Metadata for Scientific Data: An Analysis Targeting Challenges and Opportunity in Our Global Information Ecology

CODATA Conference

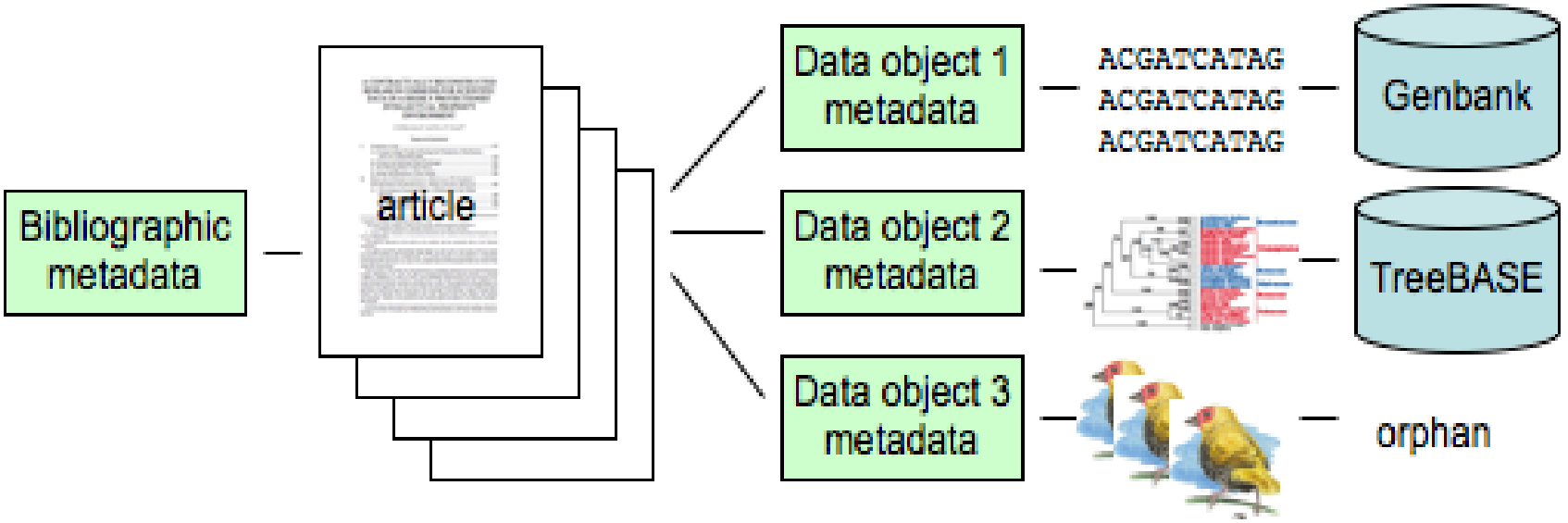
October 2010, Cape Town, South Africa

Jane Greenberg, Professor
Craig Willis, Graduate Research Assist.
Hollie White, Doctoral Research Fellow
Metadata Research Center
University of North Carolina at Chapel Hill

