Describing Scholarly Works with Dublin Core: A Functional Approach

Julie Allinson

Abstract

This article describes the development of the Scholarly Works Application Profile (SWAP)—a Dublin Core application profile for describing scholarly texts. This work provides an active illustration of the Dublin Core Metadata Initiative (DCMI) "Singapore Framework" for Application Profiles, presented at the DCMI Conference in 2007, by incorporating the various elements of Application Profile building as defined by this framework—functional requirements, domain model, description set profile, usage guidelines, and data format. These elements build on the foundations laid down by the Dublin Core Abstract Model and utilize a preexisting domain model (FRBR—Functional Requirements for Bibliographic Records) in order to support the representation of complex data describing multiple entities and their relationships. The challenges of engaging community acceptance and implementation will be covered, along with other related initiatives to support the growing corpus of scholarly resource types, such as data objects, geographic data, multimedia, and images whose structure and metadata requirements introduce the need for new application profiles. Finally, looking to other initiatives, the article will comment on how Dublin Core relates to the broader scholarly information world, where projects like Object Re-use and Exchange are attempting to better equip repositories to exchange resources.

Introduction

This article describes work in 2006 to devise a metadata application profile for describing scholarly works. The work was funded by the UK's Joint Information Systems Committee (JISC) (http://www.jisc.ac.uk/) to solve

issues with metadata quality and consistency identified by a number of previous JISC projects and to provide richer, more functional metadata for use in the Intute Repository Search project, a National initiative to provide a comprehensive aggregation and search service for UK repositories. JISC recognized both that metadata quality would have a major impact on the success and efficacy of this project and that, with repositories becoming more widespread, metadata has a key role to play in the discovery, exchange, and reuse of scholarly information in the widest sense. The article examines the process of constructing an application profile within the developing Dublin Core framework for such activities and also looks in some detail at SWAP itself, highlighting why it was developed and what decisions were taken along the way.

For repositories, metadata is a valuable asset that needs to be shared with external systems. For its internal metadata, a repository or other metadata-rich system need only consider its own requirements. Yet, it is when we start to think about sharing metadata and objects between systems that application profiles become a crucial consideration. This is central to the SWAP work and to the existence of metadata standards and profiles. Without agreement on standards, without consistent approaches, sharing information would be a laborious mapping process and users would be presented time and again with new and conflicting information on non-standard interfaces. Confusion abounds.

JISC initially scoped and defined work on a new application profile for describing EPrints. They commissioned it outside the bounds of a specific project in order to free the work of associations with time-limited project activities, hoping that the profile would become embedded into the repositories community. Initially UK-focused, the authors recognized from the outset that for genuine uptake the profile ought to be of international relevance. The outputs and processes used have indeed engendered interest from across the world and a Dublin Core Scholarly Communications community has been established to foster interest in this and other scholarly metadata (DCMI, 2007). This community currently numbers some 195 members. SWAP is also being cited as an exemplar within DCMI for application profile development as will be discussed later in this article.

SWAP is not without challenges and many of these will be discussed further in the ensuing sections. Uptake has, so far, been slow, perhaps because of the seeming complexity of the SWAP model, the movement away from flat metadata descriptions and the demands it places on repositories to review their current practice. The profile may need future revision and further work to embed and test its use. Yet SWAP was created to tackle real issues inherent in using metadata that is constrained by its simplicity and, as such, introduces some necessary complexity that looks to a future where metadata more effectively and efficiently describes the resources about which repositories need to share information.

Issues with Simple Dublin Core

Dublin Core and DCMI, has become synonymous with "Simple Dublin Core"—the fifteen elements defined its Dublin Core Metadata Element Set (DCMI, 2008). Simple DC records are mandated for exchange by the OAI-PMH harvesting protocol and understood by many systems on the Web. Yet a number of projects and services such as OAIster (http://www.oaister.org/) have cited the poor quality and inconsistency within these simple metadata descriptions as barriers to providing richer services. Some of these issues were summarized by the ePrints-UK project final reports:

> Simple DC is not targeted at describing eprints specifically so there is more to the description of an eprint than simple DC will allow. To get round these limitations of simple DC, some repositories try to put more information than necessary into the Dublin Core fields. This varying use of metadata can lead to difficulties for end-users who are trying to discover eprints across multiple repositories. (Powell, Day, & Cliff, 2005)

Similar conclusions were reported by the PerX project (MacLeod, 2007), which implemented a cross-search for resources in the engineering subject discipline.

