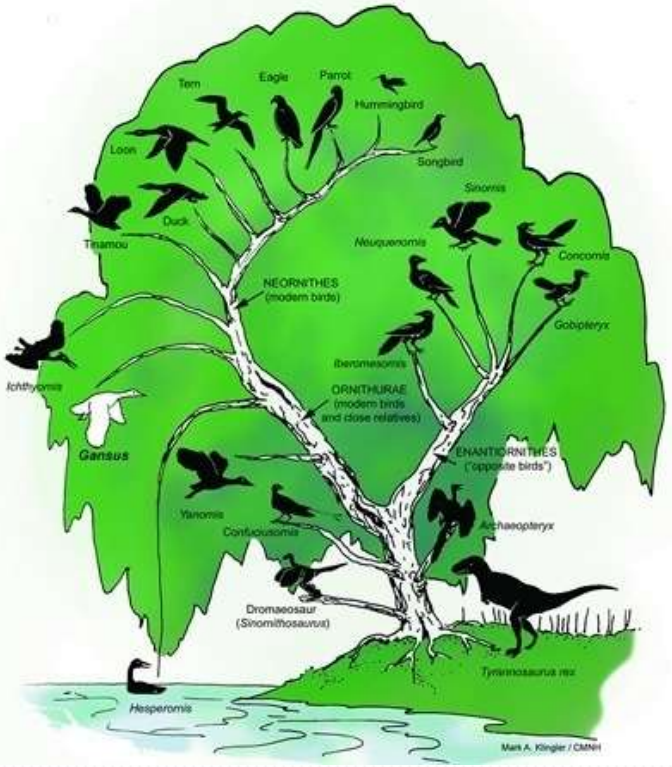
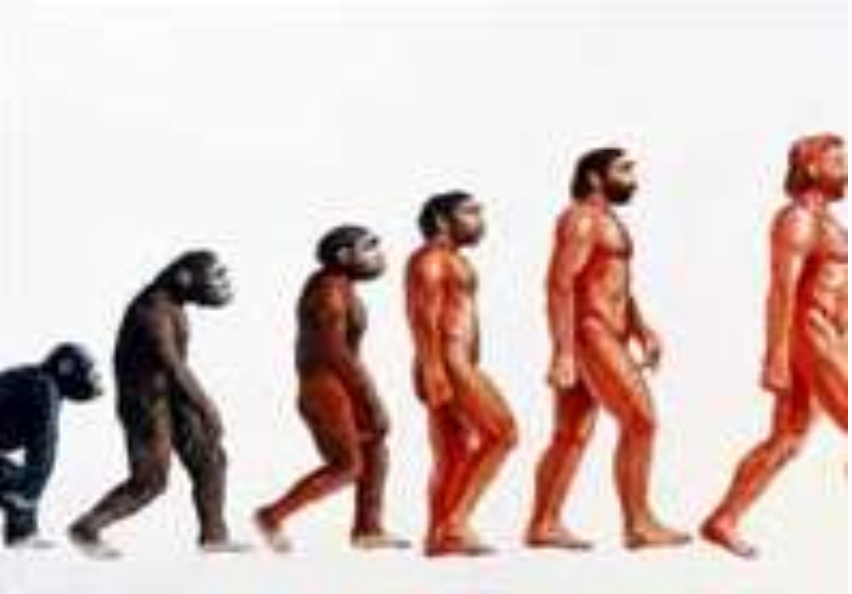
The DRYAD Repository

~~~~~~

Librarians and e-Science:
Focusing Toward 20/20

CIC 2008: May 12





Jane Greenberg
SILS/Metadata Research Center
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
janeg@email.unc.edu
~~~~~~

Overview

- DRYAD
  - Formerly: DRIADE – (Digital Repository of Information and Data for Evolution)
- NESCent / SILS Metadata Research Center <MRC> collaboration
  - Research
- CIC context
- Conclusions