Overview

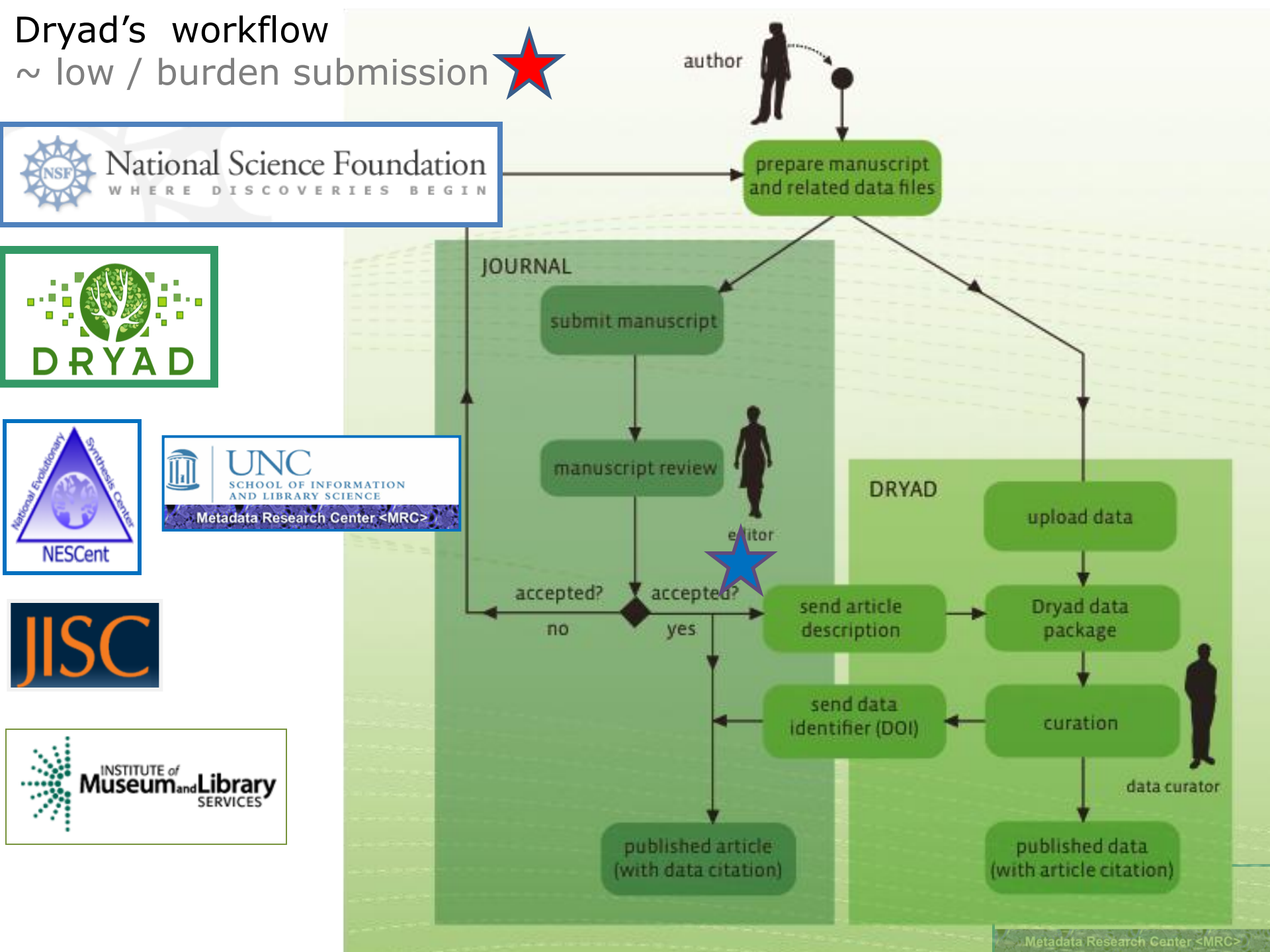
- Dryad: Metadata application profile
  - Linked data/Semantic Web
- Research on metadata schemes
  - Observations and motivation
  - Objectives and methodology
- Conclusions
- Q & A