Poor quality and inconsistent metadata are a real problem for aggregator services, for a variety of reasons. Within the work on SWAP, these were analyzed to help inform the requirements specification process. Identification is a particular issue, with inconsistent practices employed in the use of <dc:identifier>, <dc:relation>, and <dc:source> to capture identifiers for full-text resources, metadata records and other related resources in a way that cannot be easily disambiguated. Additionally, simple Dublin Core does not allow for the specification of Syntax Encoding Schemes, so where a particular identifier scheme has been used, for example, URI, DOI, ISBN, these cannot easily be identified. Other issues include: use of multiple <dc:title> elements and the inability to specify the language of the element contents, that is, for translated titles; where name elements such as <dc:creator> and <dc:publisher> are used it is not possible to indicate whether a normalized form has been used or whether the name is of a person or organization; use of <dc:date> can be ambiguous without the ability to express what kind of date, that is, publication, modification; and where <dc:subject> is used there is no means of indicating whether controlled terms from a specific vocabulary have been used.

Brief Background to SWAP

It is with these, and other, issues in mind that the SWAP activity was funded. Originally known as the Eprints Application Profile, the name was later changed to avoid confusion with the EPrints repository software. SWAP is software neutral. For those familiar with the earlier work, SWAP and the Eprints Application Profile are synonymous. It is an important point,

though, that misunderstandings about terminology can genuinely affect uptake and interest—a term used by one community can be used in quite a different way by another—a significant challenge for metadata creators and application profile developers and for communities like DCMI.

JISC provided the initial scope for the work, which was examined and expanded in the functional requirements work. From the outset it was decided that the profile should be grounded in Dublin Core principles and, wherever possible, use Dublin Core metadata properties. JISC is a DCMI affiliate and has a long history of supporting work on DCMI and inputting into its goals and outputs so it was a natural choice.

Identifying what we meant by *eprints* or *scholarly works* was an important first step, and for this we used the definition provided by the Budapest Open Access Initiative (n.d.) of "a scientific or scholarly research text, for example a peer-reviewed journal article, a preprint, a working paper, a thesis, a book chapter, a report, etc." (Suber, 2007). Interestingly, this list included theses, materials for which dedicated application profiles exist. There was no intention here to replace the richer, dedicated profiles available, merely to allow theses to be described along with other scholarly works if so desired, while leaving repositories free to expose their metadata using other metadata formats as necessary.

Also integral to the scope of the work was the need to support the improved availability of open access resources. Open access is defined as:

> free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. (Suber, 2007)

JISC engaged Andy Powell from the Eduserv Foundation and Julie Allinson, then from UKOLN, to lead on the work. We devised a program of work for the next three months, assembled a working group of invited experts and began to develop deliverables in the open, collaborative arena of the UKOLN Repositories Research Team wiki (JISC, 2008, October 16). A decision was taken early in the process not to establish a formal DCMI group for this work, principally because of the short timescale allotted by JISC. By August 2006, the project had produced the following deliverables:

- Functional requirements
- Application model
- Application profile (with embedded usage guidelines)
- XML schema
- "Dumb-down" guidelines
- Community acceptance plan

Each of these will be discussed and illustrated in the following sections.

Being Part of a Dublin Core Framework

At its core, SWAP is a Dublin Core application profile, but what does this mean? Dublin Core is perhaps still most often associated with the fifteen core properties that comprise the DCMI Element Set (ISO 15836) (DCMI, 2008), known as *Simple DC*. Any Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant repository will be familiar with exposing simple Dublin Core records in the oai_dc format. Yet Dublin Core is much more than the original fifteen, and increasingly it can support flexible and extensible metadata that is compatible with the Semantic Web, metadata, which can capture added information, references, vocabulary details, etc. Application profiles, which can build on this richness and draw terms from other schemes, fit well and are being formalized by the work of the DCMI Usage Board and Architecture Forum.

