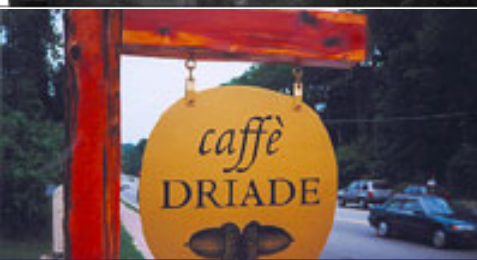


Darwin's 200th Anniversary



Digital Repository of Information and Data for Evolution



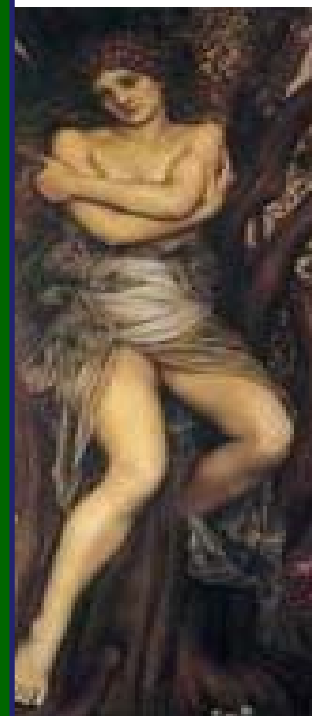
Metadata Research Supporting the Dryad Data Repository

Metadata Working Group
Cornell University Library

April 17, 2009

Jane Greenberg, Associate Professor
School of Information and Library Science
University of North Carolina

janeg@email.unc.edu



Overview

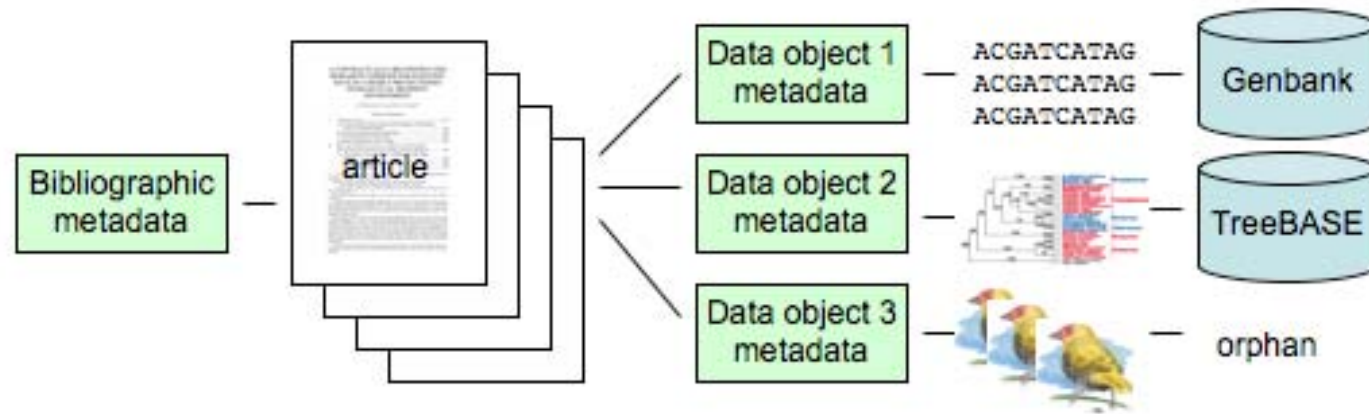
1. DRYAD: Motivation and Goals
 - Functional requirements
2. Metadata Research and Development
 - Application profile development, instantiation work, etc.
3. HIVE (Helping Interdisciplinary Vocabulary Engineering)
4. Q&A
5. DCMi Science and Metadata (SAM)



DRYAD: Motivation and Goals



The published data package

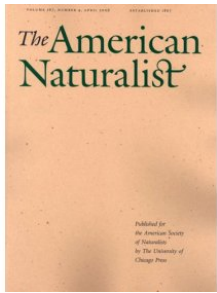


- Some data are archived in specialized repositories (e.g. Genbank)
- But most are *orphaned*
- *All* data should be available for
 - Validation of results
 - Development of improved methodologies
 - Meta-analysis
 - Synthesis

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

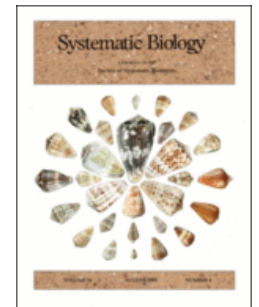
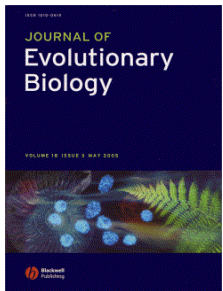
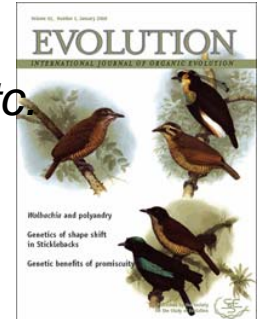
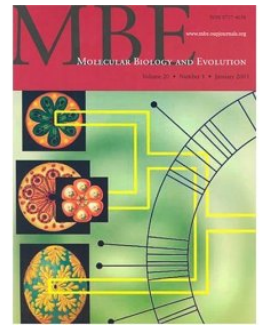
Society for Systematic Biology

Systematic Biology

Commercial journals

Molecular Ecology

Molecular Phylogenetics and Evolution



SCHOOL OF INFORMATION AND LIBRARY SCIENCE

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...

~ 200 data objects, 48 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

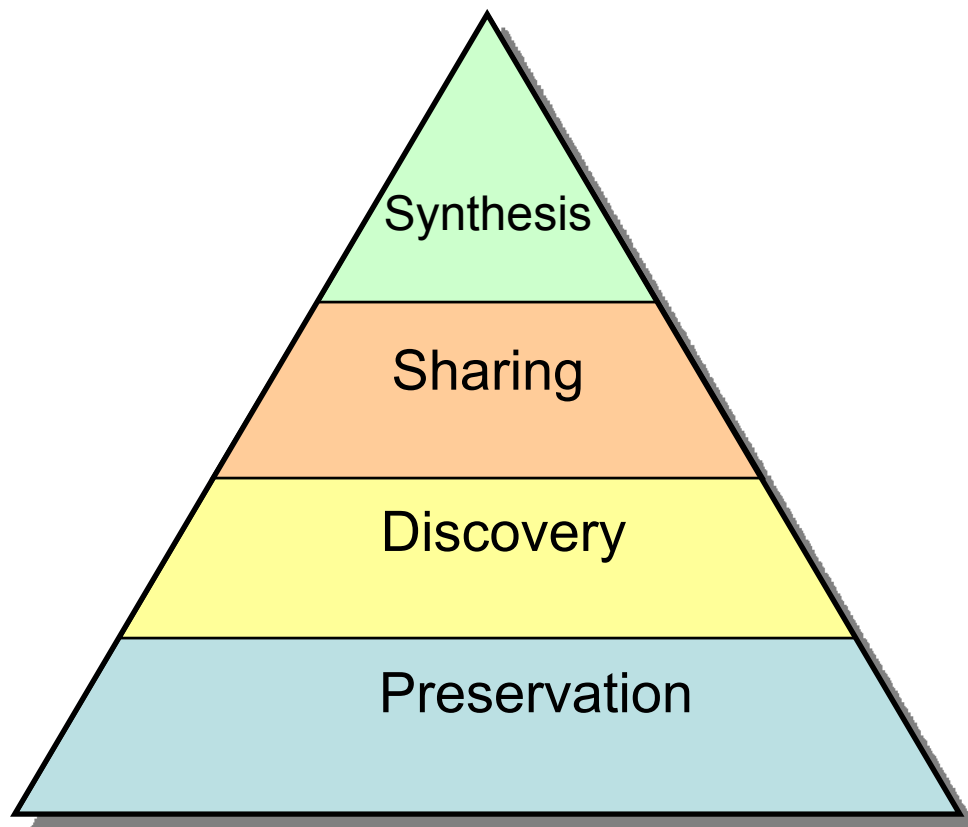
- Todd Vision, Director of Informatics and Associate Professor, Biology, UNC
- Hilmar Lapp, Assistant Director of Informatics
- Ryan Scherle, Data Repository Architect

