Evolution of an Application Profile: Advancing Metadata Best Practices through the Dryad Data Repository

Edward M. Krause, Erin Clary, Adrian Ogletree, & Jane Greenberg

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

@all_metadata

www.cci.drexel.edu/mrc

Speaker Introduction

Edward Krause

- Master's student in Health Informatics

- Data Curator, Dryad Repository, Metadata Research Center

- College of Computing & Informatics, Drexel University, Philadelphia, PA, USA

2

Overview

- Dryad
  - DSpace Architecture
  - Metadata Architecture
- Research Objectives

- Methods
  - Crosswalk Analysis
  - Domain Models
- DCAP V3.2
- Next Steps

What is Dryad?

- A curated digital repository for data underlying peer-reviewed, scholarly publications

- Considered a general purpose repository, containing data from a wide variety of disciplines

- Dryad's goal is to make data discoverable, accessible, reusable, and citable

Dryad, cont.

- Community and stakeholder oriented

- Data-centric model

- Metadata application profile
  - Focused on capture, preservation and retrieval functions
  - Domain-specific elements
  - Heterogenous data types
  - Link publications with underlying datasets

Research Objectives

- Understand the growing representation of data from wide range of disciplinary fields

- Assess changes in Dryad application profile over time (v1.0 - v3.1)

- Document current metadata practices as version 3.2 of the application profile

- Propose next steps to update the application profile

Early Application Profile Work
(2007-2012)

- DSpace Framework (version 1.8)

- Metadata Best Practice:
  - Dublin Core Abstract Model
  - Singapore Framework for Dublin Core Application Profiles

- Two pronged approach:
  - Short-term Goal: Make content available in DSpace through an XML schema
  - Long-term Goal: Align Dryad with Semantic Web standards

- Definition of the repository's functional requirements & development of a domain model

Further Application Profile Development (2013-2015)

- V3.1 most current published version of Dryad application profile

- Namespaces include:
  - Dublin Core, Darwin Core, Bibliographic Ontology Specification (bibo), Dryad

- Components include:
  - bibliographic metadata
  - scope and coverage of data files
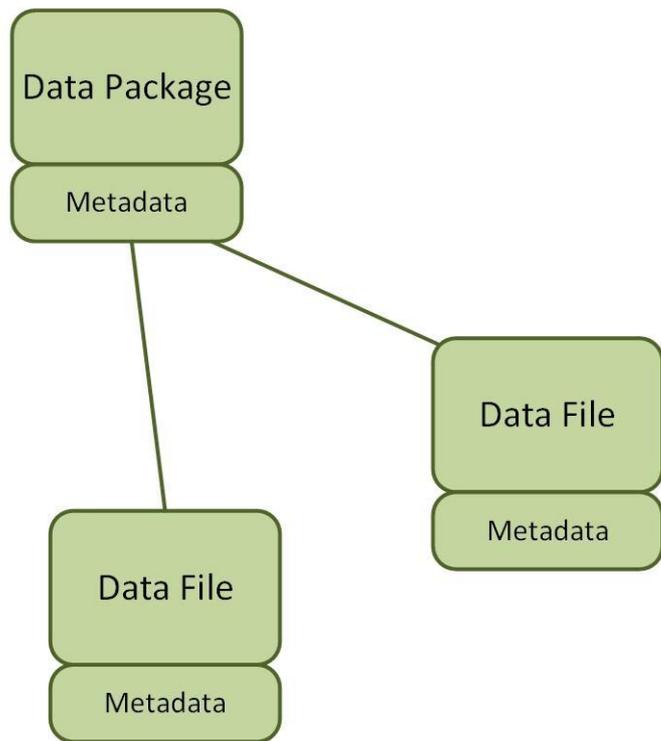 - relationships among data files, data package, and publication

Dryad DCAP v3.1: Metadata Elements

Data Package	Data File
dcterms:type	dcterms:type
dcterms:creator	dcterms:creator
dcterms:dateSubmitted	dcterms:dateSubmitted
dcterms:available	dcterms:available
dcterms:title	dcterms:title
dcterms:identifier	dcterms:identifier
dcterms:description	dcterms:description
dcterms:subject	dcterms:subject
dwc:scientificName	dwc:scientificName
dcterms:spatial	dcterms:spatial
dcterms:temporal	dcterms:temporal
external	embargoedUntil
dcterms:references	dcterms:rights
bibo:pmid	dcterms:format
bibo:Journal	dcterms:provenance
dcterms:hasPart	dcterms:isPartOf

