Evolution of a Metadata Application Profile for a Digital Data Repository

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

Metadata Research Center, College of Computing and Informatics, Drexel University

ABSTRACT

Dryad is a curated, digital archive for data associated with scholarly publications. In an effort to facilitate the discoverability, reusability, and interoperability of archived content, Dryad has implemented a standardized set of metadata elements in the form of an application profile. Dryad metadata captures information about data packages, which are comprised of individual data files, the associated scholarly publication, and the relationships among these entities. This research examines the evolution of Dryad's application profile from its inception in 2007 as version 1.0 through the last update in 2013 as version 3.1, and documents current practice as version 3.2. We model the relationships between data packages, data files, and publications for each version of the application profile and perform a crosswalk analysis to map equivalent metadata elements across each version. Results covering versions 1.0 to 3.0 show an increase in the number of metadata elements used to describe data objects in Dryad. Results also confirm that Version 3.0, which envisioned separate metadata element sets for data package, data files, and publication metadata, was never fully realized due to constraints in Dryad system architecture. Version 3.1 subsequently reduced the number of metadata elements captured by recombining the publication and data package element sets. Version 3.2 represents the current metadata practices in Dryad and demonstrates changes in the content and functionality of the repository. This work aligns the application profile with current Dryad practices and informs a larger effort to meet the needs of Dryad's diverse community of stakeholders and its expanding scope.

BACKGROUND

What is Dryad?

Dryad is a curated digital repository for scientific data underlying peer-reviewed scholarly literature, which accepts data from a wide variety of disciplines, including medical and social sciences. Dryad's chief mission is to make data discoverable and reusable for scientific endeavors.

Dryad Structure

The primary entity represented in Dryad's structure is the data package. Each data package is linked to its associated publication, and Dryad stores metadata related to the data package and its files, in addition to metadata derived from the publication. Dryad's data package model associates a single data package with one or more data files. The initial goals of developing an application profile for Dryad were twofold; an immediate short-term concern was to make content available in the DSpace framework through an XML schema, and in the long-term, to align with the Semantic Web [1]. Dryad requires metadata representing publication and bibliographic information, subject domains, relationships among entities, and identifiers [2].

RESEARCH OBJECTIVES

1. Evaluate changes to the Dryad Application Profile over time
2. Align Dryad's metadata element set with current metadata practices
3. Document current practices as Version 3.2 of the application profile

METHODS

1. Crosswalk Analysis – map equivalent metadata elements across each application profile version to examine changes in metadata usage
2. Domain Model – demonstrate the relationships between the entities that represent data package, data files, and publication information for each version of the application profile
3. Summary of Version Changes – compare each version of the application profile to the previous iteration and document changes and constancy in metadata element usage
4. Develop Dryad Application Profile Version 3.2 – compile the current elements and attributes that correspond with each domain entity in the repository

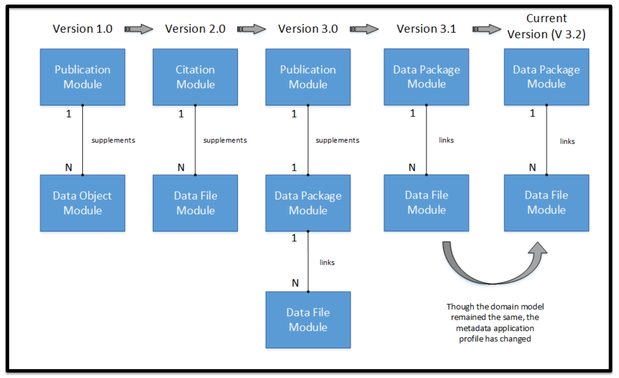
CROSSWALK ANALYSIS

Dryad Metadata Application Profile Elements Version 1.0	Dryad Metadata Application Profile Elements Version 2.0	Dryad Metadata Application Profile Elements Version 3.0	Dryad Metadata Application Profile Elements Version 3.1	Current Metadata Elements Version 3.2	Definitions
dcterms:available	dcterms:available	dryad:embargoedUntil	dryad:embargoedUntil	dcterms:date.embargoedUntil	Embargo date - a date after which the dataset will be made public
dcterms:issued	dcterms:issued	dcterms:issued		dcterms:date.issued	Date of journal article publication
		dcterms:dateSubmitted	dcterms:dateSubmitted		An automatic timestamped date when a depositor finalizes their submission
		dcterms:provenance	dcterms:provenance	dcterms:description.provenance	Information related to the origin and integrity of the file
dcterms:description	dcterms:description	dcterms:description	dcterms:description	dcterms:description	Description of entity (data package, data file, data object, publication, etc.)
dcterms:abstract	dcterms:abstract	dcterms:abstract			Abstract of associated Journal Article (data package) or description of data file
dcterms:extent	dcterms:extent	dcterms:extent		dcterms:format.extent	Size of the file storage
dcterms:format	dcterms:format	dcterms:format	dcterms:format		Format in which the data set is stored - can also represent software format
dcterms:bibliographicCitation				dcterms:identifier.citation	Journal article citation
				dcterms:identifier.manuscriptNumber	Manuscript number of associated journal article
				dcterms:identifier.uri	URL which links to the web location of the data package in Dryad
dcterms:identifier	dcterms:identifier	dcterms:identifier	dcterms:identifier	dcterms:identifier	DOI of the dataset (data package)
dcterms:isPartOf	dcterms:isPartOf	bibo:doi		dcterms:relation.isreferencedby	DOI of published journal article associated with the data package
dcterms:hasPart	dcterms:hasPartOf	dcterms:hasPart	dcterms:hasPart	dcterms:relation.haspart	Associated Dryad data file record identifier (doi:###/1 ; doi:###/2 ; etc.)
		dcterms:isPartOf	dcterms:isPartOf	dcterms:relation.ispartof	Associated Dryad Data Package Identifier (doi:###) - the "root" doi of the package

