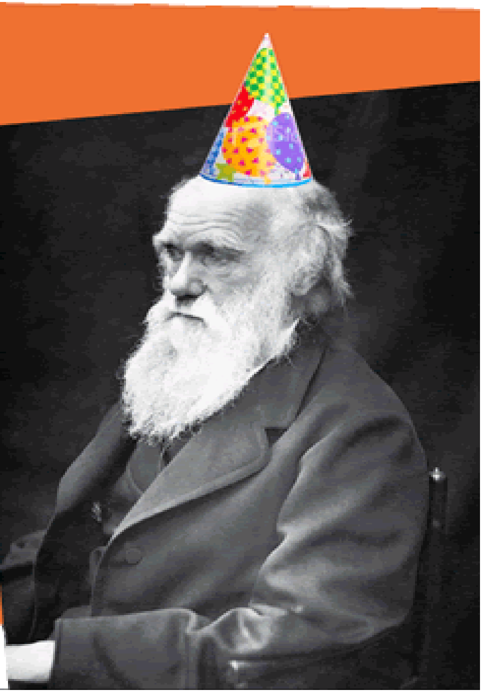


Celebrating Darwin Day + 1: Metadata Research and the Dryad Repository Project

Sarah Carrier
Jane Greenberg
Hollie White

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

Overview

- DRYAD: Motivation and Goals
- Dryad Research and Development
    - Functional requirements
    - Application profile development
    - Vocabulary analysis
    - Instantiation study
    - Briefly… a few items…
    - PIM and KO
    - HIVE – Helping Interdisciplinary Engineering
- Q&A

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

NESCent

National Evolutionary Synthesis Center

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>

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

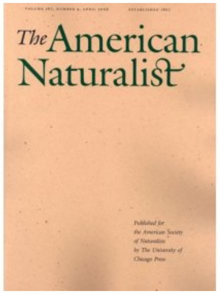
NESCent

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

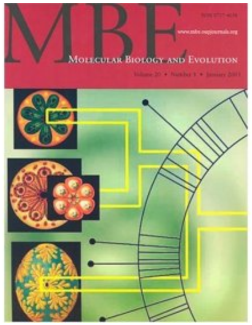
UNC/SILS/MRC

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

New vital link: Peggy Schaeffer, Coordinator/manager

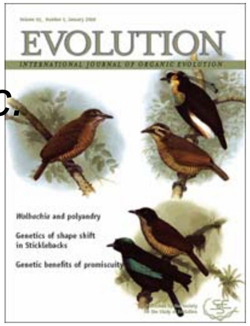


Partner Journals



American Society of Naturalists

American Naturalist



Ecological Society of America

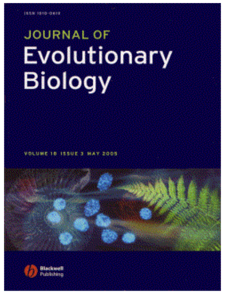
Ecology, Ecological Letters, Ecological Monographs, etc.

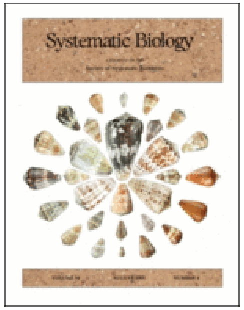
European Society for Evolutionary Biology

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

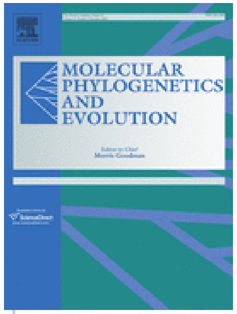
Society for Systematic Biology

Systematic Biology

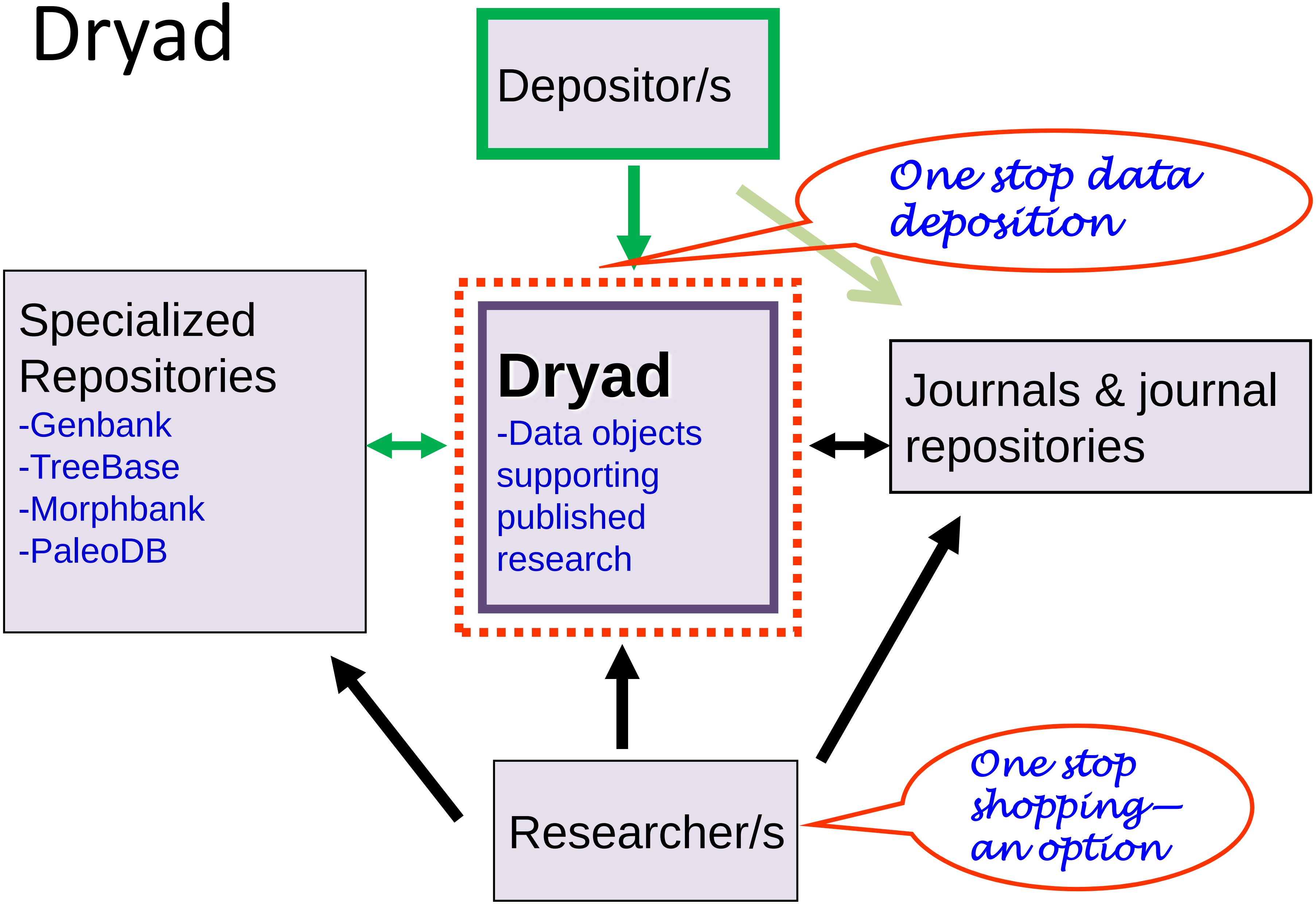
Commercial journals

Molecular Ecology

Molecular Phylogenetics and Evolution



Dryad



Depositor/s

One stop data deposition

Specialized Repositories
-Genbank
-TreeBase
-Morphbank
-PaleoDB

Dryad
-Data objects supporting published research

Journals & journal repositories

Researcher/s

One stop shopping— an option

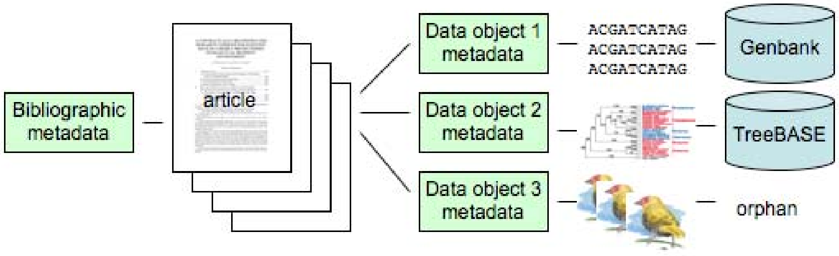
Research and Development

- Functional requirements
- Application profile development
- Vocabulary analysis
- Instantiation study
- Briefly… a few items…
- PIM and KO
- 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)

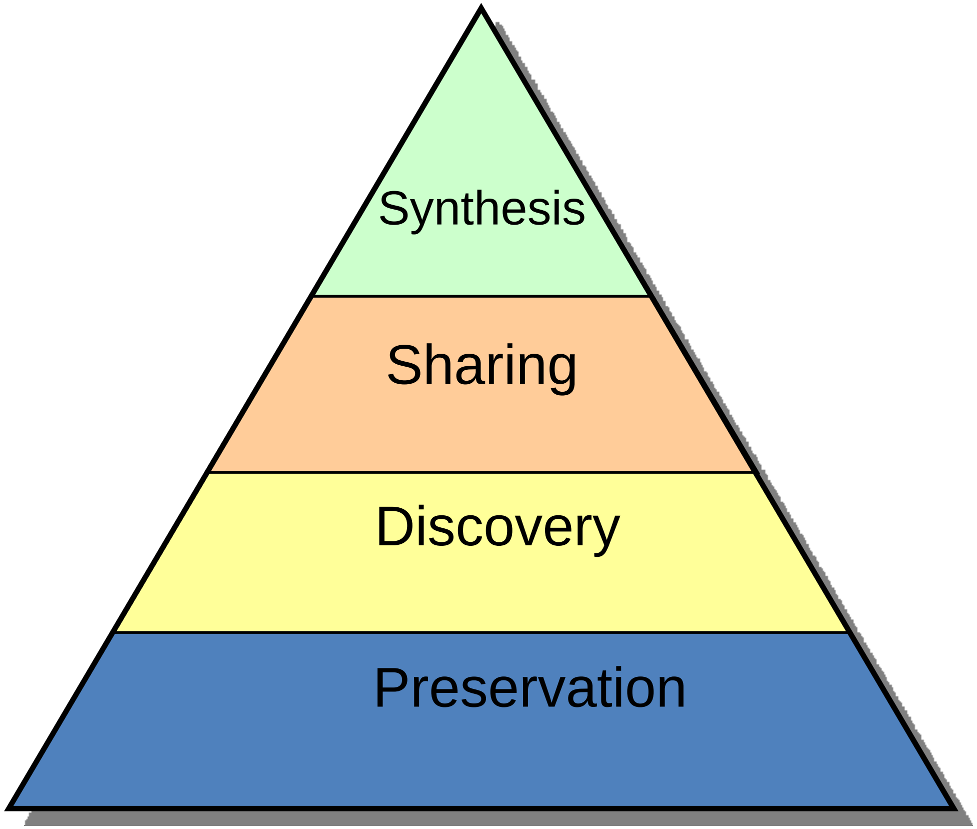
- computer-aided metadata generation and augmentation
- linking data submission and publication
- support for identity, authority and data security
- support for basic metadata repository functions, such as resource discovery, sharing, and interoperability



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					

A hierarchy of goals



R & D: Accomplishments and Activities

- Metadata architecture / Application profile, ver. 1.0

Modular scheme:

1. **Journal citation**
2. **Data objects**

(Carrier, et al., 2007)

Namespace schemas:

1. **Dublin Core**
2. **Data Documentation Initiative (DDI)**
3. **Ecological Metadata Language (EML)**
4. **PREMIS**
5. **Darwin Core**

<DRIADE application profile, version 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

R & D: Accomplishments and Activities

- 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

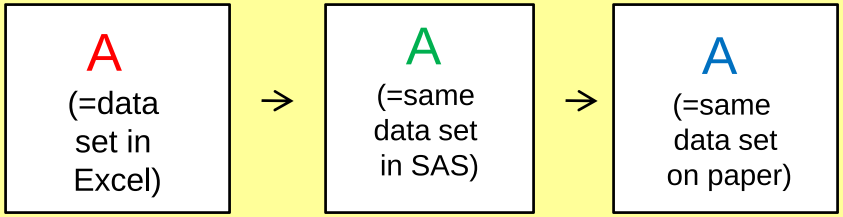
    LCSH, 22% found exact matches, 25% partial

    Need multiple vocabularies + which vocabularies

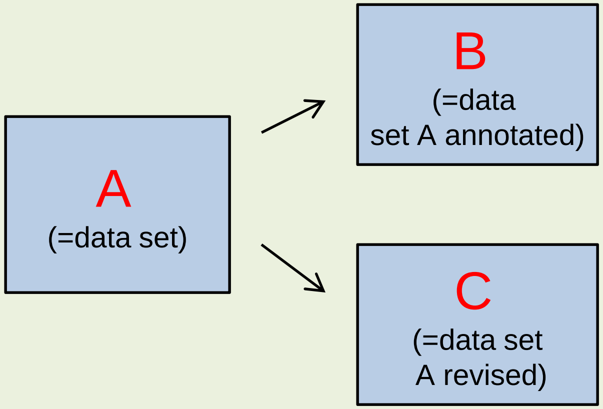
- Instantiation study
  - Bibliographic relationships for life-cycle management (Tillett, 1992, 1992; Smiraglia, 1999, 2000+.; Coleman, 2002; FRBR, DCAM)

Data object relationships

Equivalence


A (=data set in Excel) → A (=same data set in SAS) → A (=same data set on paper)

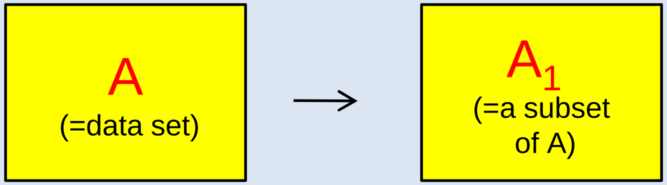
Derivative


A (=data set)

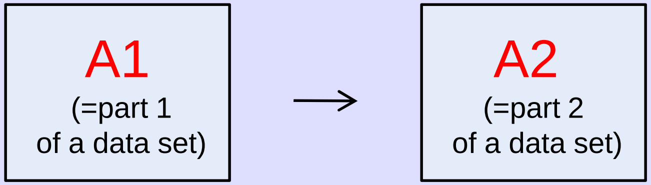
B (=data set A annotated)

C (=data set A revised)

Whole-part


A (=data set) → A_1 (=a subset of A)

Sequential


A1 (=part 1 of a data set) → A2 (=part 2 of a data set)

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"

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>

R & D: Accomplishments and Activities

- Use-case study (Sarah Carrier)

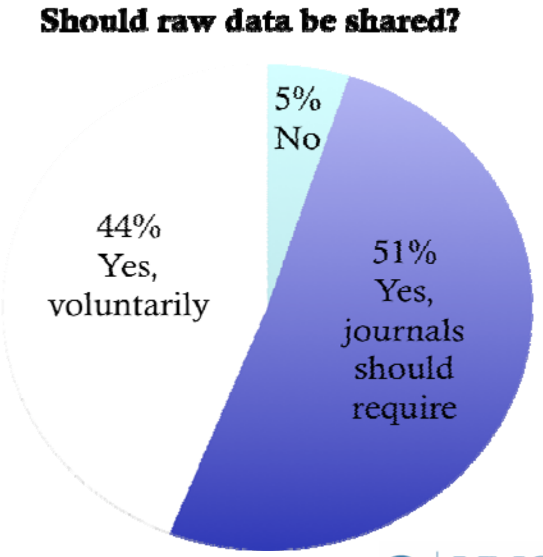
  Intensive interviews about data sharing

- Survey (*team*)

  International survey, launched via evoldir

  ~ 400 respondents

- PIM Exploratory study
  - (Hollie White)



The Dryad Repository: Where PIM and KO meet

Interviews

- **Method:** Exploratory, ethnographically-inspired, free-flowing interviews.

- **Interview Length:** 15 minutes to 1 hour and 25 minutes.

- **Interview Focus:** Interviews addressed the following topics:

  - type of data collected
  - organizational style and motivation
  - perception of sub-domain organizes trends
  - organizational style preference and rational underlying that preference

Participants

- **Participant Description:** 7 Evolutionary Biologists
  - 5 male and 2 female
  - lab and field foci
  - various age and experience levels
  - all have published works

- **Sub-Domains Represented:**
  - botany
  - genetics
  - palentology
  - mammology
  - entomology

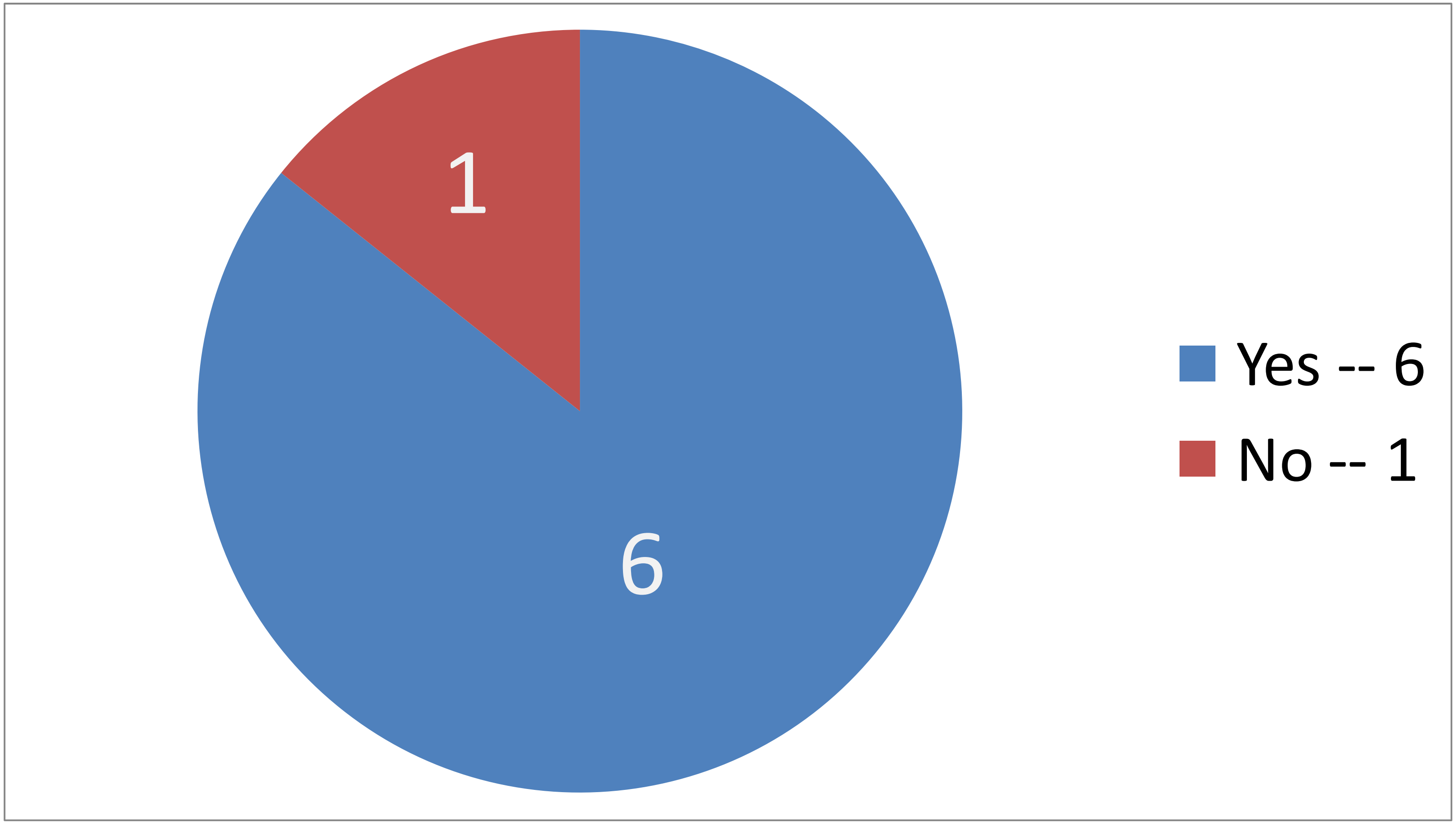
What do scientists collect?

Or what makes up a dataset?

Lots of electronic, but some is non-electronic.

- GPS data
- Gene Sequences
- Herbarium samples
- Fossils
- Mammal life histories
- Photographs
- Insect Measurements

Do scientists use metadata for their own research?



Finding a few answers:

- Are there trends in how scientists organize their data?
- Why do scientists organize their data the way they do?
- What do scientists think about the way data is organized?

More Questions to look into

- Just how “personal” is research data?
- What are the differences/similarities in the way scientists and information professionals organize research/scientific data?
- Where does personal organization end and knowledge organization begin?

HIVE

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

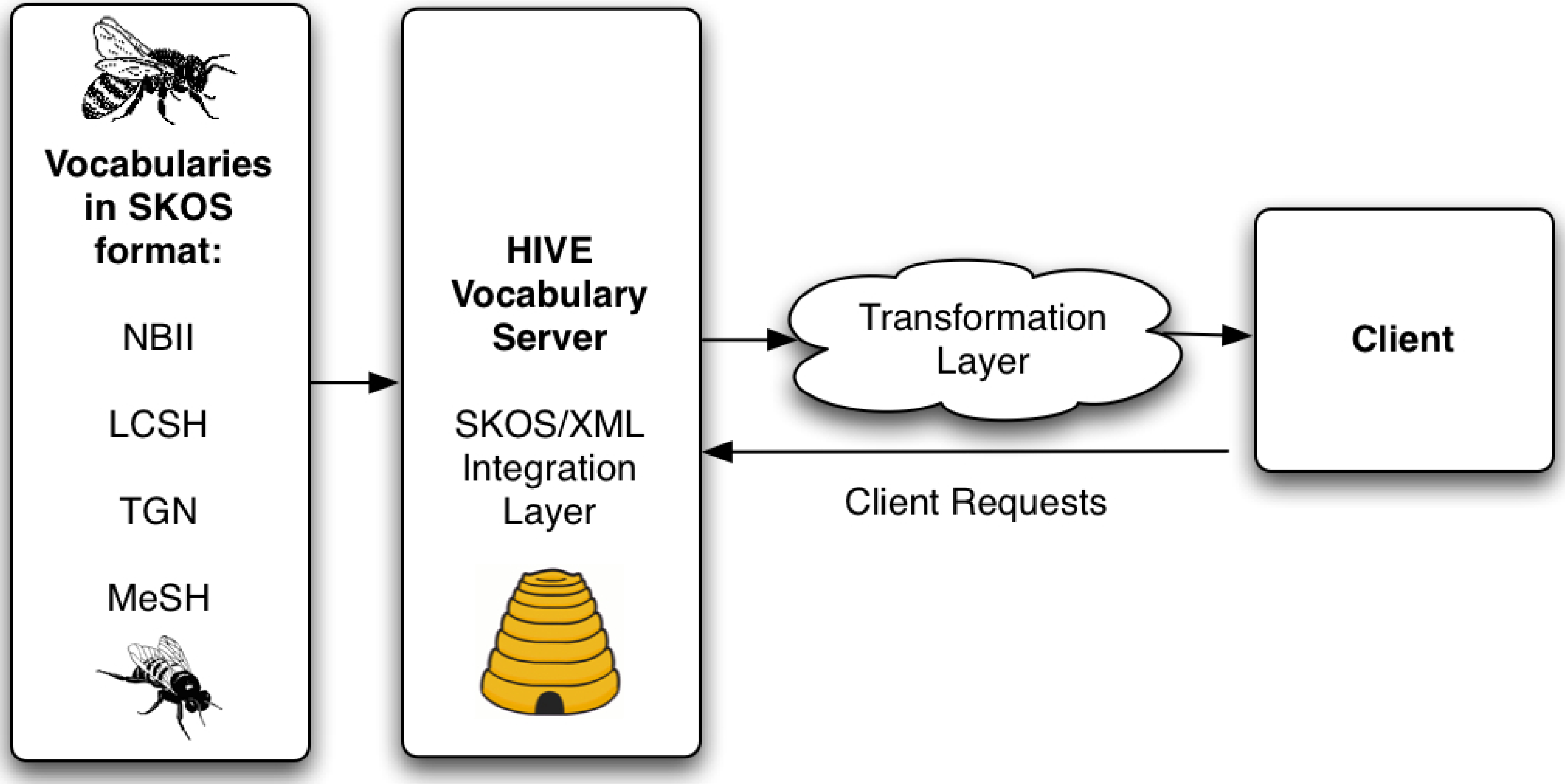
Vocabulary Partners

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

Advisory Board

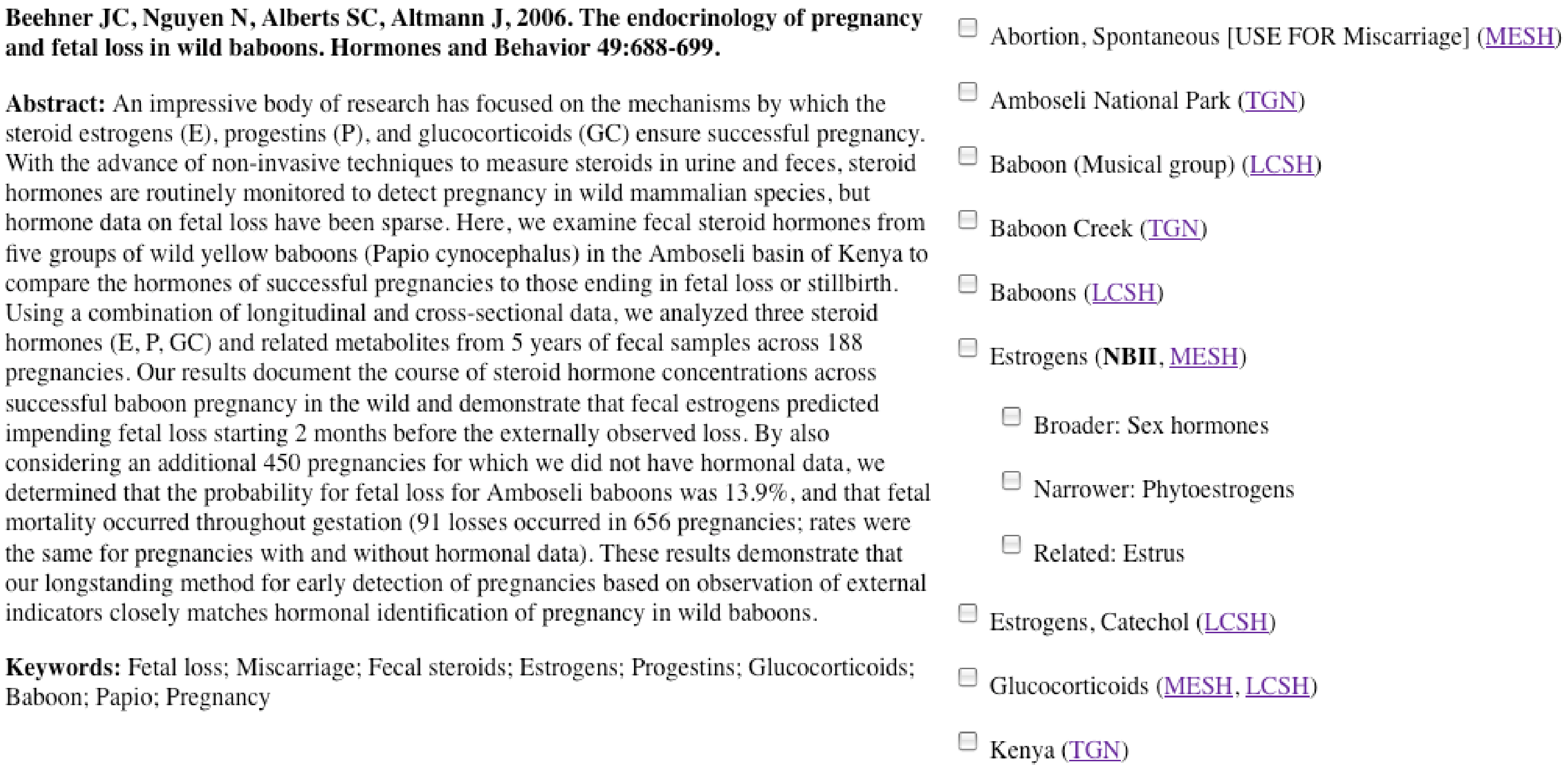
- Jim Balhoff, NESCent
- Mike Frame, USGS
- Alistair Miles, CCLRC Rutherford Appleton Laboratory
- William Moen, University of North Texas
- Eva Méndez Rodríguez, University Carlos III of Madrid
- Joseph Shubitowski, Getty Research Institute
- Barbara Tillett, Library of Congress
- Kathy Wisser, UNC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

HIVE model


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

HIVE

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
Metadata Research Center <MRC>

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.