DSpace Architecture

- Open source repository framework

- Latest version is v5.0, but Dryad implemented on v1.8

- No plans to upgrade due to risk of incompatibilities resulting from Dryad customizations

- Internal DSpace metadata element structure not supported by Dublin Core

- Automated metadata harvesting converts internal elements to Dublin Core compliant properties

Dryad Metadata Architecture

- One data package associated with a single peer-reviewed, scholarly publication

- One or more data files associated with each data package

- Linked by relationships expressed in the metadata

10

Methods

- Crosswalk Analysis

- Analysis of domain model changes

- Comparison of each application profile to previous iteration

- Documentation of changes in element usage

- Construction and documentation of application profile version 3.2

Crosswalk Analysis

Comparison of each application profile version to its next iteration (v1.0-v2.0, v2.0-v3.0, v3.0-v3.1)

Four possible cases for each metadata element:

- Element and concept did not change and is present through all iterations
- Concept did not change, but the element representing a concept changed
- Element and concept is added
- Element and concept is phased out

Crosswalk Analysis, cont.

Elements and concepts which did not change:

- descriptions
- spatial and temporal characteristics
- digital identifiers
- entity types
- relationships
- subjects
- taxonomic classification

Crosswalk Analysis, cont.

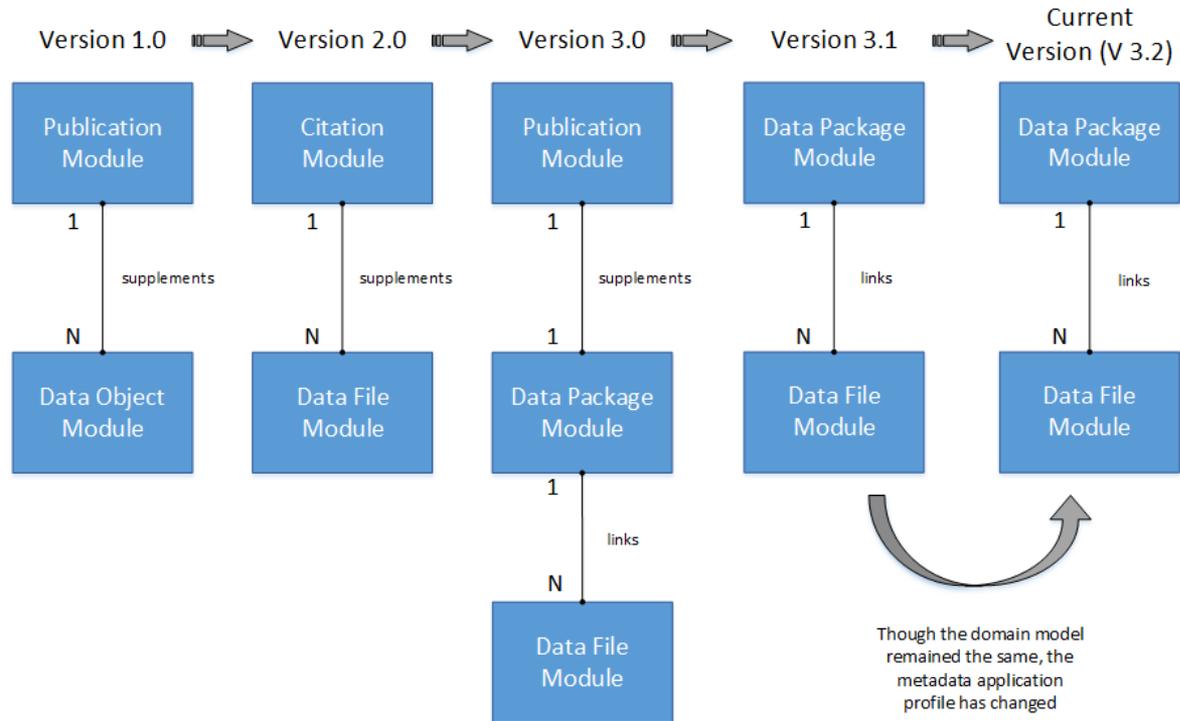Concepts static, but elements representing the concepts were modified:

- embargo date
  - dcterms:available ⟹ dcterms:embargoedUntil

New concepts and elements added:

- dcterms:date.blackoutUntil
- dcterms:identifier:manuscriptNumber
- dcterms:type.embargo

Domain Model Evolution

- Expansion of metadata terms and granularity of usage from v1.0 - v3.0

- v3.0 never fully implemented due to architectural constraints

- V3.1 consolidated domain entities and metadata

Version 1.0 ⟹ Version 2.0 ⟹ Version 3.0 ⟹ Version 3.1 ⟹ Current Version (V 3.2)

Version 1.0	Version 2.0	Version 3.0	Version 3.1	Current Version (V 3.2)
Publication Module	Citation Module	Publication Module	Data Package Module	Data Package Module
1	1	1	1	1
supplements	supplements	supplements	links	links
N	N	1	N	N
Data Object Module	Data File Module	Data Package Module	Data File Module	Data File Module

Data Package Module
1
links
N
Data File Module

Though the domain model remained the same, the metadata application profile has changed

15

Domain Model Evolution, cont.

