Consensus Building and Prioritizing <Metadata> Development for Project DRIADE: A Case Study

~~~~~~

DigCCur 2007, Chapel Hill North Carolina
April 18, 2007

Jane Greenberg, Associative Professor and
Director, SILS Metadata Research Center,
School of Information and Library Science,
University of North Carolina at Chapel Hill

NESCent

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>
~~~~~~

Overview

- Introduce DRIADE
  - Motivation
- Consensus building
- Functional requirements
- Metadata framework
- Conclusions and next steps
- Implications for digital curation education

DRIADE: Digital Repository of Information and Data for Evolution

http://www.caffedriade.com/

- Internet impact / “small science”
  - Knowledge Network for Biocomplexity (KNB)
  - Marine Metadata Initiative (MMI)
- Evolutionary biology
  - Ecology, genomics, paleontology, population genetics, physiology, systematics, …genomics
  - Data deposition (Genbank, TreeBase)
  - Supplementary data
    Molecular Biology and Evolution

NESCent

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>

DRIADE's goals

- One-stop shopping for scientific data objects supporting published research
- Support data acquisition, preservation, resource discovery, data sharing, and data reuse of heterogeneous digital datasets
- Balance a need for low barriers, with higher-level … data synthesis

DRIADE Team

NESCent

- Todd Vision, Director of Informatics and Assistant Professor, Biology, UNC-CH
- Hilmar Lapp, Assistant Director of Informatics

UNC-CH/SILS/MRC

- Jane Greenberg, Associate Professor
- Jed Dube, MRC Doctoral Fellow
- Sarah Carrier, MRC Research Assistant
- Amy Bouck, UNC/Duke Biology Postdoc

Consensus building: Stakeholders' workshop

1. Unanimous support for DRIADE
   - Advance science, cultural change, policing
2. Challenges
   - Scope, representation, quality control, security, cultural change, sustainability
3. Priorities and next steps
   - Preservation – access – synthesis
     - Maslow's hierarchy of life needs!
   - Cultural change: editorials, publicizing at conferences, requirements