The Scholarly Works Application Profile predates the Dublin Core Conferences both in Mexico (2006) and Singapore (2007) and was presented in some form at both. At the latter, Mikael Nilsson introduced a framework for Dublin Core application profiles, now known as the Singapore Framework. This framework offers the following definition of an application profile (Nilsson 2007b): "A DCAM [Dublin Core Abstract Model]-conformant Application Profile ('DC Application Profile') is a packet of documentation that consists of:

- Functional requirements (mandatory)
- Domain model (mandatory)
- Description Set Profile (DSP) (mandatory)
- Usage guidelines (optional)"

This is illustrated in more detail at figure 1, which also outlines the standards on which the elements of the framework draw.

SWAP acts as an example of an application profile that complies to the above framework. In the ensuing sections, each element of this framework will be considered in more detail, including the Description Set Profile—an addition to the original suite of deliverables.

Devising a Functional Approach to Functional Requirements

There were four steps to developing functional requirements for SWAP: considering who benefits and who is interested—our stakeholders and community; reviewing literature, current practice, and other work in the area; developing scenarios and gathering user requirements; and turning these into a requirements specification to guide the next work steps.

For SWAP, our primary stakeholder community included implementers of a UK Institutional Repositories search service (Intute, n.d.), managers and administrators of UK repositories (those containing scholarly works), developers of repository software such as EPrints, DSpace, and
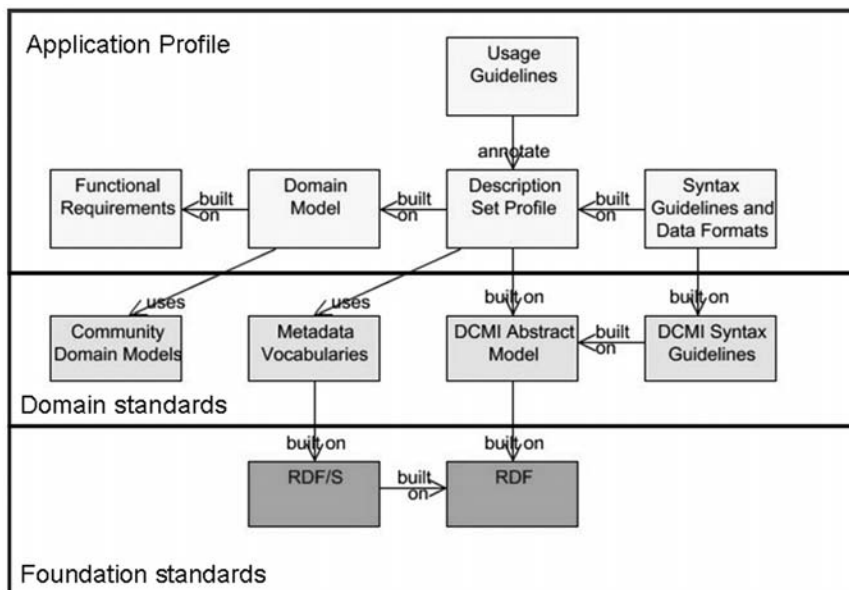
Figure 1. The Singapore Framework for DC Application Profiles

Fedora, implementers of the Depot repository (Prospero, 2006) and users of any of the above systems. Representatives of all of these communities needed to be engaged in order to ensure that the application profile is developed with its users needs in mind. All groups were represented on our project working and feedback groups. Communities such as the JISC Development Community, the international repositories community and the DCMI community were all engaged and encouraged to provide expertise and feedback.