- Expanded domain model results in increased number of metadata elements

- Concept of data package introduced later in Dryad's development

  - Contains metadata to describe both the associated scholarly publication and data file(s) and to link these different objects together

  - Dryad submissions centralized on data package, the "base" unit in Dryad to which the publication and data files are linked

Dryad Application Profile, Version 3.2

- Identifies how concepts are represented internally and how they can be represented externally when metadata are harvested

- Movement towards a more Dublin Core-centric application profile

- 24 elements in the data package module and 21 elements in the data file module
  - Some elements repeated in both modules, while others are present in only a single module

- Added elements reflect new functionalities and workflows
  - I.e., capture of manuscript numbers for associated publications supports consideration of data in the peer review process

17

Application Profile V3.2, cont.

Internal Element Representation (DSpace):		dcterms:contributor.author	
External Element Representation (Metadata Harvesting APIs):		dcterms:creator	
URI:	http://purl.org/dc/terms/creator		
Definition:	Authors on publication / Authors of data submission		
Module(s): Package & File	**Obligation:** Required	**Cardinality:**	Repeatable

**Full Application Profile V3.2 stored as a Dryad dataset and is referenced in the paper

Dryad Next Steps

- Publicly declare the Dryad-specific subproperties using the Dryad PURL domain
- Content analysis on a selected set of metadata schemas and elements (e.g. DDI)
- Re-evaluate Dryad's functional requirements
  - Consider and identify new stakeholders and data types
  - Survey and/or interviews with data submitters

19

Questions?

Contact information:

- Edward Krause, emk92@drexel.edu
- Erin Clary, eclary@datadryad.org
- Adrian Ogletree, aogletree@drexel.edu
- Jane Greenberg, janeg@drexel.edu

@all_metadata

www.cci.drexel.edu/mrc

References

Carrier, Sarah. (2008). The Dryad Repository Application Profile: Process, Development, and Refinement (Master's Paper).   University   of   North   Carolina   at   Chapel   Hill. Retrieved   from https://cdr.lib.unc.edu/indexablecontent/uuid:727a2712-f8f0-40a3-8876-e8e16b6d086d.

Dryad. (2013). Metadata Profile. Dryad Wiki. Retrieved from http://wiki.datadryad.org/Metadata_Profile.

Dryad. (2015). The Organization: Overview. Retrieved March 30, 2015, from http://datadryad.org/pages/organization.

Dryad Data Repository. (2015). Retrieved from http://datadryad.org/DSpace. (2015). DSpace. Retrieved from http://http://www.dspace.org/Dube, Jed. (2007). A Metadata Application Profile for the DRIADE Project.

Greenberg, Jane, Hollie White, Sarah Carrier, and Ryan Scherle. (2009). A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata, 9(3), 194-212. doi:10.1080/19386380903405090

Krause, Edward M., Erin Clary, Adrian Ogletree, Jane Greenberg. (2015). Data from: Evolution of an application profile: advancing metadata best practices through the Dryad data repository. Dryad Digital Repository. doi:10.5061/dryad.f0n35

Mannheimer, Sara, Ayoung Yoon, Jane Greenberg, Elena Feinstein, and Ryan Scherle. (2014). A Balancing Act: The Ideal and the Realistic in Developing Dryad's Preservation Policy. First Monday, 19(8). doi:10.5210/fm.v19i8.5415

Nevile, Liddy, and Eva Méndez. (2015). Do We Need Application Profiles? Reflections and suggestions from work in DCMI and ISO/IEC. In DC-2015: Proceedings of the International Conference on Dublin Core and Metadata Applications. São Paulo, Brazil, September 1-5, 2015.

Nilsson, Mikael, Thomas Baker, and Pete Johnston. (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved from http://dublincore.org/documents/singapore-framework/.

White, Hollie. (2008). Exploring evolutionary biologists' use and perceptions of semantic metadata for data curation. In International Conference on Dublin Core and Metadata Applications 2008, September 22-26, 2008, Berlin, Germany, pp. 202. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/936/932.