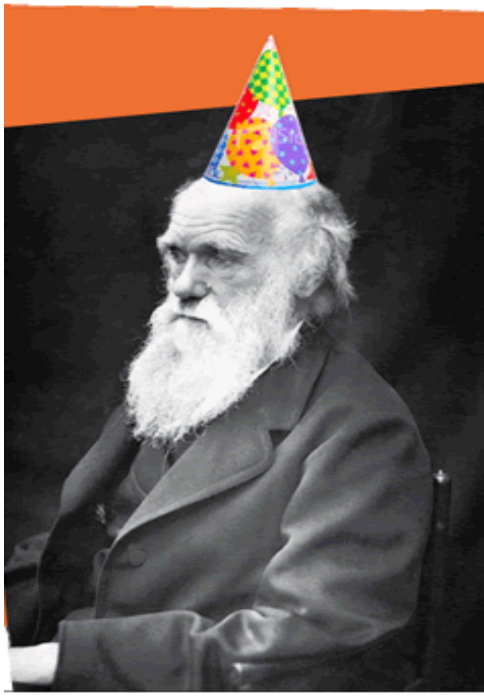
Metadata for Scientific Data Curation in the Dryad Repository

Special Libraries Association,
Annual Conference
Washington, D.C.
June 16, 2009



Jane Greenberg, Director, SILS Metadata Research Center <MRC>
School of Information and Library Science
University of North Carolina at Chapel Hill
janeg@email.unc.edu

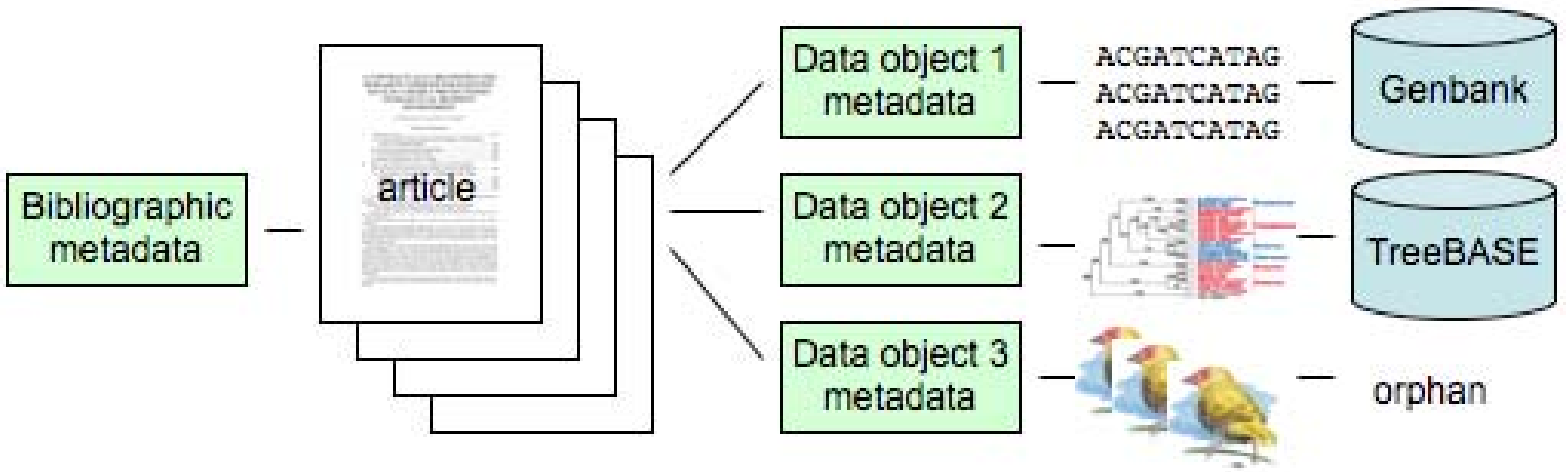
UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE
Metadata Research Center <MRC>

Overview

1. Introduce Dryad
   - Underlying motivation
   - Dublin Core Metadata Application Profile (DCAP)
   - Author curation
2. Metadata research arm
3. DCMI Science and Metadata (SAM)
4. Librarians and scientists working together

- Introducing Dryad….the underlying motivation

The published data package

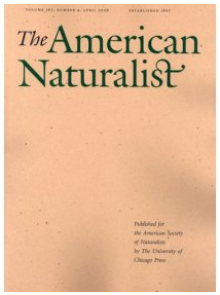


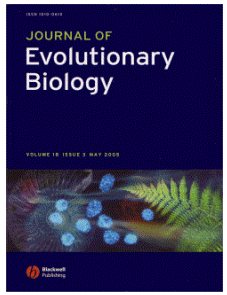
- Some data archived in specialized repositories
  - Genbank for gene sequences
  - TreeBASE for phylogenetic trees
- Most data *orphaned*
- *All* data should be available for
  - Validation of results
  - Data reuse, meta-analysis, synthesis...