UNC/SILS/MRC

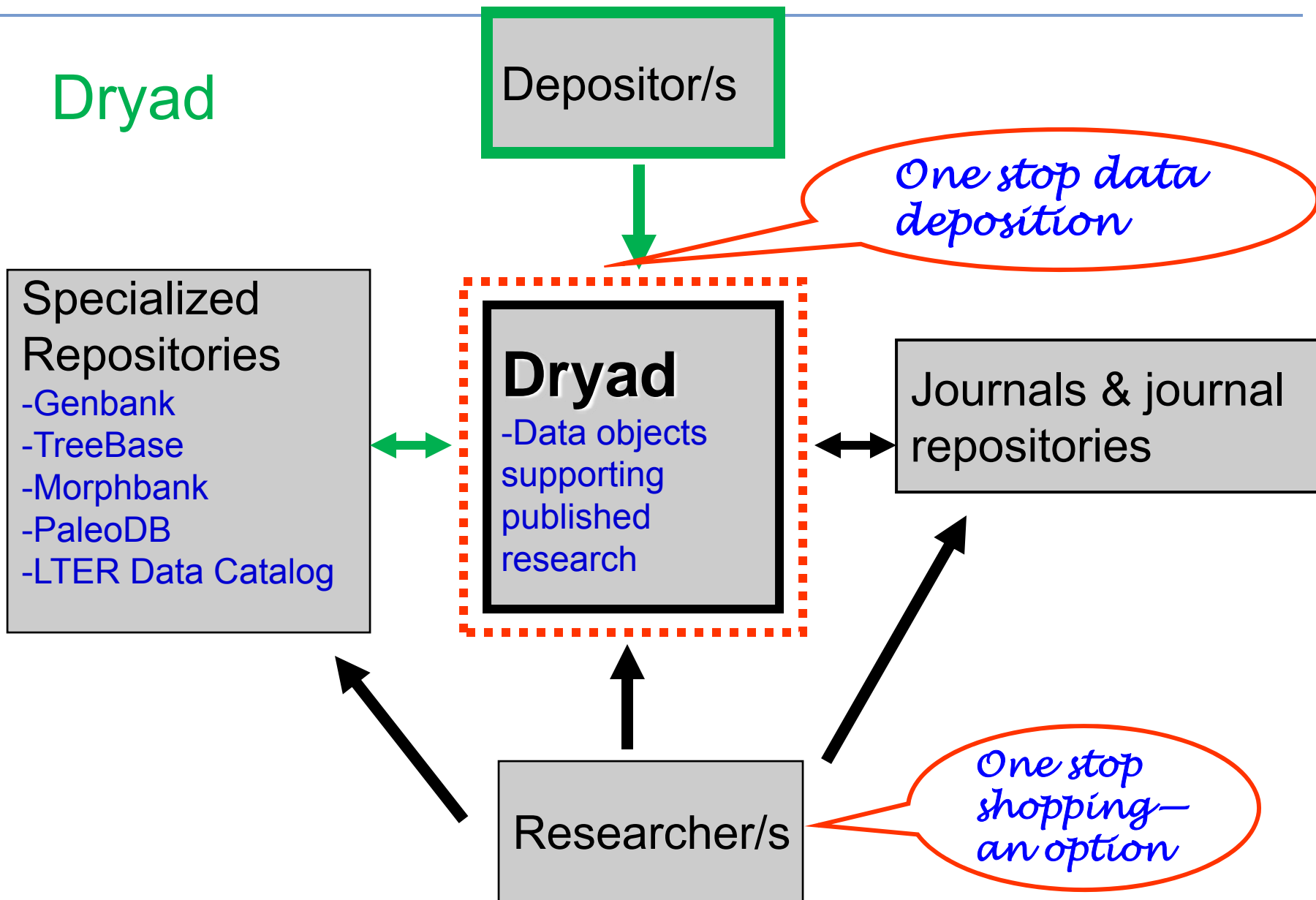
- Jane Greenberg, Associate Professor, SILS
- Bob Losee, Professor, SILS
- Sarah Carrier, Doctoral Fellow
- Hollie White, Doctoral Fellow
- Amol Bapat, Master's student

Project Coordinator: Peggy Schaeffer, Coordinator/manager

A hierarchy of goals



Dryad



Functional Requirements

Functional requirements



- Repository analysis (Dube, et al., JCDL, 2007)
- Workshops: Stakeholders (Dec. 06), SSR (May '07)
 - Resource discovery and use
 - Data interoperability
 - Automatic and semi-automatic metadata generation
 - Linking of publications and underlying datasets
 - Data/metadata quality control
 - Data security