Dryad: Digital data underlying published research



Dryad's workflow

~ low / burden submission



National Science Foundation
WHERE DISCOVERIES BEGIN

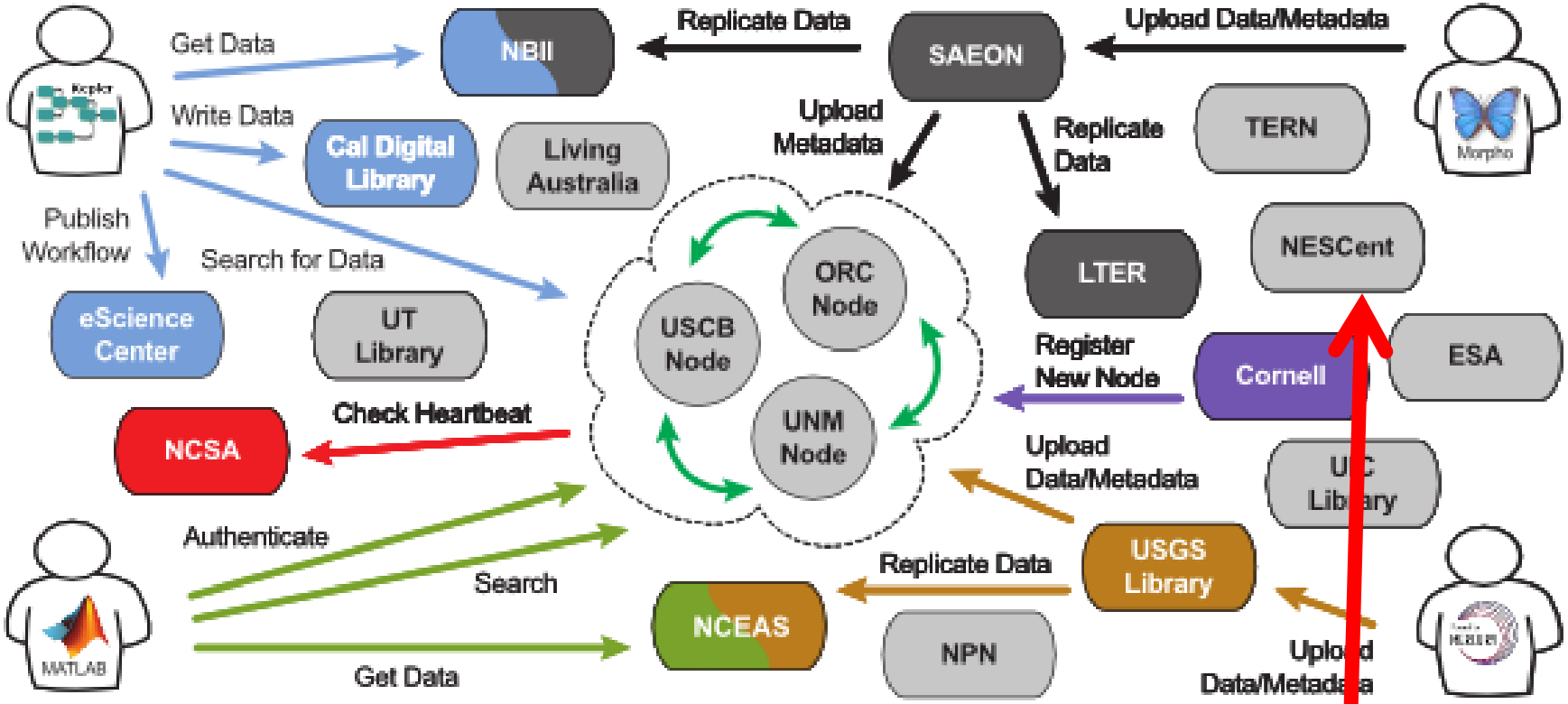
DRYAD

NESCent

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

JISC

INSTITUTE of Museum and Library SERVICES


Get Data
Write Data
Publish Workflow
Search for Data
Authenticate
Search
Get Data
Check Heartbeat
Replicate Data
Upload Metadata
Replicate Data
Upload Data/Metadata
Register New Node
Upload Data/Metadata
Replicate Data
Upload Data/Metadata
NBII
SAEON
TERN
Cal Digital Library
Living Australia
eScience Center
UT Library
NCSA
ORC Node
USCB Node
UNM Node
LTER
NESCent
Cornell
ESA
UC Library
USGS Library
NCEAS
NPN
MATLAB