Scenarios and Requirements
Our approach to the requirements specification was to identify a requirement, back it up with a brief illustrative usage scenario reflecting either limitations in current practice or desired functionality, and from this propose a solution. The following example elucidates:

- *Requirement:* Enable identification of the research funder and project code
- *Usage Scenario:* A research funder has mandated deposit of materials into repositories. In order to check this using automated means, a repository must offer details of the funder and project code for works associated with a particular funded piece of research.
- *Proposed Solution:* Funder and Grant Number properties

Our primary use case was "to develop an application profile for eprints

(scholarly works) to be used by the Intute UK repositories search service to aggregate content from repositories and that in so-doing the search service can offer a richer, better experience for users and potentially added-value services in future."[1]

Overall, we identified around thirty requirements and for each outlined a usage scenario and proposed a solution. The full list is available from the SWAP wiki pages (JISC, 2008, May 14b), but there are three areas of functionality that deserve further analysis here.

Identification: The application profile MUST implement an unambiguous method of identifying full-text(s). Why? Because services interrogating the metadata need to be able to distinguish between a link to a "splash page" or metadata record, or to the full text, so that they can perform full-text searching/indexing, alert users to the presence of metadata-only records, or make explicit statements about the location of resources. SWAP, with its multidescription model described in the next section, allows for this by identifiers being assigned explicitly to the resource that they are describing. In the same area, recommending the use of open URLs as an identifier for bibliographic citations allows machines to utilize open URL services to direct users to available materials.

Versioning: The application profile MUST offer a preliminary recommendation for version identification. Why? Because increasingly repositories are faced with different versions of research papers, for example a preprint and a postprint. In current practice, it is often impossible to tie together different versions or to answer questions such as: which of these is peer-reviewed, which is the most recent, and are these two articles the same? SWAP, again with its multidescription model allows for the capture of metadata for different versions and for the relationships that tie different versions together.

Open Access: The application profile MUST facilitate identification of open access materials. Why? Because a user wants to know whether they can actually access the resource discovered through a particular search tool. SWAP, by utilizing a preexisting metadata property and establishing a short vocabulary for access rights, enables unambiguous identification of open access materials.

What was clear at this stage is that the requirements demanded a more complex model than simple flat metadata structures could provide.

Modelling the Metadata

Why do we need a "model" for metadata? What is an application, or domain, model?

A domain model, or application model, is a conceptual model identifying the entities we want to describe, the relationships between them and the attributes necessary to effectively describe the entities. It acts as a communication tool and should be understandable by both technical and non-
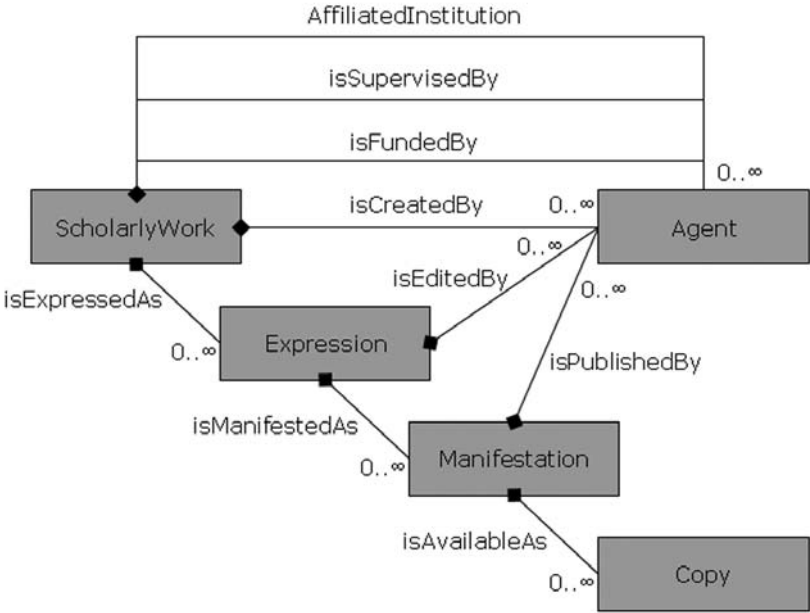
Figure 2. The Scholarly Works Application Profile Model

technical audiences. In fact, it has a role to play in providing shared understanding between different stakeholders. The model can be expressed in any way, such as a text description or a set of illustrative diagrams, or a mixture of the two. It is not intended to be machine-readable at this point although using a language like UML (Unified Modelling Language) means that it can be more easily expressed in a machine-readable way.

For SWAP, our model takes the form of a lightweight entity-relationship model expressed as a simple UML diagram, shown above. The diagram is accompanied with supporting textual information. UML is an object modeling and specification language widely used in software engineering that offers a set of definitions for a range of diagrams. The purpose of a model like this for developing application profiles is that it allows the author to better examine what is being described: the entities, relationships, and attributes. By taking this approach it is much easier to base metadata on requirements rather than on the constraints of a particular metadata solution. This approach also moves us away from the idea that metadata descriptions are to be single and flat. Traditionally metadata descriptions often contain information about a number of entities in a single description. With an entity-relationship approach, the potential for describing multiple entities and relating these to each other can be explored. It is

worth noting that at this point we are not breaking the metadata rule of one description for one resource. Later in the article it will be demonstrated how the Dublin Core Abstract Model can be used to facilitate the capture of multiple related descriptions of entities within a description set. At the modeling stage, this approach is free from dependence on a particular metadata approach or metadata vocabulary. Indeed, at this point is it important to note that we are in no way tied to using Dublin Core metadata terms—metadata vocabulary decisions come later.

Community Domain Models

There are a number of domain models that already exist for different communities and purposes, such as the CIDOC-CRM model (http://cidoc .ics.forth.gr/) for describing complex museum objects and capturing detail about provenance and the relationships between physical object and digital representation. For the Scholarly Works Application profile, we identified a couple relevant items. These were the Functional Requirements for Bibliographic Records model (FRBR) (IFLA, 1998) and the Common European Research Information Format (CERIF) (http://www .eurocris.org). The names of each demonstrate their particular focus— FRBR on bibliographic materials and library catalogs, CERIF on research information and research information systems. Work in Denmark on the DDF-MXD metadata format (DEFF, 2006) has used CERIF as the basis of a domain model to support a Danish research information service. For Scholarly Works, FRBR seemed to map more closely to the requirements identified.

FRBR is a domain model for the entities that bibliographic records describe. It defines a set of four primary entities: work, expression, manifestation, and item. In addition, there are two agent entities (corporate body and organization), and a set of "subject" entities (concept, object, event, place). The primary FRBR entities are defined as follows:

> work: a distinct intellectual or artistic creation. A work is an abstract entity; there is no single material object one can point to as the work. We recognize the work through individual realizations or expressions of the work, but the work itself exists only in the commonality of content between and among the various expressions of the work. (IFLA 1998, p. 16)

> expression: the intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, image, object, movement, etc., or any combination of such forms. An expression is the specific intellectual or artistic form that a work takes each time it is "realized." (IFLA 1998, p. 18)

> manifestation: the physical embodiment of an expression of a work. (IFLA 1998, p. 20)

> item: a single exemplar of a manifestation. The entity defined as item is a concrete entity. (IFLA 1998, p. 23)

Between all of these are relationships, and FRBR (Functional Requirements for Bibliographic Records) defines quite an extensive set of these, chief among which, in terms of SWAP, appear to be the following:

- Work—is realized through —> Expression
- Expression—is embodied in —> Manifestation
- Manifestation—is exemplified by —> Item
- Work—is created by —> Person or Corporate Body
- Manifestation—is produced by —> Person or Corporate Body
- Expression—has a translation —> Expression

Other relationships capture fine-grained relationships between entities, for example, between works, there can be relationships for imitation, transformation, complement, summarization, successor, complement, and more.

For bibliographic catalogs, the power of FRBR lies in the ability to group items logically and to facilitate the discovery of all instances of a particular work in a single search, while being able to distinguish between the different expressions, manifestations, and items and to navigate easily to the most appropriate.

FRBR-IZING SWAP

As previously stated, the SWAP model is based on FRBR. It is a simplification of FRBR as we have used fewer entities, relationships, and attributes. The reason for this being that our requirements didn't demand the same complexity or detail. FRBR is a useful model in the context of eprints because it allows us to answer questions like: What is the URL of the most appropriate copy (an item) of the PDF format (a manifestation) of the pre-print version (an expression) for this eprint (the work)? Or, are these two copies related? If so, how?

