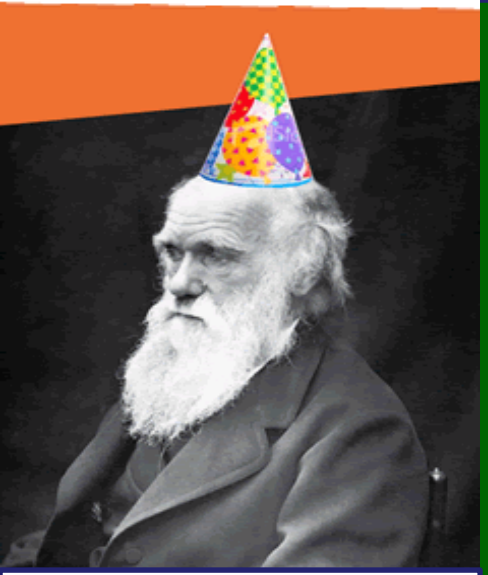


Darwin's 200th anniversary

Dig.Rep.of info.+data for Evo.

The Dryad Repository Application Profile: Groundwork Towards a Metadata Scheme for Scientific Data

DigCCurr 2009

April 2, 2009

Chapel Hill, North Carolina

- Jane Greenberg, Sarah Carrier, Hollie White , University of North Carolina
- Ryan Scherle, NESCent

DRYAD

National Evolutionary Synthesis Center
NESCent

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>

Overview

- DRYAD: Motivation and Goals
- Dryad Research and Development
  - Functional requirements
  - Metadata activities

    - Application profile development

    - HIVE – Helping Interdisciplinary Engineering
- Digital Curation Curriculum
- Q&A

DRYAD: Motivation and Goals

Motivation for Dryad

- Small science repositories (SSR)
  - Knowledge Network for Biocomplexity (KNB)
  - Marine Metadata Initiative (MMI)
- Evolutionary biology
  - Publication process

    Supplementary data (*Evolution, Amer. Nat'l*)

    "Author," "deposition date," not "subject" "species," "geo. locator"

    Data deposition (Genbank, TreeBase, Morphbank)
- NESCent & SILS/Metadata Research Center
  - NC State, Univ. of New Mexico, and Yale

ecology, paleontology, population genetics, physiology, systematics + genomics

Dryad's Goals

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

   ~ 180 data objects, 40 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

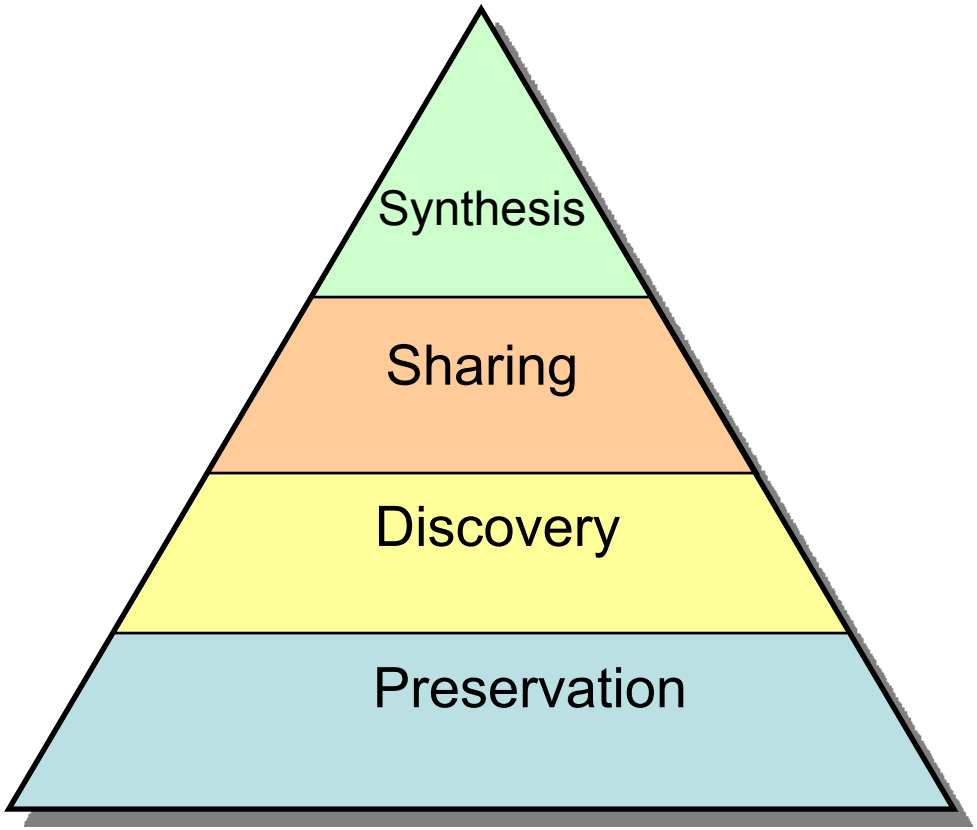
- Todd Vision, Director of Informatics and Associate Professor, Biology, UNC
- Hilmar Lapp, Assistant Director of Informatics
- Ryan Scherle, Data Repository Architect

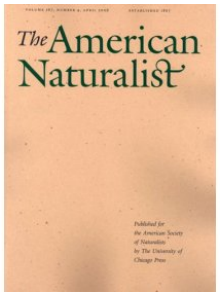
UNC/SILS/MRC

- 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





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

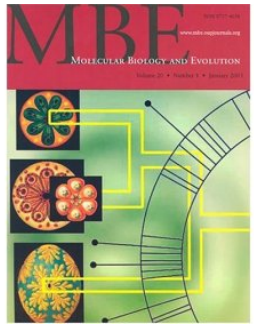
Society for Systematic Biology

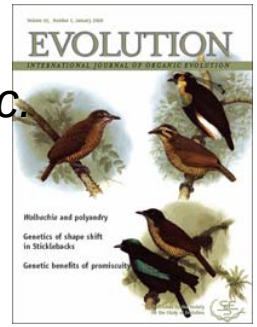
Systematic Biology

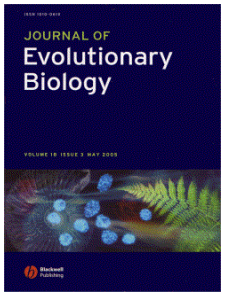
Commercial journals

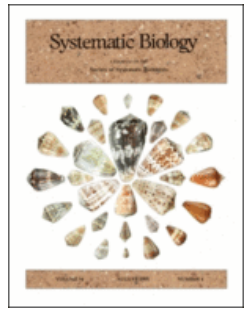
Molecular Ecology

Molecular Phylogenetics and Evolution





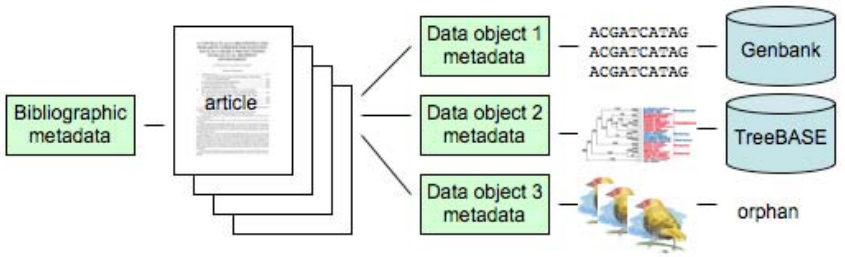




- Dryad Research and Development
  - Functional requirements
  - **Application profile development**
  - Vocabulary analysis
  - Instantiation study
  - HIVE – Helping Interdisciplinary Engineering

R & D: Accomplishments and Activities

- 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→ Goals/priorities↓	GBIF	KNB	NSDL	ICPSR	MMI
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 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:	Namespaces:
1. Journal citation 2. Data objects (Carrier, et al., 2007)	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 Compliant

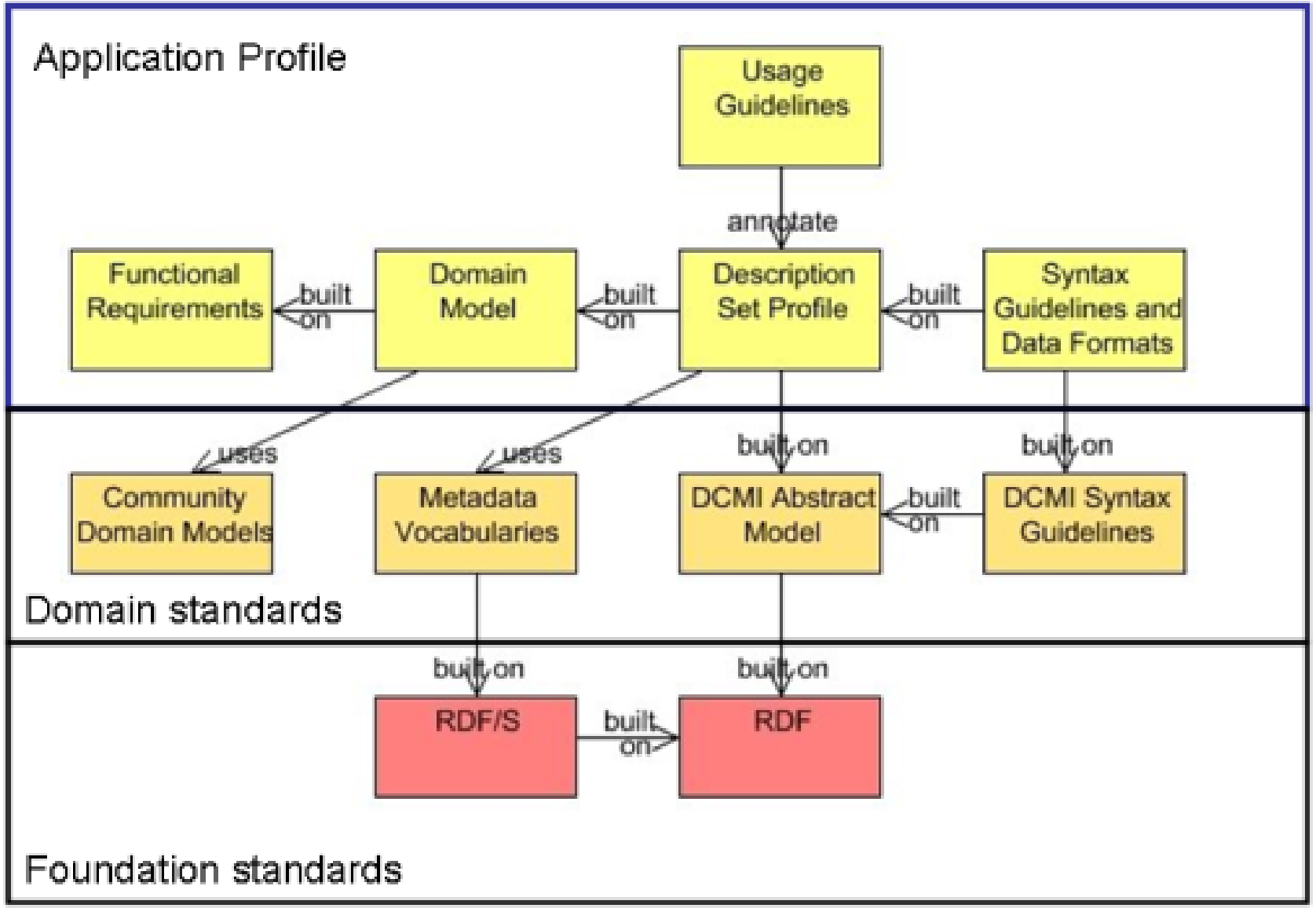
- A “loose” standard for Dublin Core “endorsed” application profiles
- Singapore framework provides guidelines for creating a DCAM-conformant Application Profile (“DC Application Profile”)
- A packet of documentation which consists of:
  1. Functional requirements (desirable)
  2. Domain model (mandatory)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

Singapore Framework

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

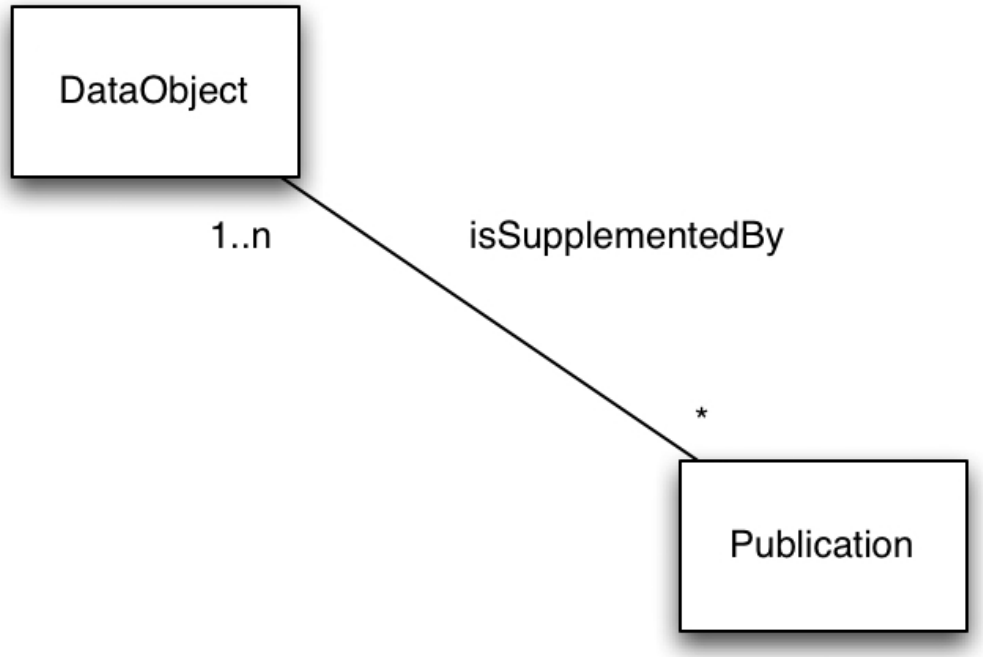
- Use of Scholarly Works Application Profile (SWAP) as a key example of an application profile in conformance with the Singapore Framework