Functional requirements

Project →	GBIF	KNB	NSDL	ICPSR	MMI
Goals/priorities ↓					
Heterogeneous digital datasets	■	■	■	■	■
Long-term data stewardship	■		■		
Tools and incentives to researchers	■	■	■	■	■
Minimize technical expertise and time required	■	■	■	■	■
Intellectual property rights	■	■		■	
Datasets coupled w/published research					

Metadata Application Profile Development



Metadata development

- Metadata architecture / **Application profile, ver. 1.0**
 - Interoperable with other schemes, why reinvent the wheel?
 - Dublin Core based
- Supports Dryad functionalities
 - Basic data/metadata storage
 - Simple retrieval and submission system

Modular scheme:

1. Journal citation
2. Data objects

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

Namespaces:

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
11. dc:coverage / Locality Required *
12. dc:coverage/Date Range Required*

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 *
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

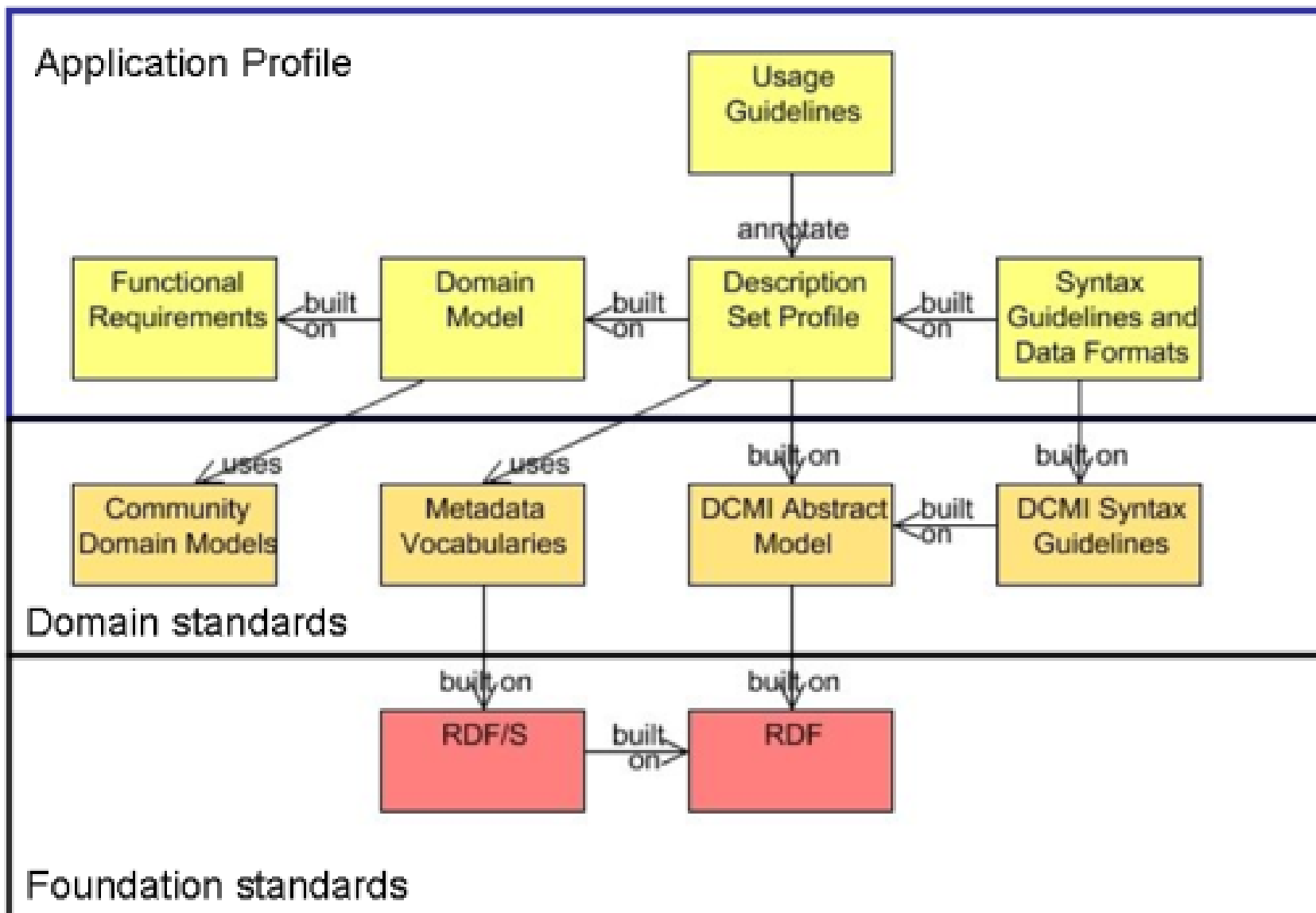
Everything else is automatic

Singapore Framework Compliant

(more recent efforts)