The SWAP model modifies FRBR in a small number of ways. In particular, as can be seen in figure 2, above, the Work entity has been renamed ScholarlyWork in order to demonstrate the refinement of FRBR. ":Item" has become "Copy," a term much more appropriate in the digital realm, where each digital object is essentially copied when retrieved by a user. The SWAP relationships have been so named for clarity and to fit in more seamlessly with Dublin Core as will be further explored in the next section. Figure 2 shows the vertical relationships in the model, yet there are also some horizontal relationships expressed in figure 3. Table 1 shows how these relationships are not intended to be a comprehensive set of relationships between entities, but they do reflect the most important relationships that exist between scholarly works, their expressions and

manifestations. Since the most common forms of Expression of an eprint are the various revisions that it might go through and its different translations, these are the main relationships captured by the model. SWAP worked closely with the UK Versions project that has identified the following version names for journal articles: Draft, Submitted Version, Accepted Version, Published Version, and Updated Version. This vocabulary has been developed since SWAP was completed and it may prove desirable that the new Dublin Core Scholarly Communications community might investigate changes to SWAP to align with the Versions outputs.
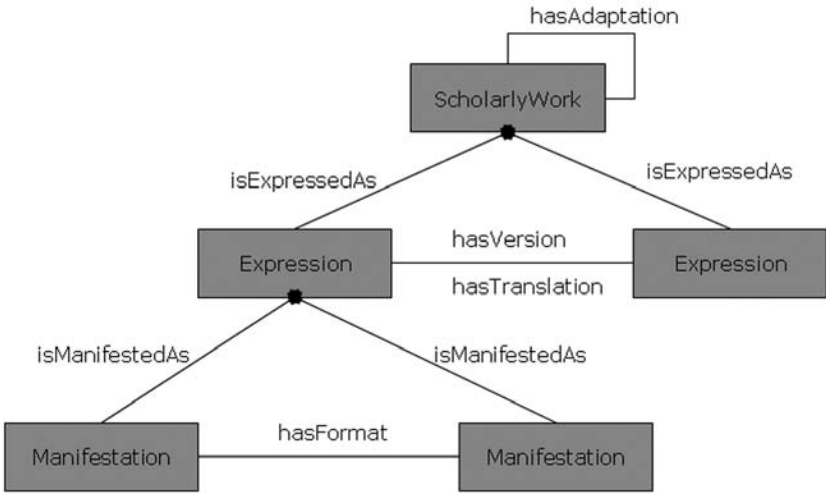
Figure 3. Horizontal relationships in the Scholarly Works Application Profile Model

The following is a list of relationships used in the SWAP model:

Table 1. SWAP Relationships

SWAP Relationship	FRBR Relationship
isExpressedAs relationship	Is realized through
isManifestedAs relationship	is embodied in
isAvailableAs relationship	is exemplified by
isCreatedBy relationship	is created by
isPublishedBy relationship	publisher attribute of a Manifestation
hasAdaptation	has adaptation
hasVersion	has a revision of
hasTranslation	has a translation of

In natural language, what the above model says is: A ScholarlyWork is expressed as zero or more Expressions. Each Expression is manifested as zero or more Manifestations. Each Manifestation is made available as zero or more Copies. Each ScholarlyWork has zero or more creators, funders, and supervisors. Each Expression has zero or more editors. Each Manifestation has zero or more publishers.

The above statement diverges from FRBR in quite a significant way, in that FRBR demands the existence of one or more expression, manifestation, and item. This issue was discussed during and after the development of SWAP and many felt this insistence on one should be retained. This remains contentious and SWAP is still open to discussion on this point but the authors felt that it should be possible to create metadata for a ScholarlyWork alone, to allow for an "idea" to be described, before any physical or digital expressions of that ScholarlyWork had come to being. Testing of the profile in repositories should help explore whether this is the correct decision or not.