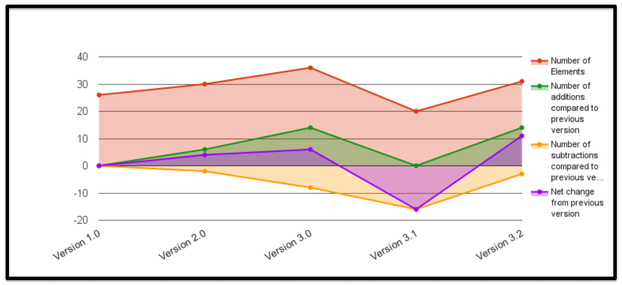
The crosswalk analysis revealed four possible cases for each metadata element in the application profile: 1) The element and the concept it represents (an element-concept pair) did not change, and were present in all iterations of the application profile. 2) The concept did not change, but the element that was used to represent that concept did change from version to version. 3) Elements and concepts are added, and 4) Elements and concepts are phased out.

CHANGING DOMAIN MODELS

The evolution from version 2.0 to version 3.0 of the application profile domain model shows an expanded set of entities, where the publication module is split from the data package module and the data package module is linked the data file module. However, version 3.0 was never fully implemented due to constraints on the Dryad system architecture, and few practical benefits to Dryad's users. Version 3.1 recombined the publication and data package modules into a single data package module, bringing the domain model more in line with earlier versions. Version 3.2 of the application profile preserves the domain model of version 3.1, but includes changes in the metadata elements it represents.



SUMMARY OF VERSION CHANGES



The summary of version changes shows the net change in the number of metadata elements between each application profile version. The increase in in elements at v3.0 corresponds with the increase in the number entities represented in the corresponding domain model. Increases in v3.2 can be attributed to new repository functionalities.

DRYAD APPLICATION PROFILE v3.2

Namespace: Name/Label	Cardinality	Module	Definition
dcterms:contributor.author	R	B	Authors on publication or data submission
dcterms:contributor.correspondingAuthor	NR	P	Name of person to contact with queries about the data
dcterms:coverage.spatial	R	B	Spatial description of the data specified by a geographic description and/or geographic coordinates
dcterms:coverage.temporal	R	B	Temporal description of the data, as geologic timespan or dates of data collection/research
dcterms:date.accessioned	NR	B	Date and time the package becomes available on DSpace
dcterms:date.available	NR	B	Date and time the package becomes available on DSpace
dcterms:date.blackoutUntil	NR	P	A date after which the dataset will automatically archive itself (move out of publication blackout)
dcterms:date.embargoedUntil	NR	F	Embargo date - a date after which the dataset will be made public
dcterms:date.issued	NR	B	Date of journal article publication
dcterms:description	NR	B	Description of entity (data package, data file, data object, publication, etc.); In the data package module, refers to abstract of associated journal article
dcterms:description.provenance	R	B	Information related to the origin and integrity of the file
dcterms:format.extent	R	F	Size of the file storage
dcterms:identifer	NR	B	DOI of the dataset (data package)
dcterms:identifier.citation	NR	P	Journal article citation
dcterms:identifier.manuscriptNumber	NR	P	Manuscript number of associated journal article
dcterms:identifier.uri	NR	B	URL which links to the web location of the data package in Dryad
dcterms:relation.haspart	R	P	Associated Dryad data file record identifier (doi:###/1 ; doi:###/2 ; etc.)
dcterms:relation.ispartof	NR	F	Associated Dryad Data Package Identifier (doi:###) - the "root" doi of the package
dcterms:relation.ispartofseries	NR	P	Associated publication/journal/article series info
dcterms:relation.isreferencedby	NR	P	DOI of published journal article associated with the data package
dcterms:rights.uri	NR	F	URL which links to the web location of the statement regarding the rights held over the resource http://creativecommons.org/publicdomain/zero/1.0/
dcterms:subject	R	B	Keywords associated with data object
dcterms:title	NR	B	Title of entity (article, dataset, package, file, etc.)
dcterms:type	NR	B	Entity type (article, package, data object, data file, etc.)
dcterms:type.embargo	NR	F	Length of Embargo (none, oneyear, custom)
dryad:downloads	NR	F	number of times the data file has been downloaded
dryad.externalIdentifer	R	P	Unique identifier for related data in Dryad partner repository (stored with prefixes e.g. GB###)
dryad:pageviews	NR	F	number of times the data file webpage has been viewed
dwc.ScientificName	R	B	Full name of the lowest level taxon to which the organism has been (may also specify other levels of biological taxonomy)
dryad.status	NR	B	Status of the metadata record
prism.publicationName	NR	P	Name of journal associated with dataset

The Dryad Application Profile v3.2 table shows the namespace and name of the element; a definition; cardinality, which can either be repeatable (R) or non-repeatable (NR); and the module in which the elements are included. Elements may be located in the data package module (P), the data file module (F) or both modules (B). The updated data package module contains 25 metadata elements and the data file module contains 22 metadata elements.

In order to re-evaluate Dryad's functional requirements, it will be necessary to identify and consider new stakeholders and more complicated curation workflows. We will also need to consider the increasingly diverse data formats and types that are used in the scientific domains represented in Dryad. New metadata elements may be needed to properly describe and preserve clinical data, social science data, and any other scientific data that Dryad could accept in the future.

ACKNOWLEDGEMENTS

We would like to acknowledge Ryan Scherle, Dryad Data Architect and Thomas Baker, DCMI.

REFERENCES

1. 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
2. Dryad. (2013). Metadata Profile. Dryad Wiki. Retrieved from http://wiki.datadryad.org/Metadata_Profile.