Motivation for Dryad

- Small science repositories (SSR)
  - Knowledge Network for Biocomplexity (KNB), Marine Metadata Initiative (MMI)
- Evolutionary biology → ecology, paleontology, population genetics, physiology, systematics + genomics
  - Publication process
    - Supplementary data (*Evolution, American Naturalists)*

      "Author," "deposition date," not "subject" "species," "geo. locator"
    - Data deposition (Genbank, TreeBase, Morphbank)
- NESCent & SILS/Metadata Research Center

Dryad's Goals

1. One-stop deposition and shopping for data objects supporting published research…
   - 108 data objects, 23 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

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

UNC/SILS/MRC

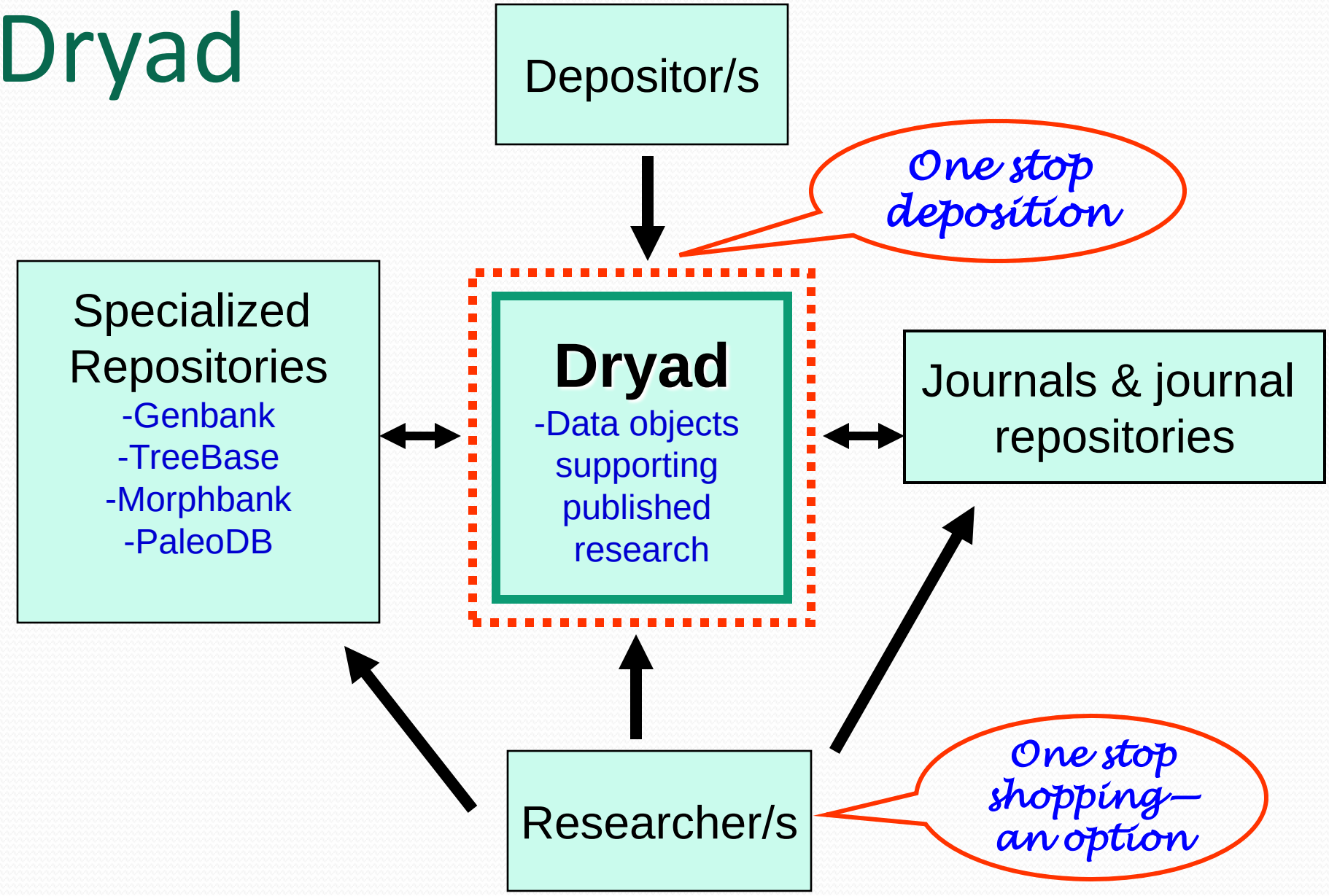
- PI: Jane Greenberg, Associate Professor, SILS and MRC
- Sarah Carrier, Research Assistant
- Abbey Thompson, DRIADE R.A./SILS Masters Student
- Hollie White, Doctoral Fellow
- Amy Bouck, Biology, Post doc

National Evolutionary Synthesis Center

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>

Dryad


Depositor/s
One stop 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

R & D: Accomplishments and Activities

- Functional requirements and model
  - Workshops: Stakeholders (Dec. 06), SSR (May '07)
  - Repository analysis (Dube, et al. JCDL, 2007)
    - OAIS (Open Archival Information System), DSpace
- Metadata architecture
  - **Level one application profile**

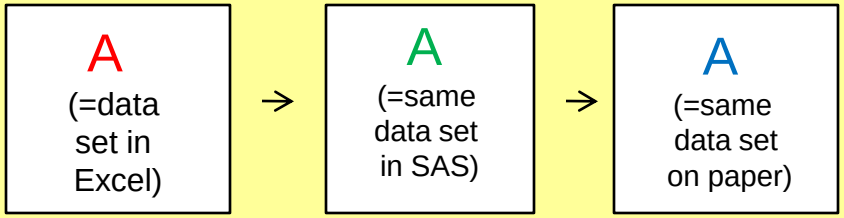
Namespace schemas:	**Modular scheme:**
1. Dublin Core 2. Data Documentation Initiative (DDI) 3. Ecological Metadata Language (EML) 4. PREMIS 5. Darwin Core	1. Journal citation 2. Data objects (Carrier, et al., 2007)

R & D: Accomplishments and Activities

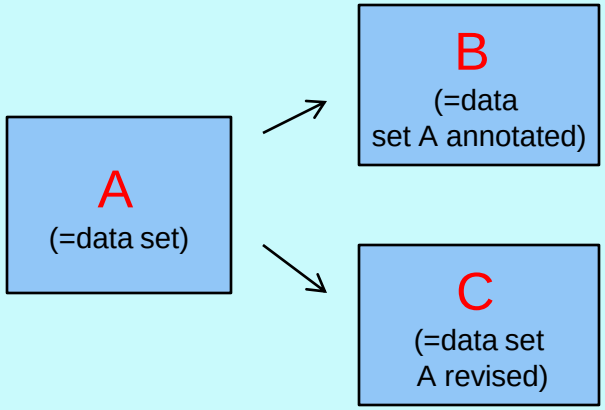
- Vocabulary analysis
  - *NBII Thesaurus*, *LCSH*, the Getty's *TGN*
    - 600 keywords, Dryad partner journals
    - Facets: taxon, geographic name, time period, topic
      - W3C SKOS (Simple Knowledge Organisation Systems)
- Instantiation study
  - Bibliographic relationships for life-cycle management (Coleman, 2002; Smiraglia, 1999, 2000, 2001, 2002, etc.; Tillett; FRBR, DCAM)

Data object relationships

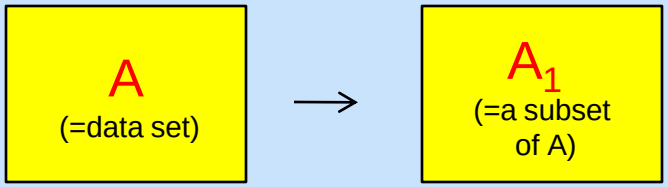
Equivalence



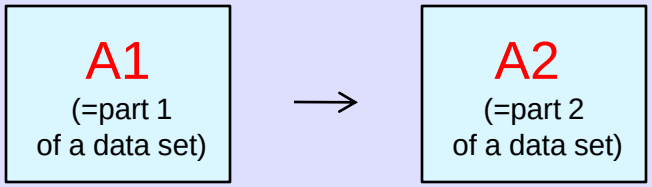
Derivative



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"