Dryad metadata application profile

Dublin Core based

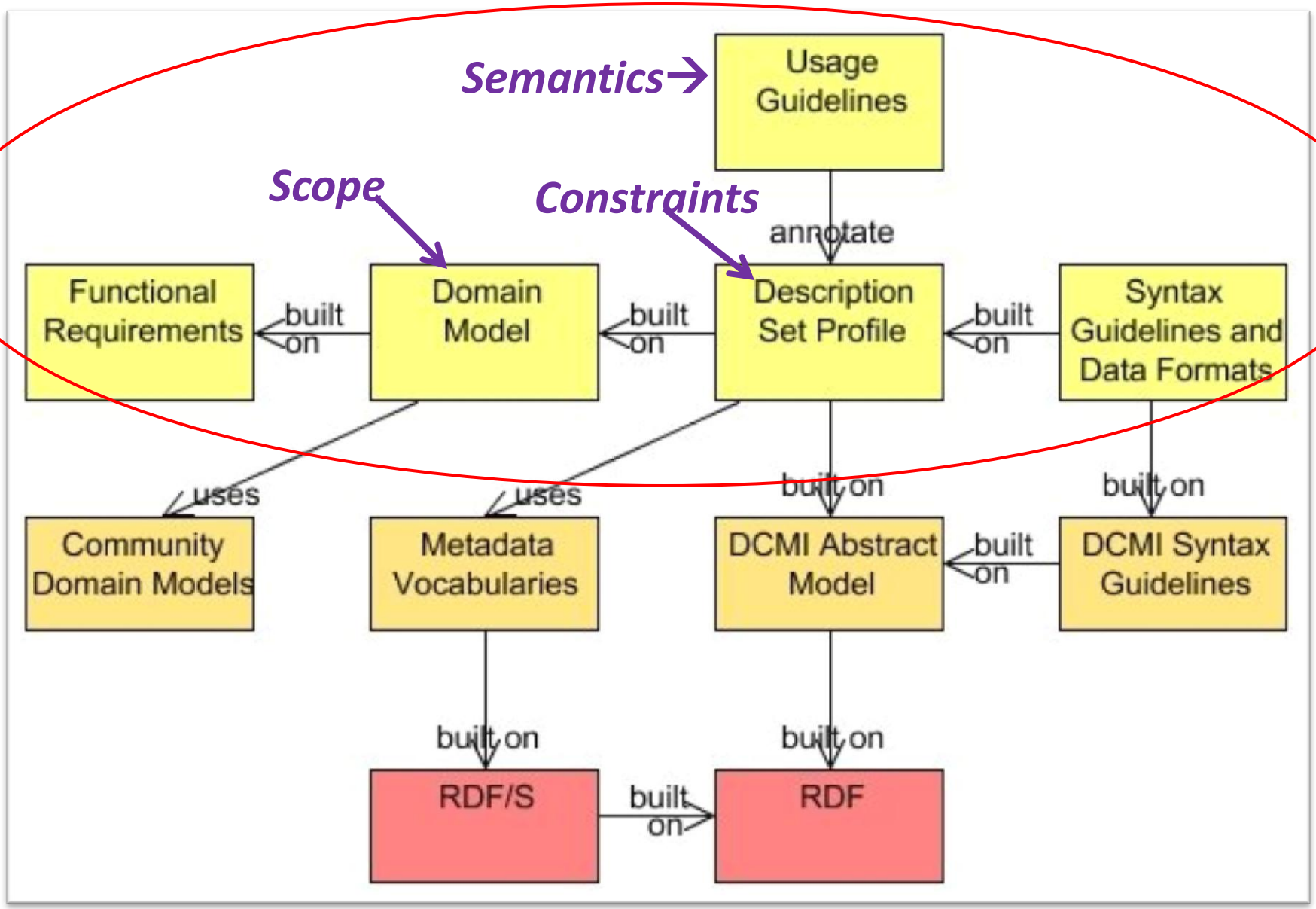
- Circumvents limitations of using a single scheme
- Interoperable with other schemes
- Why reinvent the wheel?

Modular scheme:

1. Data package
2. Journal citation
3. Data files

bibo (The Bibliographic Ontology)
dcterms (Dublin Core terms)
dryad (Dryad)
DwC (Darwin Core)

Simple: automation of metadata gen; discovery of heterogeneous datasets
Interoperable: harvesting, cross-system searching
Semantic Web Compatible: sustainable and adaptable metadata architecture, supporting machine processing



Baker, T. (2007) slides , annotated (Greenberg, 2010)

Submit Data Now!

My Account

Login
Register

Browse

Authors
Publication Date
Journal Title

Information

Depositing Data
Using Data
Dryad Partners
Archiving Policy
About Dryad
Dryad Blog

Data from: Patterns of morphological and plastid DNA variation in the Corallorhiza striata species complex (Orchidaceae)

When using this data, please cite the original article:

Barrett CF, Freudenstein JV (2009) Patterns of morphological and plastid DNA variation in the Corallorhiza striata species complex (Orchidaceae). Systematic Botany 34(3): 496-504. doi:10.1600/036364409789271245

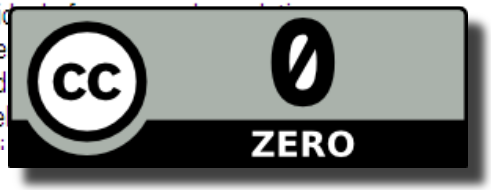
Additionally, please cite the Dryad data package:

Barrett CF, Freudenstein JV (2009) Data from: Patterns of morphological and plastid DNA variation in the Corallorhiza striata species complex (Orchidaceae). Dryad Digital Repository. doi:10.5061/dryad.1013

Dryad Package Identifier doi:10.5061/dryad.1013

Dryad Data Files http://hdl.handle.net/10255/dryad.1014
http://hdl.handle.net/10255/dryad.1015

Abstract Corallorhiza striata is a wide-ranging, morphologically variable, mycoheterotrophic species complex distributed across North America. Objectives of this study were to assess relationships and test validity of previously delimited varieties of C. striata, including the recently described C. bentleyi. Two plastid DNA regions were sequenced for indivi
North America, identifying four major clades. The large-flowe
U.S.A., southern Canada) was sister to the smaller-flowered
Mexico), and these were sister to a Californian clade with re



Observations and research motivation

- Metadata overload, (deluge?)
  - Toothbrush *scenario*
- Approached via discipline/domain, rather than function
  - EML, DDI
    - Barriers for interoperability
- Initiatives stray from core components
  - Granular metadata = rich detail
    - Can impede disciplinary and cross-domain interoperability, data discovery, access, and reuse.

Research objectives + method

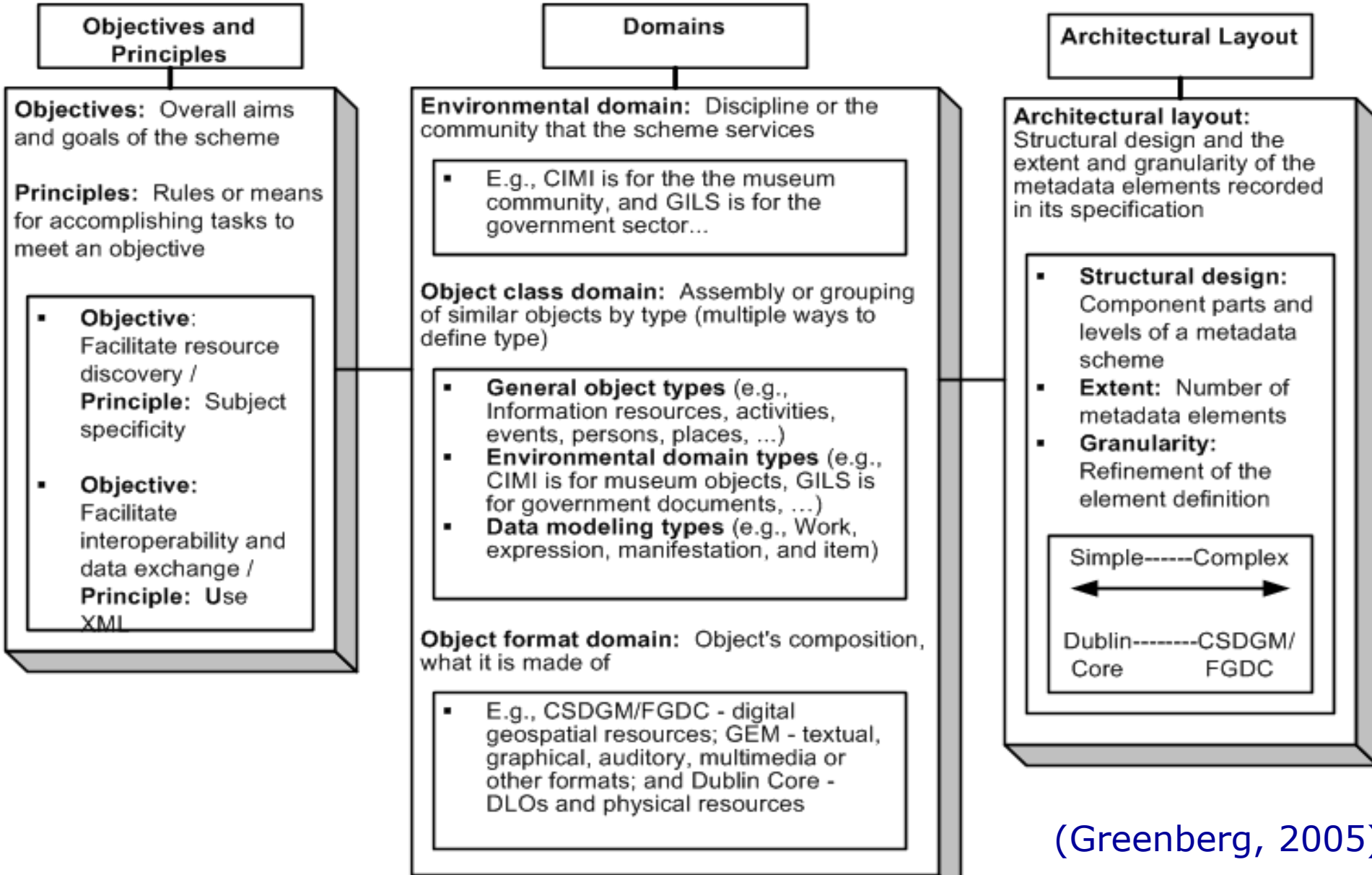
1. What is the scope of scientific metadata schemes?
2. Can we discern some similarities and differences?

Method

- Content analysis of 9 schemes, meta-analysis of literature
- Examination of 10 ontologies

The MODAL Framework

for Metadata Objectives and Principles, Domains, and Architectural Layout



(Greenberg, 2005)

Scheme selection criteria

- 9 schemes having an established relationship with publishers in one or more domain.

Additional data gathering

Review scheme documentation (codebooks, data structures), policy statements, literature, and examine use in selected databases

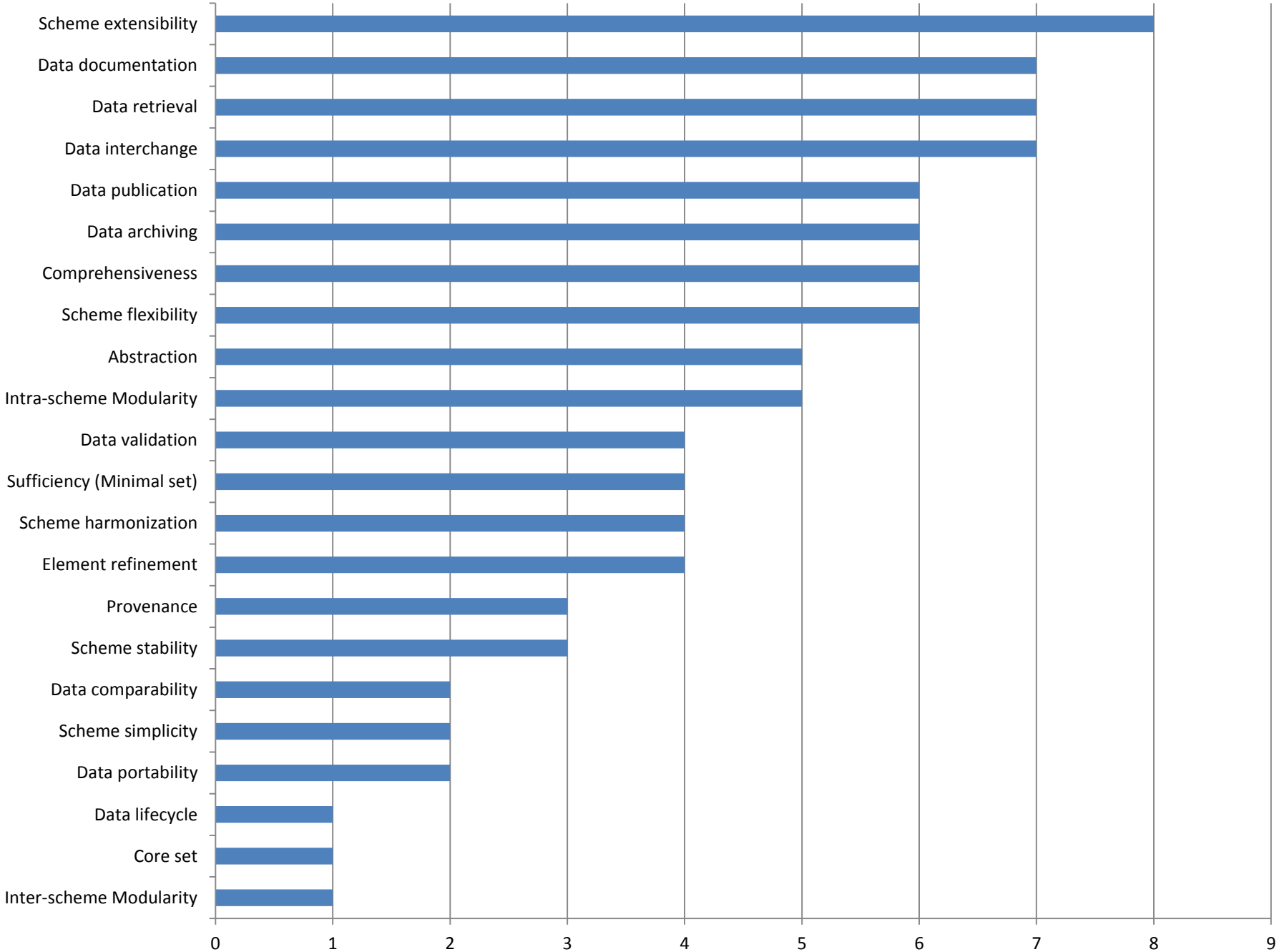
Scheme	Vers.	Initial Rel.	Maint. Body	Repository	* Pub.
1. DDI	3.1	2000	DDI Alliance	ICPSR (and others)	128[2]
2. CIF	2.4.1	1991	IUCr	Cambridge Structural Database (CSD)	130
3. DwC	App.P	2001	TDWG	GBIF	1
4. EML	2.1.0	1997	KNB	Ecological Archives	4
5. mmCIF	2.0.09	2005	wwPDB	Protein Data Bank (PDB)	NA
6. MINiML	1.16	2007?	NCBI	Gene Expression Omnibus (GEO)	53
7. MAGE	1.0	2002	FGED	ArrayExpress	53
8. NEXML	1.0	2009	NESCent	TreeBase	36
9. ThermoML	3	2002	IUPAC	ThermoML Archives	5

