE)
3. DCMI-Science and Metadata (SAM)
4. Questions, discussion...

# Dryad

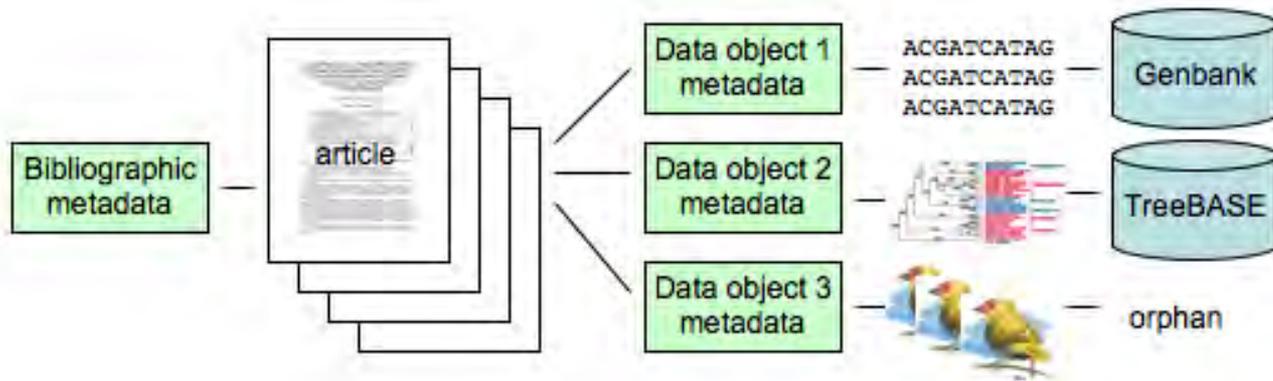


~ Evolutionary biologists use published data more frequently than they are depositing it themselves!

~ validation of results, data reuse, meta-analysis, synthesis

- 48 % based on other data
- 78% data analyzed not deposited
- Survey of 400 evol. biologists...

- Ecology
- Paleontology
- Population genetics
- Physiology
- Systematics
- Genomics



# Dryad's Goals

1. One-stop deposition and shopping for data objects supporting published research...

*Handshaking...*

2. Support the acquisition, preservation, resource discovery, and reuse of heterogeneous digital datasets

## **Dryad Team** **NESCent**



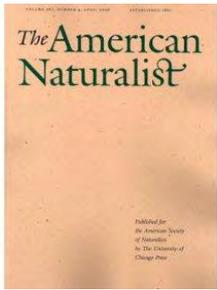
- Hilmar Lapp  
Assistant Director for Informatics
- Ryan Scherle  
Data Repository Architect
- Todd Vision, Associate Director of Informatics

## **UNC/SILS/MRC**

- Jane Greenberg, Professor
- Lina Huang, MSIS Student, Research Assistant
- Robert M. Losee, Professor
- Jose R. Pérez-Agüera, Clinical Assistant Professor
- Hollie White, Metadata Research Center Doctoral Fellow



**North Carolina State University,  
University of New Mexico/LTER, Yale  
University, + partner journals and  
societies**



# Partner Journals

**American Society of Naturalists**

*American Naturalist*

**Ecological Society of America**

*Ecology, Ecological Letters, Ecological Monographs, etc.*

**European Society for Evolutionary Biology**

*Journal of Evolutionary Biology*

**Society for Integrative and Comparative Biology**

*Integrative and Comparative Biology*

**Society for Molecular Biology and Evolution**

*Molecular Biology and Evolution*

**Society for the Study of Evolution**

*Evolution*

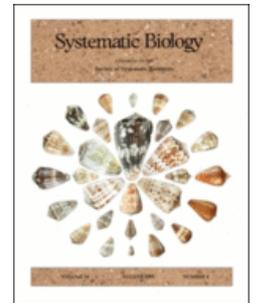
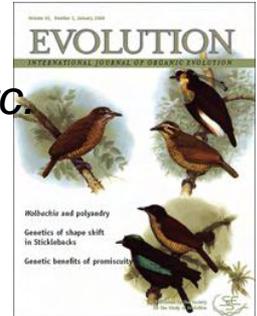
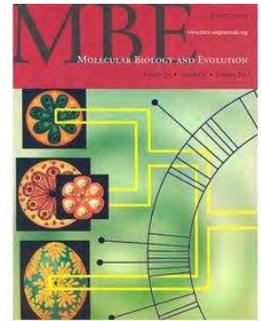
**Society for Systematic Biology**

*Systematic Biology*

**Commercial journals**

*Molecular Ecology*

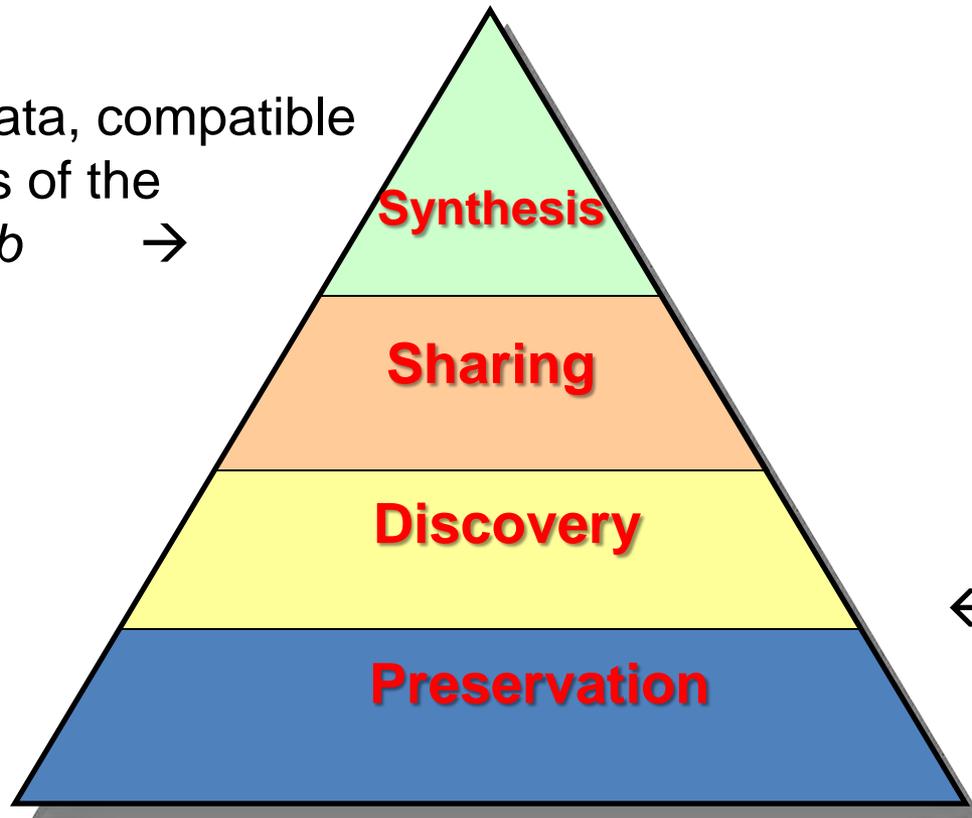
*Molecular Phylogenetics and Evolution*



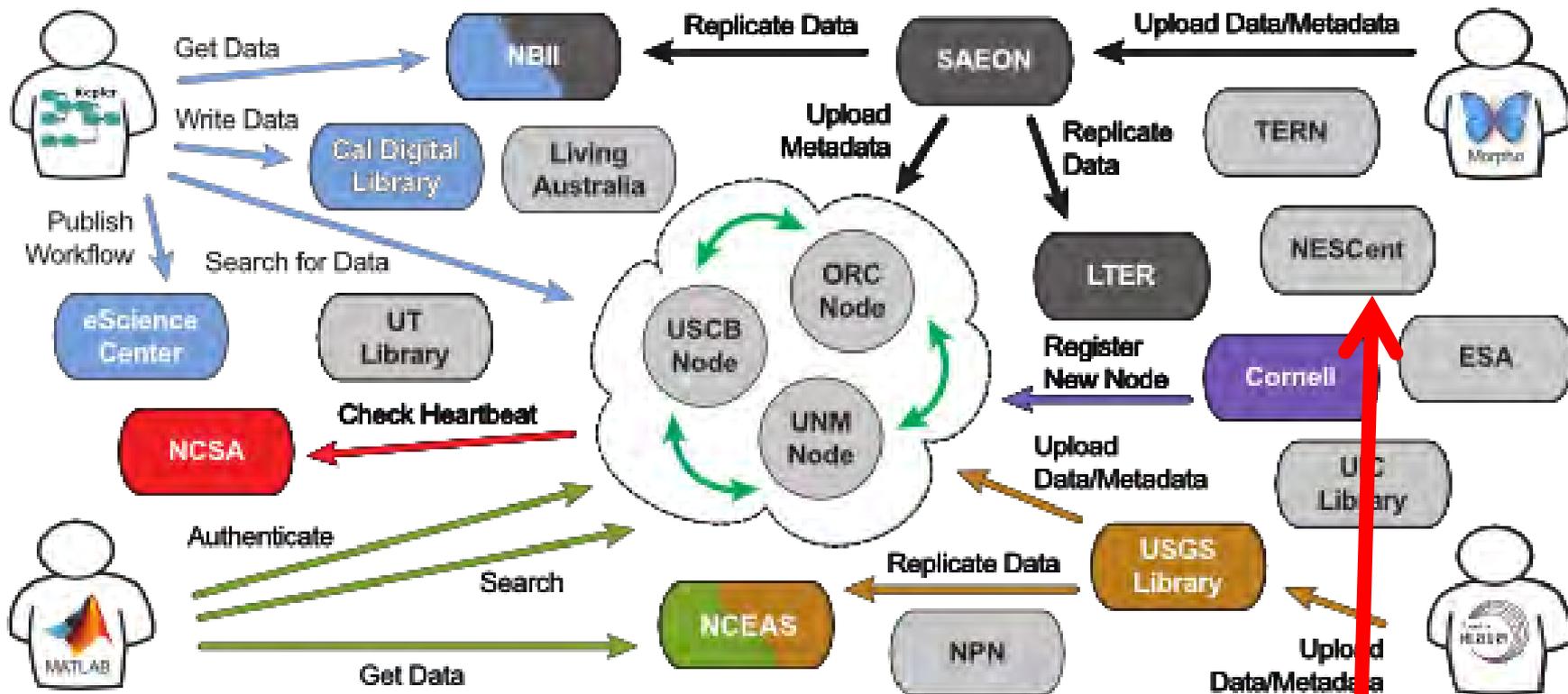
SCHOOL OF INFORMATION AND LIBRARY SCIENCE

# A hierarchy of goals

RDF, linked data, compatible  
With the goals of the  
*Semantic Web* →



← Basic resource  
description

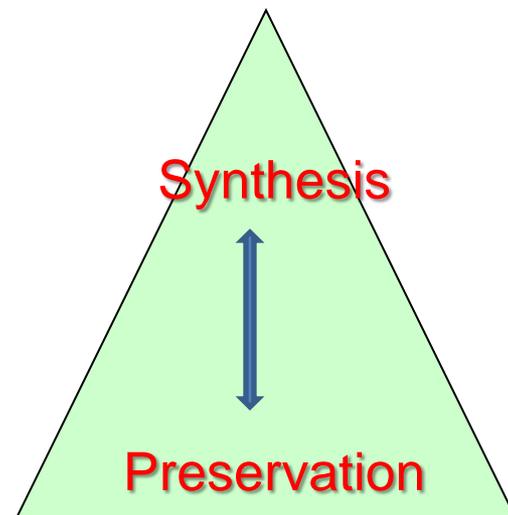


# The Dryad Repository's Metadata Framework

# Dryad's Metadata

~ METADATA, from the onset, and important part of Dryad  
Not let legacy practice hold us back

- Dryad's high-level metadata requirements
  - *Simple*
  - *Interoperable*
  - *Semantic Web compatible*
- A two-pronged approach
  1. Sustainable, long term
  2. Support immediate needs



# Dryad metadata application profile

## Dublin Core based

- Circumvents limitations of using a single scheme
- Interoperable with other schemes
- Why reinvent the wheel?

### Modular scheme:

1. Journal citation
2. Data objects

(Carrier, et al., 2007; White, et al, 2008; Greenberg et al; in press *JLM*, in press)

### Namespaces:

1. Dublin Core
2. Data Documentation Initiative (DDI)
3. Ecological Metadata Language (EML)
4. Journal Publishing Tag Set Tag Library
5. PREMIS Data Dictionary for Preservation Metadata
6. Publishing Requirements for Industry Standard Metadata (PRISM)
7. Darwin Core (DwC)
8. Dryad

# <DRYAD application profile, ver. 2.0>

## Data Object Module

1. Dryad:Status/Status \*
2. dc:creator/Name\*
3. **dc:title/Data Set #**
4. dc:identifier/Data Set Identifier
5. PREMIS:fixity/(hidden)
6. dc:relation/DOI of Published Article
7. DDI:<depositr>/Depositor \*
8. DDI:<contact>/Contact Info. #
9. dc:rights/Rights Statement
10. **dc:description/Description #**
11. dc:subject/Keywords \*
11. dc:coverage / Locality Required \*
12. dc:coverage/Date Range Required\*
13. dc:software/Software\*
14. dc:format/File Format
15. dc:format/File Size
16. dc:date/(Hidden) Required
17. dc:date/Date Modified\*
18. Darwin Core: species/ Species, or Scientific\*

### Key

\* = semi-automatic

# = manual

Everything else is automatic

# Citation metadata from a published research article in *Molecular Biology and Evolution*.

**Search Dryad**  
   
[Advanced Search](#)

[Home](#)

**Browse**

- [Communities & Collections](#)
- [Titles](#)
- [Authors](#)
- [Subjects](#)
- [By Date](#)

**Sign on to:**

- [Receive email updates](#)
- [My Dryad](#)  
authorized users
- [Edit Profile](#)
- [Help](#)
- [About Dryad](#)

[Dryad >](#)  
[Main >](#)  
[Publications >](#)

Please use this identifier to cite or link to this item: <http://hdl.handle.net/10255/dryad.162>

**Title:** Compensatory Evolution in RNA Secondary Structures Increases Substitution Rate Variation among Sites.

**Authors:** Knies, Jennifer L.  
Dang, Kristen K.  
Vision, Todd J.  
Hoffman, Noah G.  
Swanstrom, Ronald  
Burch, Christina L.

**Issue Date:** 2008

**Publisher:** Oxford University Press

**Citation:** Compensatory Evolution in RNA Secondary Structures Increases Substitution Rate Variation among Sites. 2008. Knies, Jennifer L., et. al. *Molecular Biology and Evolution*. 25(8):1-10. doi:10.1093/molbev/msn130

**Series/Report no.:** *Molecular Biology and Evolution* 25(8):1-10

**Description:** There is growing evidence that interactions between biological molecules (e.g., RNA-RNA, protein-protein, RNA-protein) place limits on the rate and trajectory of molecular evolution. Here, by extending Kimura's model of compensatory evolution at interacting sites, we show that the ratio of transition to transversion substitutions ( $j$ ) at interacting sites should be equal to the square of the ratio at independent sites. Because transition mutations generally occur at a higher rate than transversions, the model

Automatic propagation of article citation metadata, **enhanced citation**—  
including taxonomic and geographic metadata

# Data object metadata for item underlying published research in previous slide

[Dryad](#) >  
[Main](#) >  
[Data](#) >

Please use this identifier to cite or link to this item: <http://hdl.handle.net/10255/dryad.169>

**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>

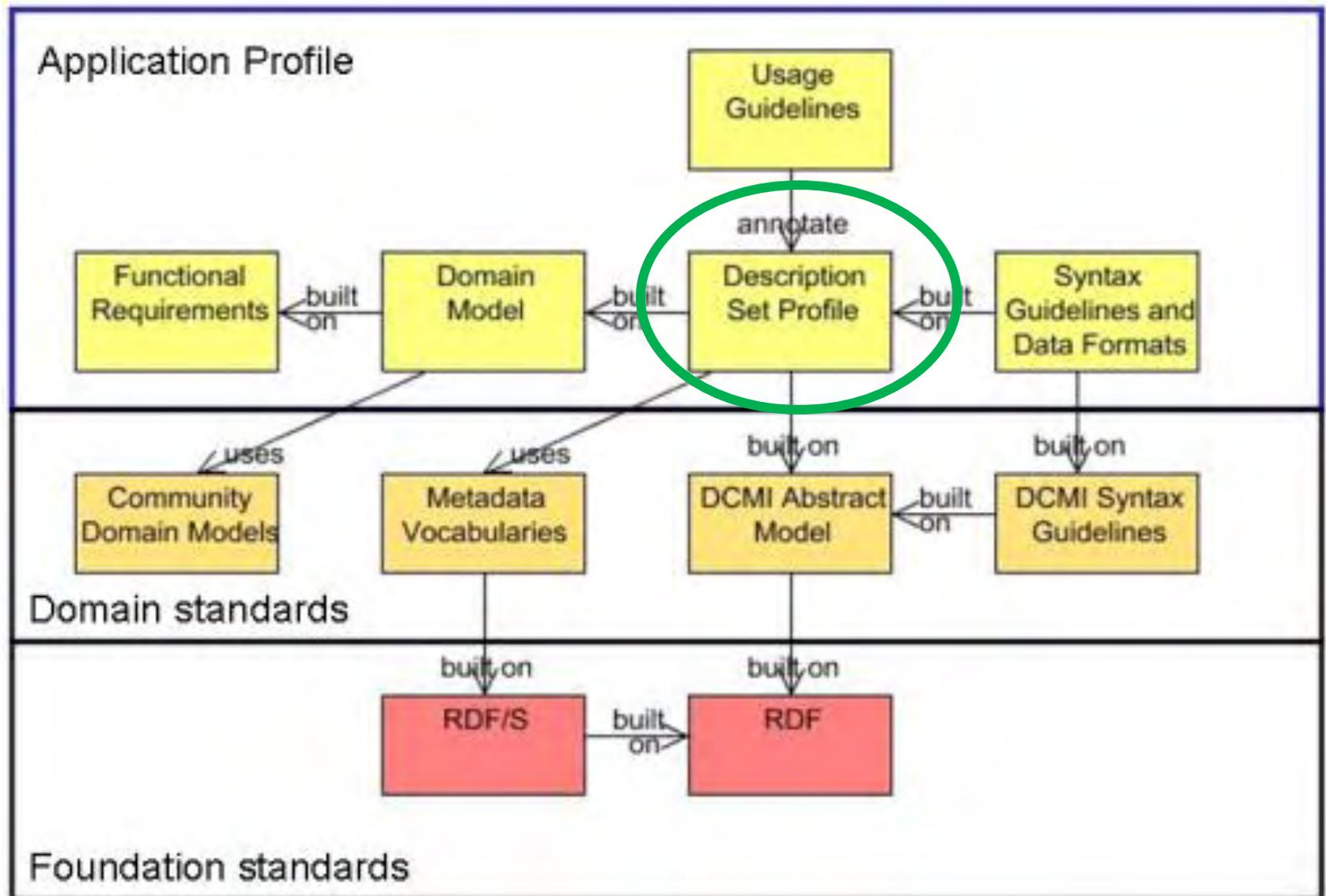
**Appears in Collections:** [Data](#)