R & D: Accomplishments and Activities

- Use-case study
  - Intensive interviews with evolutionary biologists about data sharing

    ~ show KNB, ask about metadata creation, interface issues to help w/input
- Survey
  - International survey, launched via evoldir, ~ 300 respondents

    ~ included questions on labeling practices, understanding of metadata

User perceptions and behaviors re: data sharing








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 (most useful may not be the most interesting, vice/versa)
  - priorities
- Language barriers
- Funding models: Gap research and implementation
- Understanding: Trust, Task assignment
- Not having everyone in the same building

Concluding remarks… CIC

- What is eScience and why does it matter to libraries and librarians?
  - … Matters to LIS researcher and educators too, to help advance practice and train information professionals
- What are the needs of scientists who are using large data sets?
  - … Small science has needs too, similar and perhaps distinct
- What are new ways that librarians can collaborate with and support science researchers?
  - Dryad offers an exciting model
- What are the skills needed by librarians to work successfully in this arena?
  - Bias: Research and evaluate implementations

A final quote…

A revolution is taking place in the scientific method….“Hypothesize, design, and run experiment” is being replaced by “hypothesize, look up answer in database.”

(*Towards 2020 Science*, MS Research, 2006;Lesk, M. 2004)

Thank you!

Dryad repository: http://datadryad.org/

Wiki: https://www.nescent.org/wg_digitaldata/

Jane Greenberg, Director, SILS Metadata Research Center

janeg@email.unc.edu


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


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


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

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					

<DRIADE application profile>

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 Information
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

DRIADE metadata application profile….*organic…*

- Level 1 – initial repository implementation
  - Application profile: Preservation, access/resource discovery, (limited use of CVs)
- Level 2 – full repository implementation
  - Level 1++ expanded usage, interoperability, preservation; administration; greater use of CV and authority control; data sharing and reuse
- Level 3 – "next generation" implementation
  - Considering Web 2.0 functionalities, Semantic Web

Research design

- Objective: Build an open access repository that accurately reflects the “disciplinary knowledge structures” (relationship among data objects”)
- Research questions
  1. Do scientists (developing scientists) view data objects as works?
  2. Do they understand different “instantiations”?
  3. Do they think instantiation tracking is important?
- Method
  - Instantiation identification test and survey
- Participants: Scientists, research and publication

Motivation

- DRIADE goals:
  - Data **preservation, sharing, use/re-use, validation, repeatability**
    - Is it important to know history of a data object?
    - How can we support accurate and effective tracking of the life-cycle of data object?
- Data objects = first class objects
  - Data structures as works (Coleman, 2002)
    - Units of analysis, intellectual products of activity
  - Work = “propositions expressed (ideational content)”, and “expressions of the propositions” (Smiraglia, 2001)
  - Data objects are content carriers (Greenberg, 2007)

Motivation

- Research and…to explain a *"work"*; *bibliographic families;* and *instantiation*

  - Coleman (2002)
  - Cutter (1904)
  - Leaser (1999)
  - Ranganathan: embodied/expressed
  - Smiraglia (1999,2001, 2002,etc.)
  - Tillet (1991, 1992)
  - Vellucci (1995)
  - Wilson (1983)
  - Yee (1995)

- Metadata and Bibliographic Control Models
  - FRBR (Functional Requirements for Bibliographic Records)
  - DCAM (Dublin Core Abstract Model)
  - RDF (Resource Description Framework)

Results (participant comments)

- Validity: "Whenever anything changes, there's the ability to make mistakes"
- Sheer quantity of data: "We have 30 years worth of data, tracking changes is important"
- The impact of changes in scientific nomenclature:
  - "As time goes on, datasets must be maintained to reflect current understanding of taxonomy and nomenclature, to allow connection of old data and attributes of that data to be associated correctly with new data and attributes of that data. This is a giant problem."