- Benefits
 - Consistency
 - Long-term quality control
 - **Interoperability** with other metadata structures
 - Aligns **w/Semantic Web** and **linked data** developments
 - Research arm of our project

- Use of Scholarly Works Application Profile (SWAP) as a key example of an application profile in conformance with the Singapore Framework



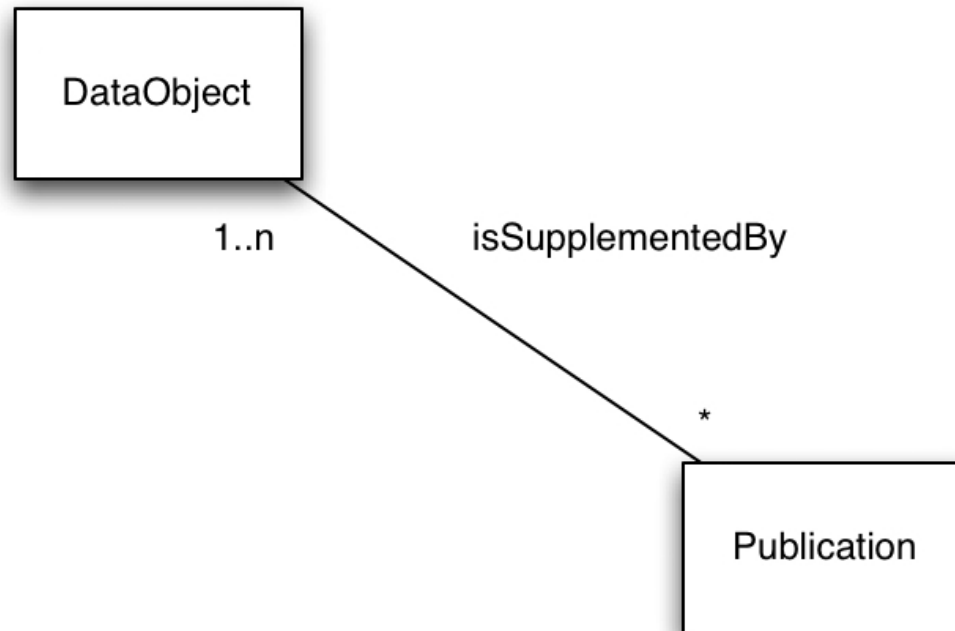
Singapore Framework Compliant

- 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)



Domain Model

- Dryad application profile version 1.0 accomodates one publication associated with multiple datasets



Description Set Profile and Usage Guidelines

- 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>

- Usage guidelines are optional

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

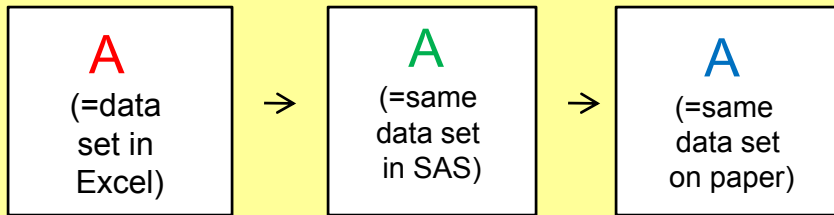
- Infrastructure not all there... (a lot is not in RDF)
 - Registered Dryad “purl”
- Proof of concept difficult
- Time consuming
- Documentation lacking

Instantiation (works)

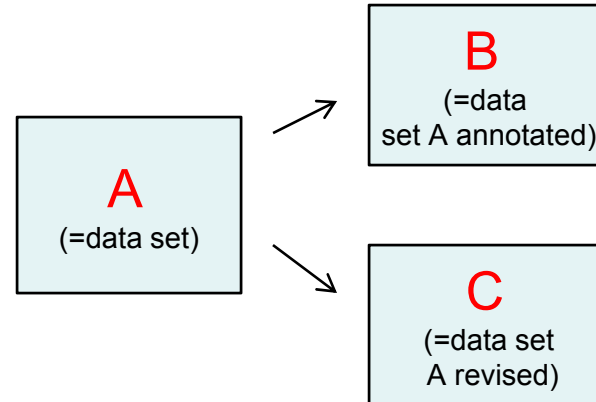
- Instantiation study
 - Bibliographic relationships for life-cycle management (Tillett, 1992, 1992; Smiraglia, 1999, 2000+.; Coleman, 2002; FRBR, DCAM)