NESCent

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

Metadata Research Center <MRC>

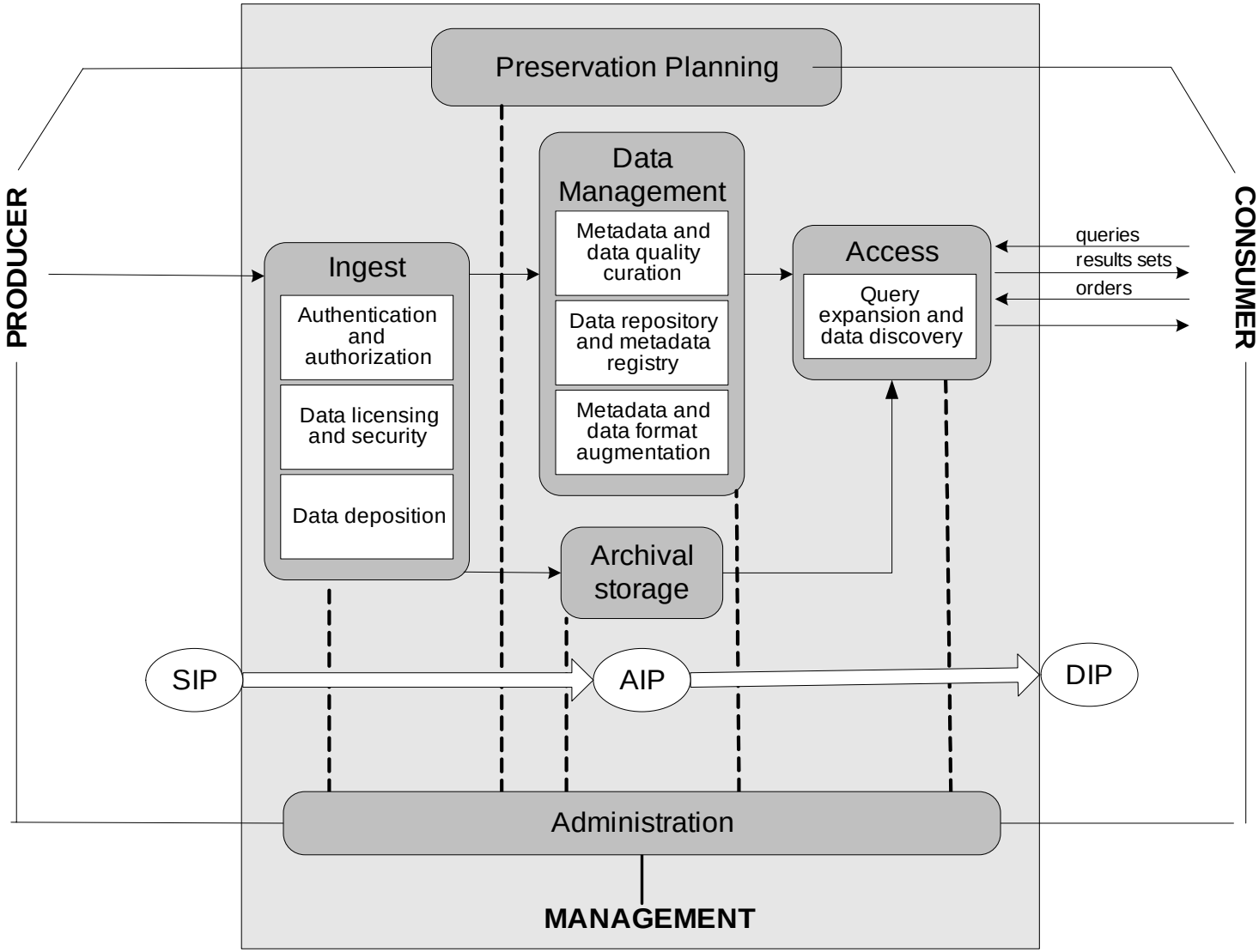
Functional requirements

	GBIF	KNB/ SEEK	NSDL	ICPSR	MMI
Heterogeneous digital datasets	▪	▪	▪	▪	▪
Long-term data stewardship	▪		▪		
Tools and incentives to researchers	▪	▪	▪	▪	▪
Minimize technical expertise and time required	▪	▪		▪	▪
Intellectual property rights	▪	▪		▪	
Published Datasets					

Functional requirements

- Support:
  - Computer-aided metadata generation / augmentation
  - Specialized modules linking data submission and manuscript review
  - Data and metadata quality control by integrating human and automatic techniques
  - Data security
  - Basic metadata repository functions, such as resource discovery, sharing, and interoperability

DRIADE's functional model based on OAIS



DRIADE metadata framework

- Level 1 – initial repository implementation
  - Preservation, access, and basic usage of data, (limited use of CVs)
- Level 2 – full repository implementation
  - Level 1 plus expanded usage, interoperability, preservation, administration, etc., greater use of CV and authority control
- Level 3 – “next generation” implementation
  - Considering Web 2.0 functionalities

Application profiles

- "…consist of data elements drawn from one or more namespace schemas combined together by implementors and optimised for a particular local application." (Heery& Patel, 2000)
  - Data Elements: Title, Name, Coverage, Identifier, etc.
  - Namespace schemas:
    - Dublin Core
    - Data Documentation Initiative (DDI)
    - Ecological Metadata Language (EML)
    - PREMIS
    - Darwin Core

Why create an Application Profile?

- Single existing schemes are often not sufficient
  - Dublin Core scheme doesn't meet all of DRIADE needs
  - Do not need all elements in a single scheme (e.g. in DDI or EML)
- Don't want to re-invent the wheel
- Interoperability

Why DRIADE needs an application profile?

- Evolutionary biology data requires a range of metadata to effectively support:
  - Unstructured datasets, non-standard formats
  - Varied data relationships, methods, software
  - Varied data object relationships (i.e. part of larger studies, linkages to publications, etc.)
  - Immediate and future dataset preservation

Level 1+ Application Profile

Module 1: Bibliographic Citation

- dc:title / Title*
- dc:creator / Author*
- dc:subject / Subject*
- dc:publisher / Publisher*
- dcterms:issued / Year*
- dcterms:bibliographicCitation / Citation information*
- dc:identifier / Digital Object Identifier*

Level 1+ Application Profile

Module 2: Data Object

- dc:creator / Name
- dc:title / Data set title
- dc:identifier / Data set identifier♦
- fixity (PREMIS) / (*hidden*)♦
- dc:relation / Digital Object Identifier of published article
- DDI: <depositr> /Depositor or submitter name*
- DDI: <contact> / Contact information for <depositr>*
- dc:rights / Rights statement♦
- dc:description / Description of the data set *
- dc:subject / Keywords describing the data set *
- dc:date / Date modified ♦
- dc:date / (*hidden*) ♦
- dc:format / File format ♦
- dc:format / File size ♦
- dc:software / Software
- dc:coverage / Locality
- dc:coverage /Date range

Level 3, *brainstorm…*

- Personalization, query results, workflow "macros", user interface
- Virtual societies utilizing "social tagging"
- Integration and extension of existing ontologies
  - Implementation of emerging standards
    - Minimal Information About a Phylogenetic Analysis (MIAPA)
- Harvesting metadata (pull) / Exposing metadata (push)
- Visualizations: topic clustering data relationship maps

Conclusions and next steps

- Conclusions
  - Team work required
    - stakeholders (scientists and journal representatives), metadata experts, and sustainability partner
  - Late to the game, benefit from what's been accomplished (e.g., application profile, models)
  - Need to understand DRIADE's unique goals
- Next steps:
  - Survey and use-case/life-cycle studies
  - Metadata application profile experiment

Implications for digital curation education

- Students participation, service learning
- Curriculum needs to address the whole picture –
  - Digital resource life-cycle
  - Metadata life cycle
  - IA components
  - Human factors
- Language barriers and communication skills
  - Metadata facets… woo woo???
- Conferences like DigCCur

References

- Application profiles: mixing and matching metadata schemas http://www.ariadne.ac.uk/issue25/app-profiles/
- Application Profiles, or how to Mix and Match Metadata Schemas http://www.cultivate-int.org/issue3/schemas/
- Dublin Core Element Set: http://dublincore.org/documents/dces/
- Data Documentation Initiative (DDI) http://www.icpsr.umich.edu/DDI/
- Ecological Metadata Language (EML) http://knb.ecoinformatics.org/software/eml/
- PREMIS http://www.oclc.org/research/projects/pmwg/
- Darwin Core Wiki: http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome