The DRIADE project: Phased application profile development in support of open science

Sarah Carrier
School of Information and Library Science, University of North Carolina at Chapel Hill, USA
scarrier@email.unc.edu

Jed Dube
School of Information and Library Science, University of North Carolina at Chapel Hill, USA
jdube@email.unc.edu

Jane Greenberg
School of Information and Library Science, University of North Carolina at Chapel Hill, USA
janeg@ils.unc.edu

Abstract

DRIADE (Digital Repository of Information and Data for Evolution) is a project being developed for the acquisition, preservation, sharing and re-use of heterogeneous data in support of published research in the field of evolutionary biology. Metadata is a fundamental part of DRIADE's information architecture. This paper reports on DRIADE's overarching goals. We describe our phased approach to developing an application profile, which supports three phases of DRIADE's development. We present a multi-method approach to developing the application profile. Our methods included a *requirements assessment*, *content analysis*, and *crosswalk analysis*. The paper concludes by identifying next steps and discussing the applicability of DRIADE's work to other initiatives seeking to tightly couple published research and data.

Keywords: application profile; open science; data objects; published data; metadata; data sharing; interoperability.

1. Introduction

The Internet is dramatically impacting processes and practices for publishing, distributing, and sharing of scientific research. For example, GRID computing efforts, such as TeraGrid for large scale scientific data, are accessible via web interfaced portals that offer centralized data access and services. Equally significant is the growing number of "small science" digital repositories and initiatives supporting data preservation and sharing. Examples include the Knowledge Network for Biocomplexity (KNB) (http://knb.ecoinformatics.org/) for ecology and the Marine Metadata Interoperability Project (MMI) (http://marinemetadata.org/) for marine biology. The field of evolutionary biology is another scientific discipline in the category of small science that is moving toward an open science model.

Evolutionary biology is an interdisciplinary field drawing from a wide range of scientific disciplines (ecology, paleontology, population genetics, physiology, systematics, and new biological subdisciplines such as genomics). The current publication process requires evolutionary biologists to deposit certain data in specialized data repositories (e.g., GenBank at http://www.ncbi.nlm.nih.gov/Genbank/index.html and TreeBase at http://www.treebase.org/treebase/index.html) and additional supplementary data in journal repositories (e.g., *American Naturalist* at http://www.journals.uchicago.edu/AN/ and *Molecular Biology and Evolution* at http://mbe.oxfordjournals.org/). This requirement advances the preservation of data and makes data accessible for reuse, although the discovery of data via these current venues is limited. The field of evolutionary biology will greatly benefit from a central metadata repository that would bring these disparate data pieces together and link data objects to published research.

The National Evolutionary Synthesis Center (NESCent) (http://www.nescent.org/) recognizes this need and has launched DRIADE (Digital Repository of Information and Data for Evolution)—a repository for published research and data in the field of evolutionary biology. DRIADE is a partnership between NESCent and the School of Information and Library Science, Metadata Research Center (SILS/MRC) (http://ils.unc.edu/mrc/), at the University of North

Carolina at Chapel Hill. Metadata, including an application profile, are fundamental components of DRIADE's information architecture. This paper reports on DRIADE's overarching goals and describes the phased approach to developing DRIADE's application profile, which supports three phases of development. The paper concludes by identifying next steps and discussing the applicability of DRIADE's work to other initiatives seeking to tightly couple published research and data.

2. Identification of Key Components of an Open Science Publication and Data Repository

Expedient and easy access to information via the Internet, compared to the extreme time delays and costs of traditional publication venues, explain, in part, the growing number of digital repositories primed for sharing research. Among one of the most successful and global efforts for sharing research is the Open Archives Initiative (http://www.openarchives.org/), which focuses, primarily, on pre-publication off-prints in addition to electronic dissertations and theses via the Networked Digital Library of Theses and Dissertations (NDLTD) (http://www.ndltd.org/). Creating metadata for electronic "documents" such as prepublications, dissertations, and theses is fairly straightforward**,** drawing from standard bibliographic control practices. Metadata generation becomes more complicated, however, when a repository includes multiple object types, such as "publications" and "data objects", and desires a scheme addressing this type of diversity and even wants to support linking among related objects—beyond subject relationships.

Ongoing discussions of digital object life cycle management and the identification of data types were key components informing the early stages of designing the DRIADE application profile. Repository developers needed to understand the life cycle of all of the items being represented in the repository. Hodge (2000) has identified six stages for digital resources: 1. creation, 2. acquisition, 3. cataloging/identification, 4. storage, 5. preservation, and 6. access. These stages provide a useful framework for understanding the phases of a digital object's life—whether it is a publication or a data set.

We have found that another important requirement in developing a publication/data repository is to define the "types of information objects" and "data types" that are going to be represented or possibly contained in the repository. It is also imperative to identify where linking among objects and data types is desired. In the context of the Dublin Core, linking can be supported by the "relation" element. A preliminary analysis by Carrier, Dube and Greenberg (2007), presented in Tables 1 and 2, provides a framework for helping to build a repository linking published resources and their data.

An initial step in developing the DRIADE project was to identify the above components. This work, combined with DRIADE's goals (discussed in Section 3), helped us determine that we should develop an application profile, and that our resource description and management efforts would benefit from existing metadata developments.

TABLE 1. Object Types.

Publication (e.g., journal article, conference paper)
Published piece of data in the publication (e.g. a table)
Dataset behind the published data (e.g. supplemental data)
Initial data source (e.g., American Ornithologists' Union checklist)
Newly created data (e.g., data derived from any of above)

TABLE 2. Data Types.

Structured labeled data (e.g., tabular data with column and row headings)
Structured unlabeled data (e.g., tabular data *without* column and row headings, or with undecipherable headings)
Unstructured textual data (e.g., readable text)
Unstructured non-textual data (e.g., maps, graphs, images)

3. The DRIADE Project

DRIADE is a collaboration between NESCent (National Evolutionary Synthesis Center) and the SILS Metadata Research Center (MRC) at the University of North Carolina at Chapel Hill. DRIADE is being developed to support data acquisition and ensure long-term preservation of data objects that support published research in the field of evolutionary biology. Overarching goals include promoting resource discovery, data sharing, and data reuse of heterogeneous digital datasets. We have developed a set of functional requirements based on an analysis of existing data repositories (Dube, Carrier & Greenberg, 2007). A synopsis of these requirements include support for the following:

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

4. The Application Profile Approach for DRIADE

A number of namespaces have been developed to serve the data preservation and sharing needs of biologists, including Darwin Core (http://wiki.tdwg.org/DarwinCore) and Ecological Metadata Language (EML) (http://knb.ecoinformatics.org/software/eml/). Despite the existence of these relevant and well-documented schemes, the DRIADE metadata team realized early in the planning stages that a unique modularized application profile was the most logical approach. We use the term "modularized" to describe multiple components within the same application profile. DRIADE's modular (multi-part) scheme provides access to: 1. published research (journal articles and potential conference papers) and 2. data objects supporting the published research. Specific factors impacting our approach included DRIADE's goal to couple published research and supporting data, and the plan to support data preservation, discovery, and reuse. A more detailed discussion of the three key reasons why the application profile approach was selected follows.

First, there are a number of metadata schemes applicable to DRIADE, and can support at least some of the desired functionalities (e.g. data preservation, discovery, use/reuse). For example, the Dublin Core supports resource discovery of published resources (or articles, in DRIADE's case) and selected Dublin Core elements support resource discovery of data objects. It does not make sense to reinvent the wheel, and it has been practical and productive to evaluate and draw from existing schemes.

Second, although there are already schemes developed that support aspects of DRIADE, we have not found a single scheme that satisfies all of the desired functionalities. As indicated above, selected Dublin Core elements support resource discovery of data objects, although it does not support other desired functionalities that are also important to DRIADE, such as preservation and

data integrity. Examples of elements from other schemes important to DRIADE include "fixity" from PREMIS and "depositor" from the DDI.

Third, evolutionary biology is an interdisciplinary field, and a goal of our project has been to create an interoperable environment, one where DRIADE can "shake-hands" with other repositories (e.g., Genbank and TreeBase) used by evolutionary biologists. Zhang (2006) explains, "Establishing metadata interoperability has long-term benefits for resource discovery and retrieval, especially in increasingly interdisciplinary science research."

5. Methodology and Procedures

The DRIADE team used a multi-method approach to develop the modularized, multi-leveled application profile. Our methods included a *requirements assessment*, *content analysis*, and *crosswalk analysis*. We conducted this work following best practices defined by Hillmann (2006), Heery and Patel (2000), and Dekkers (2001).

The *requirements assessment* involved identifying DRIADE stakeholders. Stakeholders include evolutionary biologists, journal publishers in the field of evolutionary biology, professional societies in evolutionary biology, and NESCent—a research center for synthetic research addressing fundamental questions in evolutionary biology. The needs and goals of these individuals and groups were identified at a stakeholders' workshop held in December 2006 at NESCent in Durham, North Carolina. Among initial questions addressed at the workshop were: What is the minimum number of metadata elements required? What functions will the DRIADE scheme support? Answers to these questions have informed the development of DRIADE's functional requirements and the metadata framework.

The *content analysis* involved the application of a social science technique designed for the systematic examination of content (Krippendorf, 2004). We examined various metadata schemes and employed the content analysis methodology to identify relevant elements. For each schema, we asked the following questions in the following sequence:

1. Which schema is being analyzed and what elements are included?
2. How is the schema defined?
3. In what context was the schema designed, and how is it currently applied?
4. How does the context relate to DRIADE? Where would it fit into the application profile, and at what level? What function(s) does the element support? How useful is it for us, useful for the users?

Finally, we conducted a *crosswalk analysis*, which involved the mapping of selected elements from various namespaces (NISO, 2004). Metadata standards selected for the crosswalk included the Dublin Core (http://dublincore.org/documents/dces), PREMIS (http://www.oclc.org/research/projects/pmwg/), EML, DDI (http://www.icpsr.org/DDI/) and Darwin Core. The crosswalk was constructed using established methods (Dekkers, 2001). Intersections in meaning and utilization were noted. During the normalization process, redundancy amongst the chosen elements was eliminated, extraneous elements were discarded, and where possible, Dublin Core elements were chosen. We prioritized the Dublin Core standard because it allows for maximum interoperability and flexible mapping possibilities.

The following steps summarize our application profile development process. We:

1. Reviewed NESCent Stakeholders' workshop outcomes and DRIADE's overall goals.
2. Assessed the information lifecycle for data objects.
3. Researched the use of standards, recommended best practices, case studies, and development processes of several scientific repositories.
4. Identified potential metadata schemes/elements: Dublin Core, Data Documentation Initiative (DDI), Darwin Core, PREMIS and Ecological Metadata Language (EML).
5. Developed a list of metadata required for support of DRIADE's functional requirements.

6. Mapped the required metadata to elements (crosswalk).
7. Chose the metadata elements (Dublin Core where possible--mandatory and required elements from each scheme were considered a priority).

6. Stages of the DRIADE Application Profile

The DRIADE application profile is being developed in three stages to support the three levels of DRIADE's development. An immediate stated goal, coming from the Stakeholders workshop in December 2006, is to preserve as much data as expediently as possible, due to threats of data loss. To address these immediate needs, we have developed the first level of the application profile to serve basic acquisition and preservation functions. Level one of the application profile, addressed in this paper, is in the pre-implementation phase, although we will be implementing this work very soon. We have also been developing levels two and three of the application profile--each with a more granular and sophisticated approach to preserving data and serving the sharing needs of users. The second level of DRIADE's application profile will extend level one functionalities by capturing the complex relationships that exist among data objects. Finally, the third level of the application profile will support "next generation"/Web 2.0 functionalities in the DRIADE repository, including semantic web functionalities. In the next section of this paper, we will review the application profile levels in more detail.

6.1. Level One Application Profile

The level one application profile is intended for initial repository implementation and is currently being refined. At this level, metadata supports preservation, access, and basic usage of data. In order to record and trace the relationships between data sets and published articles, two modules were created within the application profile: the bibliographic citation module and the data object module. Bibliographic information for the published article is linked to the associated data sets via the Dublin Core "relation" and "identifier" elements. Tracking reuse of data sets for further research and publication is accommodated with these two modules, with multiple article Digital Object Identifiers linked to individual data sets across time. Such a structure is also intended to facilitate resource discovery. Throughout level one, automatic metadata generation techniques are employed where possible. Automatic methods are desired because they are more expedient and efficient and less costly than manual approaches, although we are aware of the shortcomings as well (Greenberg, Spurgin & Crystal, 2006). Level one will employ controlled vocabularies for selected elements. Several metadata elements will not be displayed to the public, such as fixity (PREMIS), simply because they are for administrative use. Those elements that are manual have been by and large also designated as optional. This decision is driven by stakeholder concern regarding buy-in and participation. The level one application profile is seen in Table 3.

6.2. Level Two Application Profile

Level two of the application profile is in the planning stages and will satisfy full repository implementation. Building upon the functionalities of level one, this stage of the project will support expanded usage, interoperability, preservation and administration. Most importantly, level two will expand upon the concept of known linkages, namely between data object and publications, as well as capture sophisticated, subtle relationships between the data objects themselves. These relationship instances, called instantiations, are essential for the future functionality of the DRIADE project. Instantiations stem from the information life cycle described above, and are captured by the Dublin Core relation element. At this stage of the development, feedback and assessment strategies will be implemented to determine metadata quality. Furthermore, automation in the process of deposition will streamline user workflow. The establishment of user profiles accommodates this process and will also prepare DRIADE for the third stage of implementation.

TABLE 3. Level One Application Profile.

Namespace:Name /Label	**Obligation**	**Generation Method**	**Occurrences** R=Repeatable NR=Non-repeatable
Module 1: Bibliographic Citation			
dcterms:bibliographic Citation/Citation Information	Required	Automatic	R
dc:identifier/Digital Object Identifier	Required	Automatic	NR
Module 2: Data Object			
dc:creator/Name	Required	Semi-Automatic	R
dc:title/Data Set Title	Optional	Manual	NR
dc:identifier/Data Set Identifier	Required	Automatic	NR
PREMIS:fixity/(hidden)	Required	Automatic	NR
dc:relation/DOI of Published Article	Optional	Semi-Automatic or Automatic	R
DDI:<depositr>/Depositor	Required	Manual, then Automatic after profile creation	NR
DDI:<contact>/Contact Information	Required	Manual, then Automatic	R
dc:rights/Rights Statement	Required	Semi-automatic or Automatic	NR
dc:description/ Description of the Data Set	Optional	Manual	NR
dc:subject/Keywords Describing the Data Set	Required	Manual and Automatic	NR
dc:coverage / Locality	Required	Semi-automatic	R
dc:coverage/Date Range	Required	Semi-automatic	R
dc:software/Software	Optional	Semi-automatic	R
dc:format/File Format	Required	Automatic	NR
dc:format/File Size	Required	Automatic	NR
dc:date/(Hidden)	Required	Automatic	NR
dc:date/Date Modified	Required	Automatic	NR
Darwin Core: species/ Species, or Scientific Name	Optional	Semi-automatic	R

6.3. Level Three Application Profile

Level three of the application profile will support next-generation semantic web functionalities for the DRIADE system, such as:

- Personalization: in which multiple (optional) levels of personalization and recommendation are made possible for users. System functions such as query results, workflow "macros", and user interface could be optimized for individual users.

- User community "virtual societies" utilizing "social tagging" (Web 2.0) functionalities: in which the system's user-members contribute new value to the repository holdings by sharing classifications, evaluations, usages, and other communications.
- Syntactic interoperability for data: in which interrelationships between datasets and data elements are exposed to the users. This can be effected via extensive hypertext linking, "standardization" of data labels and formats, and implementation of emerging standards such as Minimal Information About a Phylogenetic Analysis (MIAPA) (Leebens-Mack et al., 2006).
- Data and collection visualizations: topic clustering and data relationship maps will be developed, and utilized for access, discovery, and system administration.
- User feedback: Extensive collection and analysis of feedback from users will be employed for evaluation purposes, followed by design and development of revisions.

We are in the early phases of developing level there functionalities. A second DRIADE workshop to be held later this year will help us further define our level three goals.

7. Conclusion and Future Work

This paper presents our phased approach to developing a modularized application profile supporting the three phases of DRIADE's development. We reviewed our multi-method approach, which included a requirements assessment, content analysis, and crosswalk analysis, and we presented our level one application profile. DRIADE is being developed for the field of evolutionary biology. Although this interdisciplinary field can be viewed as a specific domain, the techniques and methods are applicable to other application profile initiatives, particularly small science initiatives wanting to link published research and supporting data. We also believe our work has implications for other disciplines beyond the scientific domain, which have similar goals to couple published research and supporting data. A major challenge has been balancing stakeholder interests with the realities of system development. Stakeholder interest led to the phased implementation in an effort to help satisfy their immediate need of data object preservation. Learning about the needs and concerns of this community, as well as exploring their work behaviors, has greatly informed the development of the application profile and other aspects of the system. Contact with the community will help assure a project like DRIADE short-term buy-in and long-term acceptance with such a "ground-up" approach.

Another challenge of our work has been linking the application profile development to the stages of DRIADE's implementation. DRIADE's metadata needs cannot be fully defined without defining each phase of DRIADE's development. And, yet, in some cases, it has been useful to project desired metadata functionalities first, and then take a step back to determine in which phase an activity fits best. For example, the DRIADE team spent a considerable amount of time discussing the metadata element "description" and its place as a level one or level two requirement. While this is considered a simple Dublin Core element by basic abstracting and indexing practices, the DRIADE team determined that asking scientists to write an abstract or description was too labor intensive for level one, where our goal is, partially, to incite cultural change and generate interest. The "description" element will be integrated into either phase two or three of the application profile.

Our next steps include a DRIADE workshop to review phase two of DRIADE's development and further define phase three goals, an experiment testing the level one application profile, and a survey/use case study. The application profile experiment will require evolutionary biologists to create metadata for their data objects with our application profile. The survey/use case study will allow us to gather data from evolutionary biologists about their experience with and support of data sharing, data repositories, and the DRIADE initiative. The workshop and research activities will further inform the development of DRIADE's functional requirements and, in turn, impact the development and the modularized application profile during the later planned phases.

Acknowledgements

This work is supported by National Science Foundation Grant # EF-0423641. We would like to also acknowledge contributions by the DRIADE team members; Todd Vision, Associate Director of Informatics, NESCent, and Assistant Professor, UNC; and Hilmar Lapp, Assistant Director of Informatics, NESCent.

References

Carrier, S., J. Dube, and J. Greenberg. (2007). A metadata application profile for the DRIADE project. Presented at *NESCent, March 13, 2007*. Retrieved July 7, 2007, from http://ils.unc.edu/mrc/driade-project/AppProfile-NESCent-13mar2007.ppt.

Dekkers, M. (2001). Application profiles, or how to mix and match metadata schemas. *Cultivate Interactive, 3*.

Dube, J., S. Carrier, and J. Greenberg. (2007). DRIADE: A data repository for evolutionary biology. *JCDL 2007 Poster*. Retrieved July 7, 2007, from http://ils.unc.edu/mrc/driade-project/jcdl2007.pdf (temporary URL).

Greenberg, J., K. Spurgin, and A. Crystal. (2006). Functionalities for automatic-metadata generation applications: A survey of metadata experts' opinions. *International Journal of Metadata, Semantics, and Ontologies*, *1*(1), 3-20.

Heery, R., and M. Patel. (2000). Application profiles: Mixing and matching metadata schemas. *Ariadne, 25*.

Hillmann, D. (2006). Application profiles: A tutorial. *International Conference on Dublin Core and Metadata Applications* (pp. 3-6).

Hodge, Gail M. (2000). Best practices for digital archiving: An information life cycle approach. *D-Lib Magazine, 6*(1).

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd). Thousand Oaks, CA: Sage.

Leebens-Mack, J., et al. (2006). Taking the first steps towards a standard for reporting on phylogenies: Minimum information about a phylogenetic analysis (MIAPA). *OMICS: A Journal of Integrative Biology, 10*(2), 231-237.

NISO (National Information Standards Organization). (2004). *Understanding metadata*. Bethesda, MD: NISO Press. Retrieved July 7, 2007 from http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

Zhang, Xiaolin. (2006). Cross-domain metadata interoperability to support integrated digital service environment. *20th International CODATA Conference: Scientific Data and Knowledge within the Information Society, 22-25 October 2006, Beijing.*