Published works in evolutionary biology...convincing societies and journals

- ★ 27 papers from 5 different journals
- 41% had supplemental materials
- Genbank submission was generally honored
- 78% analyzed data not deposited in any repository
- 48% were based at least in part on data from other publications

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





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

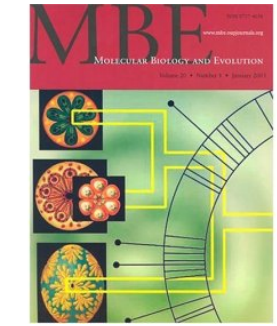
Society for Systematic Biology

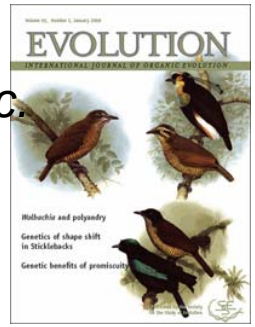
Systematic Biology

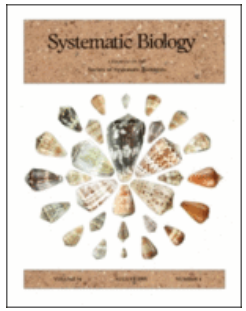
Commercial journals

Molecular Ecology

Molecular Phylogenetics and Evolution







Joint Data Archiving Policy

- DEPOSIT AT PUBLICATION
  - As a condition for publication, all data used in the paper should be archived in an appropriate public archive.
- REPEATABILITY
  - The data should be given with sufficient detail that, together with the contents of the paper, each result in the published paper may be re-created.
- EMBARGO
  - Authors may elect to have the data publicly available at time of publication, or, if the archive allows, may opt to embargo access to the data.
- EXCEPTIONS
  - Exceptions may be granted at the discretion of the editor, especially for sensitive information such as the location of endangered species.
- COORDINATION
  - The aim is for the consortium of journals to adopt this policy simultaneously.

Michael Whitlock, American Naturalist

Dryad's Goals

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

   ~ 220 data objects, 50 pubs; *American Naturalist, Evolution,…*

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

3. Balance a need for low barriers, with higher-level … data synthesis

Dryad Team

NESCent

- Todd Vision, Director of Informatics and Associate Professor, Biology, UNC
- Hilmar Lapp, Assistant Director of Informatics
- Ryan Scherle, Data Repository Architect

UNC/SILS/MRC

- Jane Greenberg, Professor, SILS
- Bob Losee, Professor, SILS
- Sarah Carrier, Doctoral Fellow
- Hollie White, Doctoral Fellow
- Amol Bapat, Master's student

Project Coordinator: Peggy Schaeffer, Coordinator/manager

- Dublin Core Metadata Application Profile (DCAP)

SUSTAINABILITY…

*...a partnership is well and good, but who is going to **curate**?*

...WHO is going to create the metadata? How? When...?

Dryad Metadata Application Profile to support author curation

Metadata development

- Dryad Metadata Application profile
  - Dublin Core: Interoperable, simple, user-generated metadata
  - Aligning with the Semantic Web
  - Basic data/metadata storage/retrieval

Modular scheme:	**Namespaces:**
1. Journal citation 2. Data objects (Carrier, et al., 2007; White, et al, 2008; Greenberg et al; in press *JLM*)	1. Dublin Core 2. Data Documentation Initiative (DDI) 3. Ecological Metadata Language (EML) 4. PREMIS 5. Darwin Core

<DRYAD application profile, ver. 1.0>

Bibliographic Citation Module

1. dcterms:bibliographicCitation/Citation information
2. DOI

Data Object Module

1. dc:creator/Name*
2. **dc:title/Data Set #**
3. dc:identifier/Data Set Identifier
4. PREMIS:fixity/(hidden)
5. dc:relation/DOI of Published Article
6. DDI:<depositr>/Depositor *
7. DDI:<contact>/Contact Info. #
8. dc:rights/Rights Statement
9. **dc:description/Description #**
10. 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

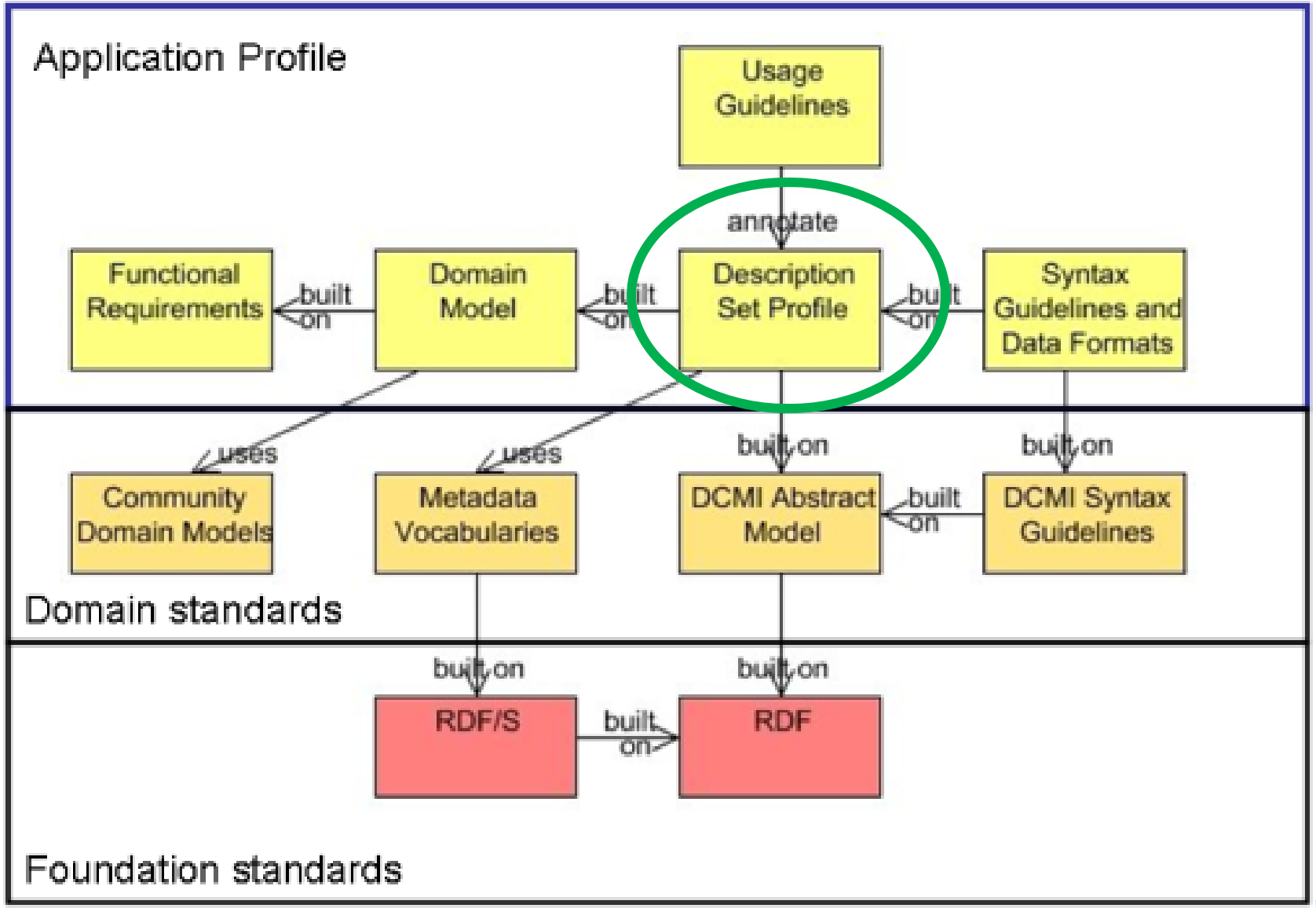
* = semi-automatic

= manual

Everything else is automatic

Singapore Framework Alignment

(machine processing, long term quality control Semantic web / linked data)



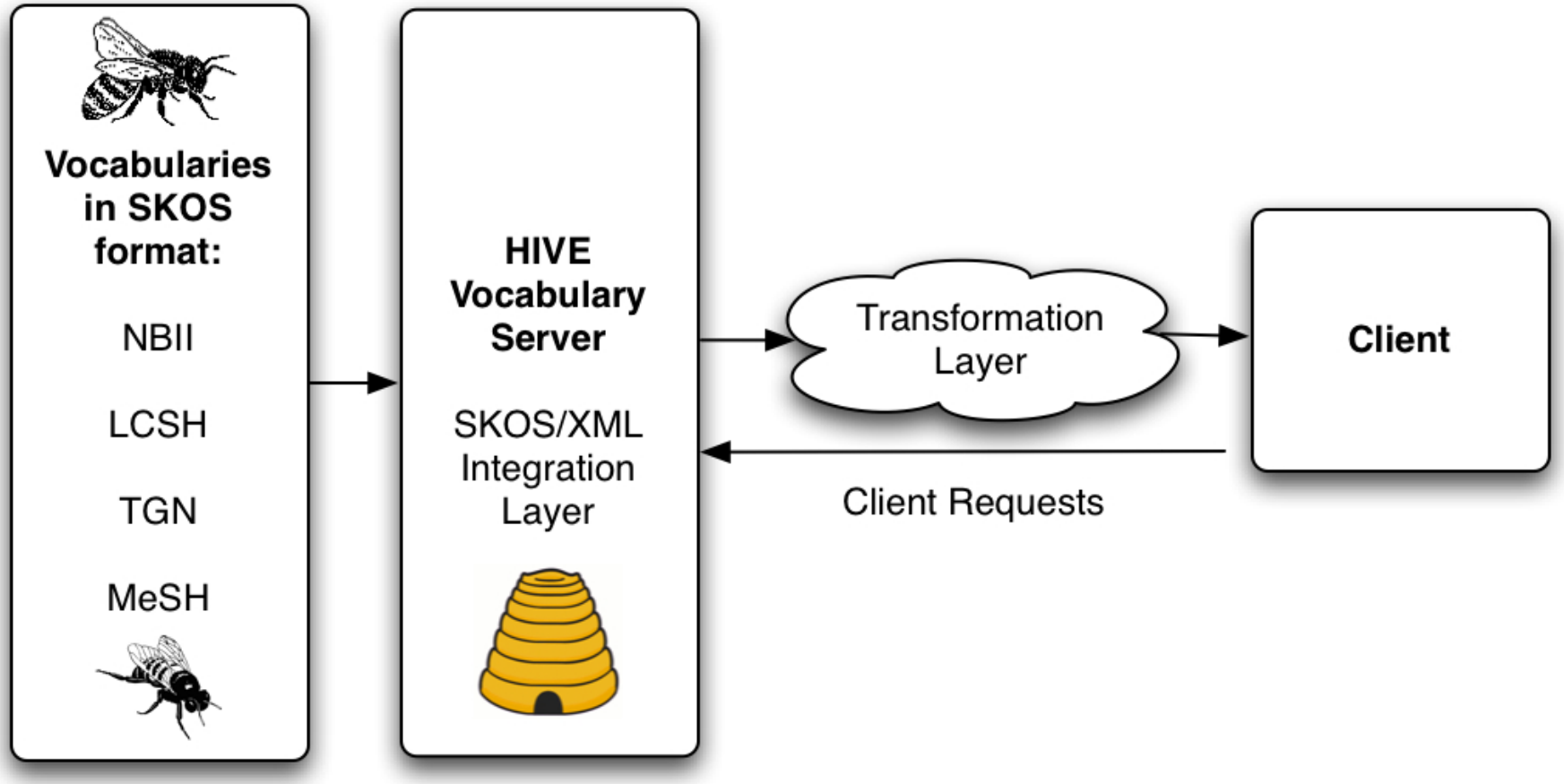
- http://dublincore.org/documents/singapore-framework

- Author curation

Author curation…

- Cost, only way this will work
- Low barriers, automation
  - Input template (**example**)
    - Pre-populate metadata template, metadata inheritence
    - Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption (Greenberg, 2009, CCQ, vol. 7, no. 3)
- HIVE (Helping Interdisciplinary Vocabulary Engineering)
  - LC, the Getty, USGS

HIVE model


Vocabularies in SKOS format:
NBII
LCSH
TGN
MeSH
HIVE Vocabulary Server
SKOS/XML Integration Layer
Transformation Layer
Client
Client Requests

Check all headings that apply to this publication.
To see broader/narrower terms, click the link for the respective vocabulary.

Beehner JC, Nguyen N, Alberts SC, Altmann J, 2006. The endocrinology of pregnancy and fetal loss in wild baboons. Hormones and Behavior 49:688-699.

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 2 months before the externally observed loss. By also considering an additional 450 pregnancies for which we did not have hormonal data, we determined that the probability for fetal loss for Amboseli baboons was 13.9%, and that fetal mortality occurred throughout gestation (91 losses occurred in 656 pregnancies; rates were the same for pregnancies with and without hormonal data). These results demonstrate that our longstanding method for early detection of pregnancies based on observation of external indicators closely matches hormonal identification of pregnancy in wild baboons.

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

- ☐ Abortion, Spontaneous [USE FOR Miscarriage] (MESH)
- ☐ Amboseli National Park (TGN)
- ☐ Baboon (Musical group) (LCSH)
- ☐ Baboon Creek (TGN)
- ☐ Baboons (LCSH)
- ☐ Estrogens (**NBII**, MESH)
  - ☐ Broader: Sex hormones
  - ☐ Narrower: Phytoestrogens
  - ☐ Related: Estrus
- ☐ Estrogens, Catechol (LCSH)
- ☐ Glucocorticoids (MESH, LCSH)
- ☐ Kenya (TGN)

UNC
SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Author curation…

- Quality control issue
  - **3 Levels**
  - Recognize some form of professional curation $------→ $$$$

- Metadata research arm

Metadata Research Arm

1. Metadata crosswalk analysis
2. Instantiation experiment
3. Vocabulary analysis (~ 600 concepts, 10 vocabularies)
4. Use-case study
5. Survey (~ 400 participants)
6. PIM Exploratory study (intensive interviews, scientists conception of KOS and metadata)

- DCMI Science and Metadata (SAM)

page | discussion | view source | history

Main Page

navigation
- Main Page
- News
- People
- Publications (private)
- Standards
- Projects
- Research

search
Go Search

toolbox
- What links here
- Related changes
- Special pages
- Printable version
- Permanent link

DCMI Science and Metadata Community

Dublin Core Metadata Initiative®
Making it easier to find information.

The DCMI Science and Metadata Community is a forum for individuals and organizations t 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 challenge 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 attent that will enhance their value.
- Better description of data and vocabularies, such that potential users may more easily determine suitability for use and repurposing, a 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 (SAM)

- Grew out of a WS meeting @ DC-2008/Berlin, Germany summary notes: http://ils.unc.edu/mrc/wp-content/uploads/2008/12/dc2008_mes_ws_summary.pdf
- 35 attendees, unanimous support for a community
- Community launched January 2009
  - ~ 200 members, international ...US,EUR,UK,Asia,Africa,Austalia,NZ,Mexico
  - Range of disciplines: biology, physics, chemistry, library/information science
  - Range of agencies: government, academic, industry, and corporations
  - Mailing list and wiki (Wiki at http://purl.org/dc/science)

Topics of discussion: Controlled vocabularies, ontologies, linked data, metadata registries, how to survey scientists, metadata quality

- Librarians and scientists working together

About the collaboration…

Pros, Benefits

- Synergy between implementation and research
- Broader familiarity with contacts & related projects (collective knowledge)
- Broader range of expertise for problem solving
- MRC: Contributing to a project that will benefit science and society
- A live lab, new research opportunities

Challenges

- Alignment of research and implementation goals (more immediate needs may not be the most interesting, vice/versa)
  - priorities
- Language barriers
- Funding models: Gap research and implementation
- Understanding…task assignment
- Not having everyone in the same building

Concluding remarks…

- Dryad offers an exciting model for making accessible data, and for library/scientist collaboration
- *Inspiration for…*DC-SAM, many others grappling with similar challenges, and opportunity for sharing solutions
- Small science has data curation needs, and metadata is an integral part of this
- Data curation matters to LIS educator/researchers, so we can help advance practice and train information professionals
- http://www.datadryad.org/repo/

Publications (project wiki: https://www.nescent.org/wg_dryad/Main_Page)

- 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)
- Greenberg, J. (2009). Theories of Evolution and Cultural Diffusion: The Dryad Repository Case Study for Understanding Changes in Organizing Information Practices. *iSociety: Research, Education, Engagement*. 2009 iConference, February, 8-11, Chapel Hill, North Carolina.
- 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
- 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.
- 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.