Functional aspects/properties (22, still refining)

1. **Core set**
2. **Data lifecycle**
3. **Data portability**
4. Scheme simplicity
5. Data comparability
6. Scheme stability
7. Provenance
8. Element refinement
9. Scheme harmonization
10. Data validation
11. Intra-scheme Modularity
12. Abstraction
13. Scheme flexibility
14. Comprehensiveness
15. Data archiving
16. Data publication
17. Data retrieval
18. Data documentation
19. Data interchange
20. Scheme extensibility

Functional aspects/properties

Criterion	Description
Core set	The scheme is intended to provide a common set of elements used to describe the most common situations.
Data lifecycle	The scheme is intended to support documentation of the data lifecycle.
Data portability	Data created using the scheme is intended to be "portable" -- software application and operating system independent. (This is generally an objective of schemes developed earlier.)


Scheme extensibility
Data documentation
Data retrieval
Data interchange
Data publication
Data archiving
Comprehensiveness
Scheme flexibility
Abstraction
Intra-scheme Modularity
Data validation
Sufficiency (Minimal set)
Scheme harmonization
Element refinement
Provenance
Scheme stability
Data comparability
Scheme simplicity
Data portability
Data lifecycle
Core set
Inter-scheme Modularity
0
1
2
3
4
5
6
7
8
9

Scheme: MODAL domains

Scheme	Environmental Domain	Object Class Domain	Object Format Domain
CIF	Crystallography	**Experimental** studies	Digital data (Crystallographic structures)
Darwin Core	Biology	**Observations** Specimen collections	Digital data (Observations) Physical specimen
DDI	Social sciences	**Experimental** studies **Observational** studies	Digital data (Social science statistical data)
EML	Ecology	**Experimental** studies **Observational** studies	Digital data (Observations)
MAGE	Molecular biology	**Experimental** studies	Digital data (Micro-array based gene expressions)
MINiML	Molecular biology	**Experimental** studies	Digital data (Micro-array based gene expressions)
mmCIF	Structural biology	**Experimental** studies	Digital data (Macromolecular structures)
NEXML	Phylogeny	**Experimental** studies	Digital data (Phylogenic trees)
ThermoML	Thermodynamics	**Experimental** studies	Digital data (Thermodynamic

Scheme	Encodings	Structural Design	Extent	# of files	# of levels
CIF	STAR DDL XSD	Data blocks Categories Data items	18 62 486	1 DDL	5
DDI	XSD	Elements Complex types Simple types	797 296 599	22 XSD	6+
EML	XSD	Resource modules Supplemental modules Elements Complex types Simple types	4 6 579 174 54	25 XSD	5+

Challenges w/ the analysis

Conclusions: Metadata analyses and application profile work, *thoughts…to date…*

Positive aspects

- Peel away at silos built via metadata development
- Intellectually engaging
- Think we are making a contribution, have to start somewhere…
- App. Profiles step preparing for machine capabilities; eScience /data synthesis

Challenges

- Time consuming
- Documentation inconsistent and lacking
- App. Profiles: Infrastructure not all there… (a lot is not in RDF)
  - Registered Dryad "purl"
  - Registries!!
- Proof of concept difficult

Metadata

- Data about data; information about information
- Structured information that supports functions (Greenberg, 2003, 2009)

Metadata functions

- Discovery/Retrieval
- Life-cycle Management
- Preservation
- Usage
  - Rights
  - Technical use
- Ratings/audience appropriateness
- Authentication
- Provenance tracking

National Science Foundation
WHERE DISCOVERIES BEGIN

INSTITUTE of Museum and Library SERVICES

DRYAD

For more information

- Metadata Research Center: http://ils.unc.edu/~mrc
  - Publications page
- Dublin Core Metadata Initiative/Science and Metadata: http://purl.org/dc/science.
- 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

Follow Dryad on Facebook and Twitter