

The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment

Hollie C. White
University of North
Carolina at Chapel Hill,
USA
hcwhite1@email.unc.edu

Sarah Carrier
University of North
Carolina at Chapel Hill,
USA
scarrier@email.unc.edu

Abbey Thompson
University of North
Carolina at Chapel Hill,
USA
abbeyth@email.unc.edu

Jane Greenberg
University of North
Carolina at Chapel Hill,
USA
janeg@email.unc.edu

Ryan Scherle
National Evolutionary
Synthesis Center, USA
rscherle@nescent.org

Abstract

This report presents recent metadata developments for Dryad, a digital repository hosting datasets underlying publications in the field of evolutionary biology. We review our efforts to bring the Dryad application profile into conformance with the Singapore Framework and discuss practical issues underlying the application profile implementation in a DSpace environment. The report concludes by outlining the next steps planned as Dryad moves into the next phase of development.

Keywords: Dryad; application profile; Singapore Framework; metadata generation; DSpace

1. Introduction

The Dryad repository³⁹ is a partnership between the National Evolutionary Synthesis Center (NESCent)⁴⁰ and the School of Information and Library Science, Metadata Research Center (SILS/MRC)⁴¹ at the University of North Carolina at Chapel Hill. The repository hosts data supporting published research in the field of evolutionary biology. Dryad is currently working collaboratively with ten leading journals that publish evolutionary biology research, including *Evolution*, *The American Naturalist*, and *Ecology*. These journals have agreed to integrate their submission systems with Dryad in the near future, eventually creating a seamless publication process from author to journal to Dryad data deposition.

Two goals informing Dryad's current metadata activities include:

1. Dryad's need to be interoperable with other data repositories used by evolutionary biologists; and
2. Dryad's need for a sustainable information infrastructure.

The first goal has inspired our development of the Dryad application profile, version 1.0; and the second goal has led to Dryad's adoption of DSpace software and technology. Current metadata activities for the Dryad development team include revising the project's application profile so that it is compliant with the Singapore Framework. The Singapore Framework is a model that was released at the 2007 Dublin Core conference approximately a year after our team created the DRIADE application profile, version 1.0 (renamed Dryad application profile, ver.1.0) (Carrier, et al, 2007). Ongoing Dryad metadata work also includes evaluating the effectiveness of

³⁹ Note that in some previous publications Dryad is referred to as DRIADE.

⁴⁰ <http://www.nescent.org>

⁴¹ <http://ils.unc.edu/mrc/>

our revised application profile and integrating it into a DSpace environment. This report reviews these two metadata focused activities, and highlights recent accomplishments and challenges.

2. Dryad's Application Profile

Dryad's metadata application profile, ver.1.0, has two modules; one module describes data objects, and the other module describes the associating publication. We developed the application profile to support basic resource and data discovery, with the goal of being interoperable with other data repositories used by evolutionary biologists. The application profile is designed to automatically capture as much metadata as possible during publication and data deposition processing. The application profile incorporates elements from the following established metadata schemes: Dublin Core, Darwin Core, Data Documentation Initiative (DDI), Ecological Metadata Language (EML), and PREservation Metadata Implementation Strategies (PREMIS). The Dryad application profile, ver. 1.0, supports Dryad's phase one functionalities that were established in a stakeholders' workshop in December 2006⁴². These functionalities include the capturing, basic preservation, and simple retrieval of datasets and metadata for associated publications. In the future, metadata elements from other metadata schemes will be needed for projected features. Dryad's phased development and corresponding functionalities are summarized in Table 1.

TABLE 1: Dryad Phased Implementation.

Phased Development/Implementation	Repository Functionalities
Phase One	• basic data/metadata storage • simple submission system
Phase Two	• integrate data deposition with publication • one-stop-deposition • data automatically and manually curated to ensure validity • automated metadata generation

3. DSpace and Dryad's Metadata Architecture

DSpace is a software package for digital repository systems⁴³. DSpace provides basic services to deposit, store, search, and retrieve digital content, but it was designed for a particular use case (storing publications, organized according to a university hierarchy), and significant modifications will be required to make DSpace suit the needs of Dryad users. Although the DSpace infrastructure has been adopted by many repositories, research on the integration of application profiles, especially those complying with the Singapore Framework, is still limited. Implementing the first iteration of the Dryad application profile in DSpace is allowing us to test the application profile, as well as evaluate the long-term applicability of DSpace for Dryad's needs.

DSpace was chosen due to its adaptability and support of Dublin Core metadata, as well as the DSpace community's support for enhancing metadata functionality, as evidenced by developments such as the SKOS module. Although most DSpace functionality revolves around qualified Dublin Core metadata, the software collects additional metadata that can be used to fill in details of the application profile, including qualifiers associated with elements drawn from

⁴² https://www.nescent.org/wg_digitaldata/Dec_5_Workshop_Minutes

⁴³ <http://www.dspace.org/>

other metadata schemes. Metadata fields not native to DSpace are configured as custom fields, which can be stored, searched, and displayed in the same manner as the native fields.

A major advantage of DSpace is its system for managing user accounts, which can be adapted for the eventual Dryad functionality of allowing end-users to submit content and create basic metadata. However, the default workflow for submitting content and generating metadata in DSpace is entirely too long and awkward for end-users, and is further complicated by the needs of the Dryad metadata model. A more configurable submission system is included in the recently released DSpace 1.5, but significant work will still be required to allow users to submit content without difficulty.

One drawback of the DSpace model is that metadata with hierarchical information (e.g., MODS) are not supported by the core repository. Hierarchical information, which is necessary for tracking data such as contact information for multiple authors of a publication, must be stored in an extra file (bitstream) attached to the object, and modifications must be made to the default DSpace functionality if any of this information is to be used beyond simple display.

Another difficulty of using DSpace is the lack of a configurable access control system, a critical feature for Dryad. One requirement of Dryad is to collect and store publications to facilitate automatic metadata generation, while simultaneously shielding these publications from end-users. Some of the content stored in Dryad will need to be placed under embargo. While others have implemented these features in DSpace, the core distribution does not include them. Modifications to the core DSpace code must be kept to a minimum if we are to take advantage of future upgrades. Therefore, it will be challenging to optimize Dryad for users and metadata creators while minimizing deviation from the core DSpace platform.

4. Progressing toward Singapore Framework Compliance

The Singapore Framework provides a model for the structure of Dublin Core application profiles (Nilsson, Baker, & Johnston, 2008). Conformance with the Singapore Framework includes the benefits of consistency, long-term quality control, and interoperability with other metadata structures. A significant effort over the last few months has been to bring the Dryad application profile, ver. 1, which is based largely on Dublin Core, in line with the Singapore Framework. Reasons for this step include the benefits noted by Nilsson, et al. (2008), as well as, our goal to comply and interoperate with Semantic Web standards.

All five Singapore Framework components have been examined for the Dryad metadata schema adaptation (Carrier, 2008). The five components include the following: 1. Functional requirements; 2. Domain model; 3. Description Set Profile; 4. Usage guidelines; and 5. Encoding syntax guidelines. With the exception of the optional encoding syntax guidelines, the other four components have been deemed appropriate for the Dryad's application profile revision. The Scholarly Works Application Profile (SWAP)⁴⁴ is a key example of an application profile in conformance with the Singapore Framework, and provides a model for the Dryad description. The results of the initial restructuring can be found online as part of the repository project wiki.⁴⁵

⁴⁴ http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

⁴⁵ https://www.nescent.org/wg_digitaldata/Level_One_Application_Profile

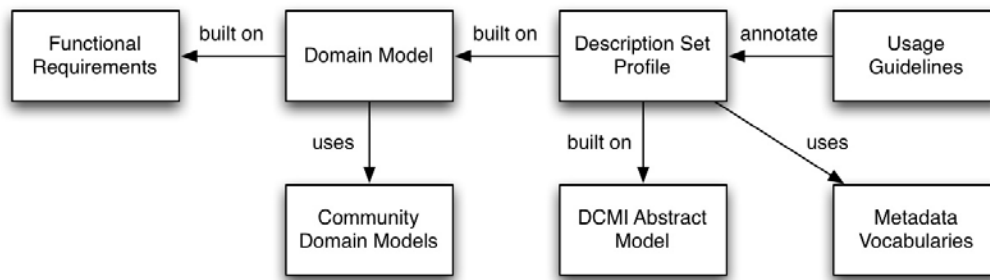


FIG. 1. Graphical representation of the Dryad Application Profile in the Singapore Framework model.

Addressing the Singapore Framework's first mandatory component, Dryad's functional requirements are based on project system requirement specifications. Using the SWAP example as a model, the Dryad's functional requirements (summarized in Table 1) address scope, stakeholders and designated community, requirements gathering, and functional requirements. Dryad's functional requirements include supporting the following operations: 1. resource discovery and use; 2. data interoperability; 3. computer-aided metadata generation and augmentation; 4. linking publications and underlying datasets; 5. data and metadata quality control; and 6. Data security. The designated community for the Dryad application profile includes researchers in the field of evolutionary biology who are generating data and reusing data for their own projects and scientists searching for datasets that are applicable to their own research. Stakeholders are evolutionary biologists, journal publishers in the field of evolutionary biology, professional societies in evolutionary biology, and NESCent. The methodology employed to gather system requirements involved assessing the needs and goals of individuals and groups identified as stakeholders and community members through a workshop held in December 2006 at NESCent in Durham, North Carolina, and more recently an ongoing use case study. Full details about the application profile functional requirements have been added to the Dryad project wiki⁴⁶. The second mandatory component of the Singapore Framework is the domain model. Unlike the SWAP example, the Dryad application profile is "data-centric" rather than document- or publication-centric. Dryad's application profile, ver. 1.0, accommodates a single publication or article with published data from one or more datasets. This relationship is represented in Figure 2.