## Files in This Item:

File	Description	Size	Format	
<a href="#">16SandTREE.nex</a>		151.74 kB	Nexus	<a href="#">View/Open</a>

# Singapore Framework

(machine processing, **metadata** → **data reuse**, Semantic web/linked data)



■ A loose standard for DC endorsed application profiles;

# Singapore Framework

<http://dublincore.org/documents/singapore-framework/>

- A “loose” standard for Dublin Core “**endorsed**” application profiles
- Singapore framework provides guidelines for creating a DCAM-conformant Application Profile (“DC Application Profile”)
- A packet of documentation which consists of:
  1. **Functional requirements (desirable)**
  2. **Domain model (mandatory)**
  3. **Description Set Profile (DSP) (mandatory)**
  4. **Usage guidelines (optional)**
  5. **Encoding syntax guidelines (optional)**

*stuff we do any how...just better!* (JLM paper)

# Description Set Profile

*Potential for metadata and ...data reuse in different contexts*

- DSP is “an information model and XML expression”

(<http://www.unc.edu/~scarrier/dryad/DSPLevelOneAppProfDraft.xml>)

– Obligation (optional, mandatory)

– **Non-literal** (thing – philosophically – *things* in the real world, known in different ways)

- <http://purl.org/dc/elements/1.1/rights> (mandatory), there are different rights
- Subject, creator, description...

– **Literals** (strings):

- <http://purl.org/dc/elements/1.1/identifier> = <http://purl.org/dc/terms/URI>,
- <http://purl.org/dc/terms/available> = <http://purl.org/dc/terms/W3CDTF>

**Toward the Semantic Web...**  
*linked data...*

# Reality...

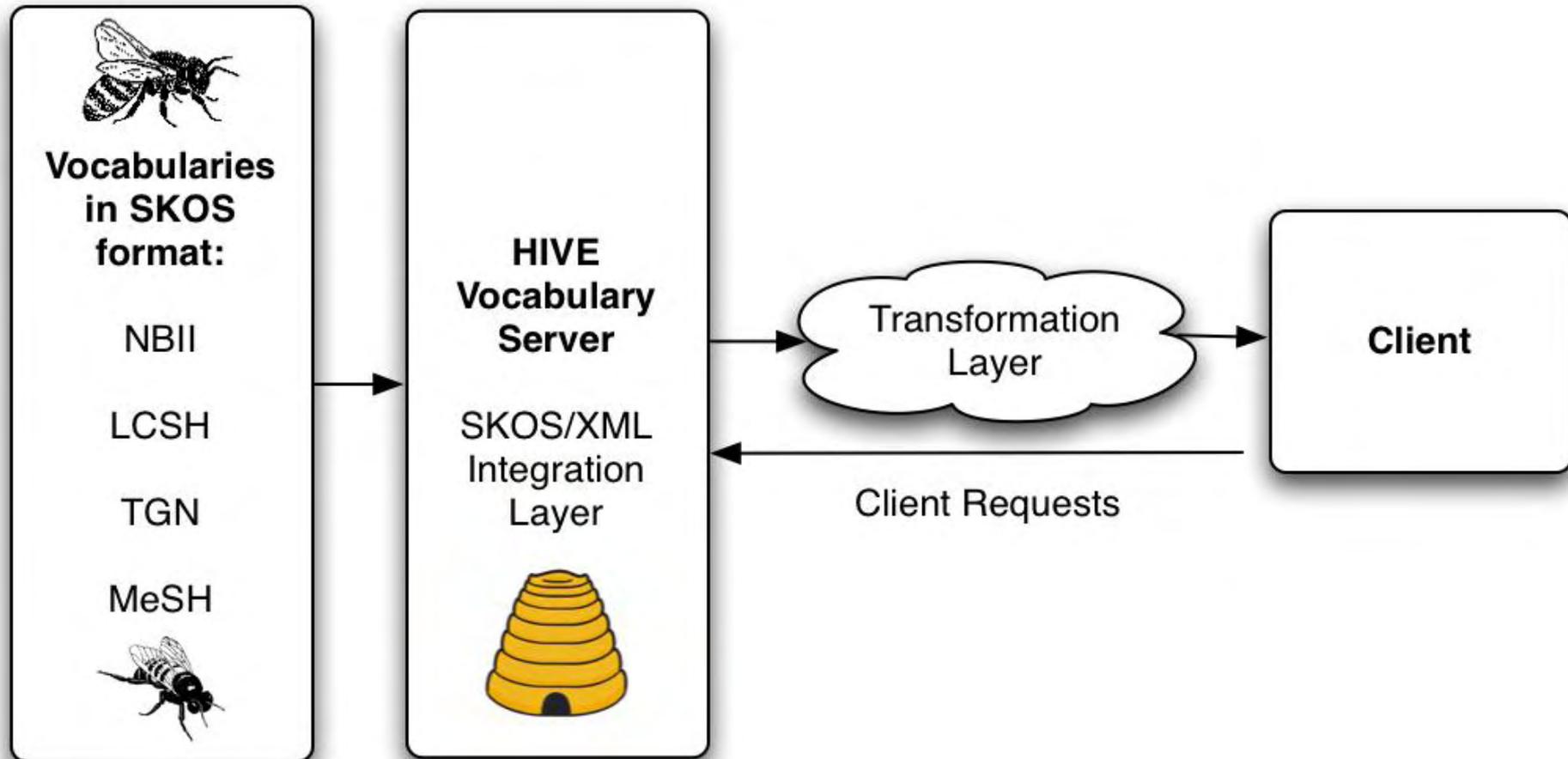
- Dryad development -- *reversed engineered* -- long-term goals first *did not have to hit the ground running*
- Commencement of data exchange plans  
DCAP work refocus → XML schema
  - Harvesting EML records from LTER Network's Metacat
  - Data exchange plans Dryad ↔ TreeBASE
- Discouraged 😞...approaches not at odds 😊
  - App. profile work instrumental in creating a sufficient XML schema
  - Renderings of the same intellectual on a continuum
  - GRDDL (Gleaning Resource Descriptions from Dialects of Languages) – enabling technology to create RDF

# Application profile work, *thoughts...to date...*

- Positive aspects
  - Intellectually engaging
  - Think we are making a contribution, have to start somewhere...
  - Machine capabilities
  - eScience/data synthesis
- Challenges
  - Infrastructure not all there... (a lot is not in RDF)
    - Registered Dryad “purl”
  - Proof of concept difficult
  - Time consuming
  - Documentation lacking

# Helping Interdisciplinary Vocabulary Engineering (HIVE)

# Helping Interdisciplinary Vocabulary Engineering (HIVE)



- <AMG> approach for integrating discipline CVs
- Model addressing **C V cost, interoperability, and usability constraints** (interdisciplinary environment)

*Building, Sharing, Evaluation* the HIVE....

# HIVE Partners

## Vocabulary Partners

- Library of Congress:  
*LCSH*
- the Getty Research Institute (GRI): *TGN (Thesaurus of Geographic Names )*
- United States Geological Survey (USGS): *NBII Thesaurus*



## Advisory Board

- Jim Balhoff, NESCent
- Libby Dechman, LCSH
- Mike Frame, USGS
- Alistair Miles, CCLRC Rutherford Appleton Laboratory
- William Moen, University of North Texas
- Eva Méndez Rodríguez, University Carlos III of Madrid
- Joseph Shubitowski, Getty Research Institute
- Ed Summers, LCSH
- Barbara Tillett, Library of Congress
- Kathy Wisser, UNC Chapel Hill
- Lisa Zolly, USGS

**WORKSHOPS HOSTS:** Columbia Univ.; Univ. of California, San Diego; Univ. of North Texas; Universidad Carlos III de Madrid, Madrid, Spain

# HIVE Construction

- HIVE's technological infrastructure *stores millions of concepts from different vocabularies* and is preparing to make them available on the Web by a simple HTTP
- Vocabularies are imported into HIVE using SKOS/RDF format
- HIVE has two modules:

## **HIVE Core**

- SKOS/RDF storage and management (SESAME/Elmo)
- Automatic Metadata Extraction and Topic Detection (KEA++ and MAUI)
- Concept Retrieval (Lucene and MG4J)

## **HIVE Web**

- Web user Interface (GWT—Google Web Toolkit)
- Machine oriented interface (SOAP and REST)

# FROM NBII

```
<rdf:Description rdf:about="http://thesaurus.nbii.gov/Mud">
  <rdf:type
rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:broader rdf:resource="http://thesaurus.nbii.gov/Sediments"/>
  <skos:prefLabel>Mud</skos:prefLabel>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Clays"/>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Mud-flats"/>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Oozes"/>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Silt"/>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Slimes"/>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Sludges"/>
  <skos:related rdf:resource="http://thesaurus.nbii.gov/Soils"/>
  <skos:scopeNote>ASF Aquatic Sciences and Fisheries LSC Life
Sciences</skos:scopeNote>
</rdf:Description>
```



## Help with Interdisciplinary Vocabulary Engineering

Home

Concept Browser

Indexing



### Welcome to HIVE Vocabulary Server!

Helping Interdisciplinary Vocabulary Engineering (HIVE) is an IMLS funded project involving the [Metadata Research Center \(MRC\)](#) at the School of Information and Library Science, University of North Carolina at Chapel Hill, and the [National Evolutionary Synthesis Center \(NESCent\)](#) in Durham, North Carolina. HIVE is an automatic metadata generation approach that dynamically integrates discipline-specific controlled vocabularies encoded with the [Simple Knowledge Organisation System \(SKOS\)](#), a [World Wide Web Consortium \(W3C\)](#) standard. HIVE Vocabulary Server is a web based system for searching and browsing concepts in interdisciplinary vocabularies, and providing cataloging aids by automatically extracting concepts for a given document.

#### Search Concept

Search

[Go to Concept Browser](#)

#### Annotate Document

Upload

[Go to Annotation](#)

#### Formal Queries

SPARQL

SERQL

Execute

#### Vocabulary Statistics

Vocabulary	Concepts	Relationships	Date Added
LCSH	10,236	5,635	08-13-2009
NBII	21,129	5,635	08-13-2009
GRI	21,129	5,635	08-13-2009
USGS	21,129	5,635	08-13-2009
MeSH	21,129	5,635	08-13-2009
ITIS	21,129	5,635	08-13-2009
WordNet	21,129	5,635	08-13-2009
TGN	21,129	5,635	08-13-2009
UBIO	21,129	5,635	08-13-2009
Gene Ontology	21,129	5,635	08-13-2009

Last updated: 08-21-2009



## Help with Interdisciplinary Vocabulary Engineering

Home

Concept Browser

Indexing

Current vocabularies sources: [x LCSH](#) [x NBII](#) [+ Add](#)

HIVE vocabulary server provides functionalities to identify concepts from a given document and automatically assign subject metadata to the document. Working with HIVE to assign metadata from authoritative vocabularies is easy and simple. You just need to upload the document and select the vocabularies you are interested in.

[Help](#)

### HIVE Automatic Metadata Extrator

Current vocabularies: [x LCSH](#) [x NBII](#) [Choose](#)

Prioritize the vocabularies: [1. LCSH](#) [2. NBII](#)  
Drag the name of the vocabulary to change the order .

#### Option 1:

Upload the document:  [Upload](#)

#### Option 2:

Enter text here:

[submit](#) [Reset](#)



## Help with Interdisciplinary Vocabulary Engineering

[Home](#)

[Concept Browser](#)

[Indexing](#)

Current vocabularies sources: [x LCSH](#) [x NBII](#) [+ Add](#)

HIVE vocabulary server provides functionalities to identify concepts from a given document and automatically assign subject metadata to the document. Working with HIVE to assign metadata from authoritative vocabularies is easy and simple. You just need to upload the document and select the vocabularies you are interested in.

[start over...](#)

### Document Summary

**Title:** The endocrinology of pregnancy and fetal loss in wild baboons

**Keywords:** Fetal loss; Miscarriage; Fecal steroids; Estrogens; Progestins; Glucocorticoids; Baboon; Papio; Pregnancy

**Abstract:** An impressive body of research has focused on the mechanisms by which the steroid estrogens (E), progestins (P), and glucocorticoids (GC) ensure successful pregnancy. With the advance of non-invasive techniques to measure steroids in urine and feces, steroid hormones are routinely monitored to detect pregnancy in wild mammalian species, but hormone data on fetal loss have been sparse. Here, we examine fecal steroid hormones from five groups of wild yellow baboons (*Papio cynocephalus*) in the Amboseli basin of Kenya to compare the hormones of successful pregnancies to those ending in fetal loss or stillbirth. Using a combination of longitudinal and cross-sectional data, we analyzed three steroid hormones (E, P, GC) and related metabolites from 5 years of fecal samples across 188 pregnancies. Our results document the course of steroid hormone concentrations across successful baboon pregnancy in the wild and demonstrate that fecal estrogens predicted impending fetal loss starting...

### Concept Cloud

[Export](#)

[+ Add your own concepts](#)



- LCSH [x](#)
- NBII [x](#)
- MeSH [x](#)

# Challenges

- Combining many vocabularies during the indexing/term matching phase is difficult, time consuming, inefficient.
  - There is some promise w/NLP and machine learning
- Interoperability = dumbing down
  - ontologies
- Proof-of-concept/ illustrate the differences between HIVE and other vocabulary registries
  - NCBO Bioportal, OBO Foundry

# Overall Conclusions

- Dryad development team has discovered that a two prong hybrid philosophy works well in this environment, keeps legacy practices from impeding development
- A bridge...
- Development has consensus driven
  - biologists, computer scientists, librarians
- New models need to be explored
- No exploration, no discovery...unsuccessful results are not BAD



**DCAP**  
**Semantic Web**  
**~ linked data...**



**XML Scheme**



[page](#) [discussion](#) [view source](#) [history](#)

## Main Page

### DCMI Science and Metadata Community



Dublin Core Metadata Initiative<sup>®</sup>  
Making it easier to find information.

The [DCMI Science and Metadata Community](#) is a forum for individuals and organizations to exchange information and knowledge about metadata describing scientific data (data methodologically collected for research, analysis, tracking, forecasting, and other uses). The Community focuses on metadata challenges specific to scientific data curation, and solutions that will benefit from the architecture and global reach of the [Dublin Core Metadata Initiative](#).

Join [the DC-SCIENCE listserv](#).

#### Background:

Funders of scientific research are increasingly attentive to the management of scientific data so that the full value of research investments can be realized and preserved. Doing so requires attention to the description and structure of datasets and to vocabularies for supporting data preservation, reuse, and repurposing.

The DCMI Science and Metadata Community is a forum for individuals and organizations to exchange information and knowledge about metadata describing scientific data (data methodologically collected for research, analysis, tracking, forecasting, and other uses). The Community focuses on metadata challenges specific to scientific data curation, and solutions that will benefit from the architecture and global reach of the Dublin Core Metadata Initiative.

The central challenges include:

- Canonical identification of datasets, critical for establishing provenance, auditing value and use, and attracting social-networking attention that will enhance their value.
- Better description of data and vocabularies, such that potential users may more easily determine suitability for use and repurposing, as well as ancillary applications for rendering and interpretation.
- Design and declaration of schemas to support reuse.

An initial deliverable of the group includes a survey of existing standards and metadata elements used to describe datasets, which will for



DCMI Science and Metadata Community

#### navigation

- [Main Page](#)
- [News](#)
- [People](#)
- [Publications \(private\)](#)
- [Standards](#)
- [Projects](#)
- [Research](#)

#### search

#### toolbox

- [What links here](#)
- [Related changes](#)
- [Special pages](#)
- [Printable version](#)
- [Permanent link](#)

## For Information

- Dryad, and Dryad wiki
  - <http://datadryad.org/>;
  - [https://www.nescent.org/wg\\_digitaldata/Main\\_Page](https://www.nescent.org/wg_digitaldata/Main_Page)
- HIVE wiki
  - <http://ils.unc.edu/mrc/hive/>
- Metadata Research Center <MRC>
  - <http://www.ils.unc.edu/mrc/>
- National Evolutionary Synthesis Center (NESCent)
  - <http://www.nescent.org/index.php>

## Publications (project wiki: [https://www.nescent.org/wg\\_dryad/Main\\_Page](https://www.nescent.org/wg_dryad/Main_Page))

- 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>
- 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).
- White, H., Carrier, C., Thompson, H., Greenberg, J., and Scherle, R. (2008). The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment. In DC-2008: Metadata for Semantic and Social Applications. *International Conference on Dublin Core and Metadata Applications*, 22-26 September, 2008, Berlin Germany, pp. 157-162.
- Carrier, S., Dube, J., and 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.
- Dube, J., Carrier, S., Greenberg, J., and White, H. (2008). Dryad: A Data Repository for Evolutionary Biology. In *Bulletin of IEEE Technical Committee on Digital Libraries*, (4) 1: <http://www.ieee-tcdl.org/Bulletin/v4n1/dube/dube.html>.
- Scherle, R., Carrier, S., Greenberg, J., Lapp, H., Thompson, A., Vision, T., and 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](http://pubs.or08.ecs.soton.ac.uk/35/1/submission_177.pdf).
- Dube, J., Carrier, S. and Greenberg, J. (2007). DRIADE: A Data Repository for Evolutionary Biology. In *Proceedings of the 2007 Conference on Digital Libraries*, Vancouver, British Columbia, Canada, June 18-23, 2007, pp. 481.