Data object relationships

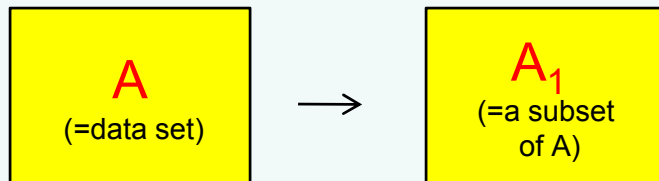
Equivalence



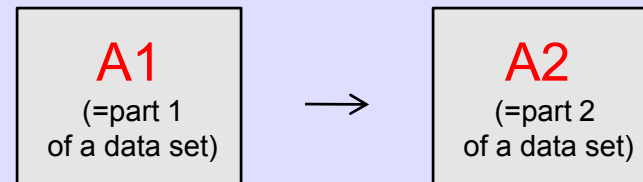
Derivative



Whole-part



Sequential



Instantiation

Scenario: Sherry collects data on the survival and growth of the plant *Borrichia frutescens* (the bushy seaside tansy)... back at the lab she enters the exact same data into an excel spreadsheet and saves it on her hard drive.

Question: What is the relationship between Sherry's paper data sheet and her excel spreadsheet?

Answer: Equivalent | Derivative | Whole-part | Sequential
(circle one)

Findings (20 participants)

- In general, more seasoned scientists better grasp
- Sequential data presented the most difficulty (less seasoned sci.)
- Unanimous support: “very → extremely important”

Use Case, Survey, and PIM research



Use case, survey, and PIM research

- **Use-case study** (Sarah Carrier)
Intensive interviews about data sharing
- **Survey** (*team*)
International survey, launched via evoldir
~ 400 respondents

Do you think the data underlying published scientific results should be made publicly accessible?"

» Yes: 395 (95.4%)

» No: 19 (4.6%)

"If yes, do you think journals should require data sharing of their authors, or should it be voluntary?"

» Required: 220 (55.6%)

» Voluntary: 176 (44.4%)

- **PIM Exploratory study** (Hollie White)

Helping Interdisciplinary Vocabulary Engineering (HIVE)



Helping Interdisciplinary Vocabulary Engineering (HIVE)

- Vocabulary analysis --- **HIVE!**

NBII Thesaurus, LCSH, the Getty's TGN, Gene Ontology

- 600 keywords, Dryad partner journals

Facets: **taxon, geographic name, time period, topic**

- 431 topical terms, exact matches

NBII Thesaurus, 25%; MeSH, 18%

- 531 terms

LCSH, 22% found exact matches, 25% partial

Need multiple vocabularies + which vocabularies



HIVE (Helping Interdisciplinary Vocabulary Engineering)

- Automatic metadata generation approach that dynamically integrates discipline-specific controlled vocabularies encoded with the [Simple Knowledge Organisation System \(SKOS\)](#)
- *provide efficient, affordable, interoperable, and user friendly access to multiple vocabularies during metadata creation activities*
- *Building HIVE*
 - *Vocabulary Development*
 - *Server preparation*
 - Primate Life Histories Working Group
 - Wood Anatomy and Wood Density Working Group
- *Sharing HIVE*
continuing education
- *Evaluating HIVE*
examining HIVE in Dryad

HIVE Partners

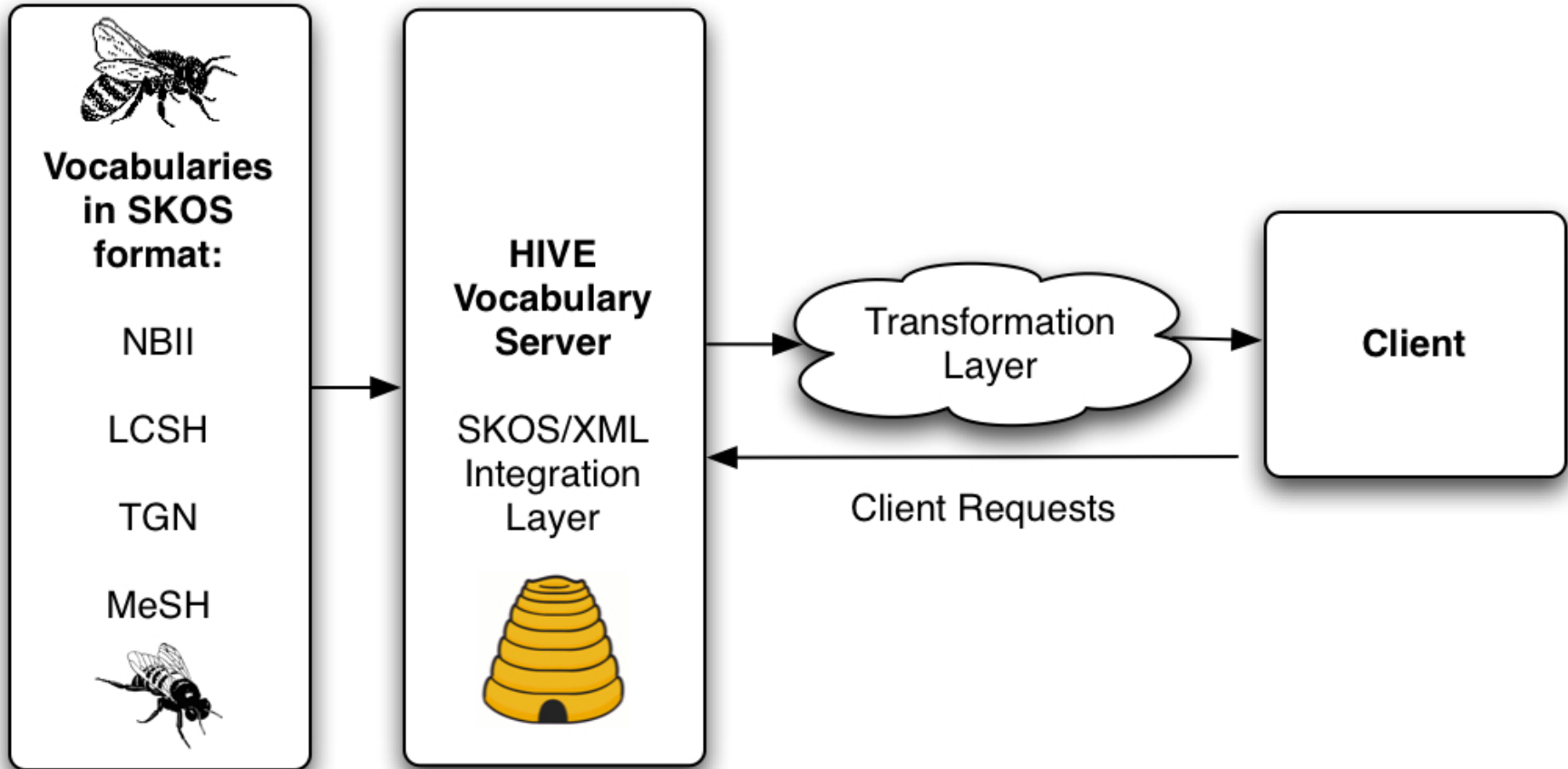
Vocabulary Partners

- 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
- 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
- Barbara Tillett, Library of Congress
- Kathy Wisser, UNC Chapel Hill
- Lisa Zolly, USGS

HIVE model



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](#))



National Science Foundation
WHERE DISCOVERIES BEGIN

- Dryad
 - <http://datadryad.org/>
- Dryad Wiki
 - https://www.nescent.org/wg_digitaldata/Main_Page
 - Includes links to publications, the application profile, and lists Dryad team members
- Metadata Research Center <MRC>
 - <http://www.ils.unc.edu/mrc/>
- National Evolutionary Synthesis Center (NESCent)
 - <http://www.nescent.org/index.php>

Exciting, competing challenges, headway, R&D, open

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

- Greenberg, J. (2009, in press). 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: <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.
- 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.

Main Page

DCMI Science and Metadata Community



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](#)

Dublin Core Metadata Initiative (DCMI) Science and Metadata (SAM) community

(<http://dublincore.org/groups/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 formally launched January 2009
- About 170 members
 - Range of disciplines: biology, physics, chemistry
 - Range of agencies: government, academic, industry, and corporations
- Mailing list and wiki (Wiki at <http://purl.org/dc/science>)
- Deciding now about DC2009/separate WS/issues, NSF DataNet, data federation