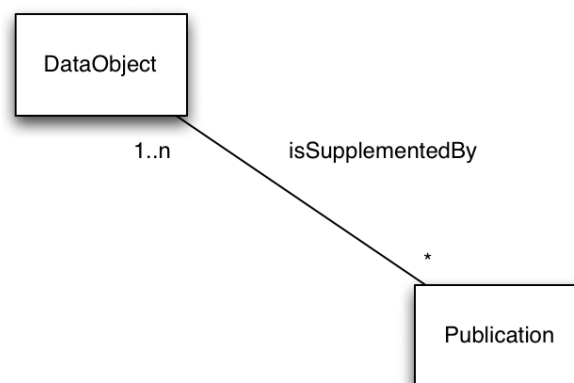


FIG. 2. Dryad Singapore Framework Domain Model

⁴⁶ https://www.nescent.org/wg_digitaldata/Level_One_Application_Profile#Functional_requirements

The third mandatory component, the *Description Set Profile (DSP)* is proving to be the most challenging aspect of the application profile revision process. As previously mentioned, the Dryad application profile is based largely on Dublin Core, but also incorporates elements from domain-specific namespaces such as PRISM, DDI, EML, and DarwinCore. None of the namespaces, except Dublin Core, are currently represented in RDF and cannot be included in the DSP. The Dryad development team has been discussing whether or not to declare unique elements for Dryad use in order to complete the Description Set Profile. Despite this challenge, the first draft of the Dryad DSP, which only includes Dublin Core elements, is available for viewing⁴⁷.

The fourth component, which is optional, is the *usage guidelines*, which have been collaboratively developed by Dryad team members and also appear online. The Dryad usage guidelines provide descriptions of each element and details regarding use⁴⁸. Additionally, the guidelines also elaborate upon the constraints defined by the DSP.

5. Challenges and Future Work

The application profile revisions undertaken to comply with the Singapore Framework has strengthened the overall metadata architecture of the Dryad repository. It has also helped the project team identify key challenges, such as limitations in the current state of citation metadata, and the project's need to encode rights metadata. Furthermore, it has aided the Dryad development team in identifying metadata issues, and clarifying those issues that require administrative or policy decision, prior to determining the appropriate metadata element or value.

The most pressing issue facing the Dryad team is to determine how or if elements from non-Dublin Core namespaces should be included in the Dryad DSP and how the elements will be represented during DSpace implementation. The inclination is to use what has already been determined by a community to be useful, and furthermore to take advantage of the work and documentation already available from other initiatives; however, the issues with interoperability remain unavoidable at this time. Therefore, the Dryad team may choose to declare unique elements for the repository project.

The benefits of moving forward in line with the Singapore Framework are critical to the long-term success of Dryad and its ability to take advantage of metadata to improve system performance. The ongoing revision of the Dryad application profile, ver. 1, will result in the release and publication of the Dryad application profile, ver. 2.0. As part of our application profile development work, we are also taking into account selected functionalities of Dryad's phase two (Table 1). Additional ongoing activities include revising Dryad's interface for entering metadata and streamlining the metadata creation and submission process to support author-depositors. As Dryad evolves, we are anticipating that the recent release of DSpace 1.5 will impact the amount of work the project is able to complete with respect to specific metadata goals and other desired functionalities. In conclusion, Dryad's metadata structure is evolving, and will be revised over time, taking into consideration Semantic Web standards and innovations that support the overall goals of Dryad.

Acknowledgements

This work is supported by National Science Foundation Grant # EF-0423641. We would like to acknowledge contributions by the Dryad team members Hilmar Lapp and Todd Vision of NESCent.

⁴⁷ <http://www.ils.unc.edu/~scarrier/dryad/DSPLevelOneAppProf.xml>

⁴⁸ https://www.nescent.org/wg_digitaldata/Dryad_Level_One_Cataloging_Guidelines

References

- Carrier, Sarah. (2008). *The Dryad Repository Application Profile: Process, development, and refinement*. Retrieved April 24, 2008 from <http://hdl.handle.net/1901/534>.
- Carrier, Sarah, Jed Dube, and Jane Greenberg. (2007). The DRIADE project: Phased application profile development in support of open science. *International Conference on Dublin Core and Metadata Applications, Singapore, 2007*.
- Nilsson, Mikael, Thomas Baker, and Pete Johnston. (2008). *The Singapore Framework for Dublin Core Application Profiles*. Retrieved April 10, 2008, from <http://dublincore.org/documents/singapore-framework/>.