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



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
  - https://www.nescent.org/wg_digitaldata/Dryad_Level_One_Cataloging_Guidelines

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

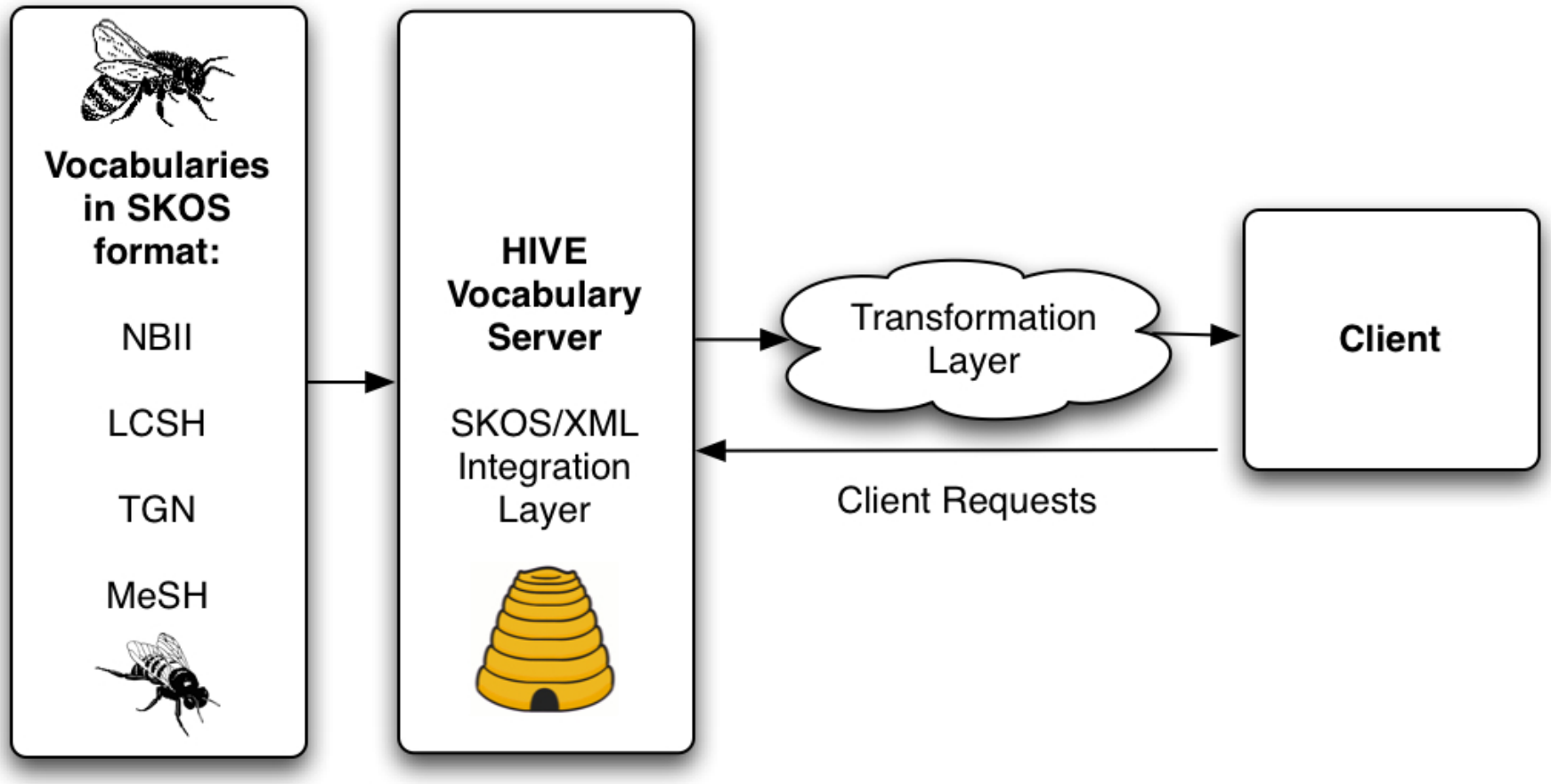
- Challenges
- Infrastructure not all there… (a lot is not in RDF)
  - Registered Dryad "purl"
- Proof of concept difficult
- Time consuming
- Documentation lacking

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 model


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

Digital Curation Curriculum

- UNC is a great place!!
- Metadata is key for digital curation, and an important part of our curriculum
- Experiential learning
  - Collaboration
  - Interdisciplinary team
  - Research

- Challenges, language, balancing priorities…

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

- 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

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