The next step in developing the domain model is to identify the key attributes that will be used to describe each entity in the model. These attributes, along with the relationships defined above, are realized as metadata properties in the application profile proper. As stated earlier, at this stage we are agnostic about the metadata vocabulary from which these are taken, for example, title of a ScholarlyWork is one attribute, but this does not have to be expressed as a Dublin Core title property. It is in the next step of profile development that these decisions are taken.

Capturing This with Dublin Core
The model outlined above is based around describing a number of entities. It is the Dublin Core Abstract Model that allows us to do this, by introducing the notion of "description sets"—a group of "descriptions" of individual entities. The Dublin Core Abstract Model is illustrated in figure 4.

There is more to the DCAM, and readers are encouraged to consult the DCAM documentation available from the Dublin Core web site (Powell, Nilsson, Naeve, Johnston, and Baker (2007) for further information. For the purposes of this paper, it is enough to state that the DCAM enables us to create a description set containing descriptions of each of our entities and to capture, as statements containing property/value pairs, the set of relationships and attributes defined by the SWAP model.

The following table shows the full list of metadata properties used to capture both the attributes and relationships in the model. The table also identifies the vocabulary encoding schemes (VES) used both for metadata properties, including Dublin Core, Friend of a Friend (FOAF)[2] and new eprint metadata properties, and for the capture of specific metadata values, such as the eprint entity type and access rights vocabularies. Syntax
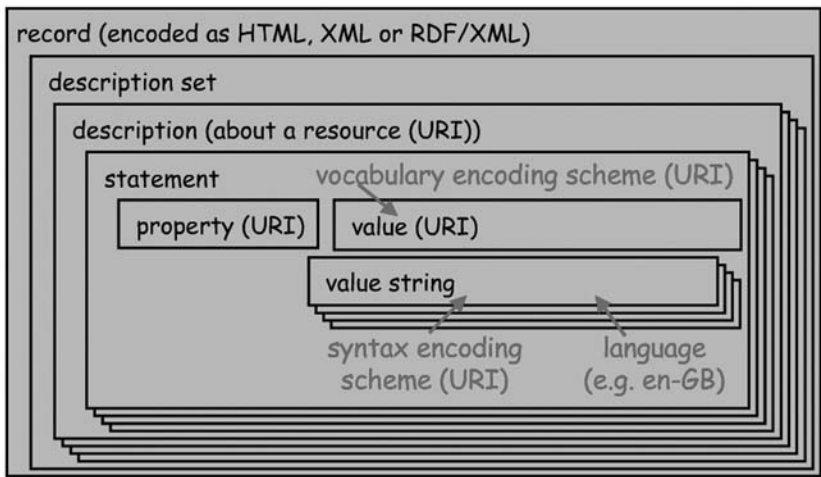
Figure 4. The Dublin Core Abstract Model illustrated

Encoding Schemes (SES) are used to standardize the use of specific syntax for properties such as language, date, and identifiers to ensure consistency.

Wherever possible, metadata properties have been taken from either the Dublin Core Element Set ("simple" Dublin Core) or Dublin Core Metadata Terms (often called "qualified" Dublin Core). Dublin Core cannot provide all of the terms needed to describe scholarly works in such a way as to fulfill our functional requirements. Two further metadata vocabularies have supplied properties: the MARC relator codes supply us with standard terms for "agent" roles, while the Friend of a Friend scheme supplies properties for describing agents and brings some semantic Web flavor to the profile. A small number of properties were defined anew. By declaring these within the "eprint" namespace they are open for further reuse in other application profiles. Table 2 shows the properties used and the Encoding Schemes used for the properties and their values.

Using metadata vocabularies allows us to standardize the properties used to describe the SWAP entities. Additionally, specifying a set of vocabularies has allowed the capture of consistent metadata values for resource type, entity type, status, and access rights. Figure 5 shows how the type vocabulary extends the DCMIType "Text" value to provide a richer set of terms for describing scholarly works.

The Access Rights vocabulary contains only three terms: open, restricted, and closed, and provides a very simple mechanism for supporting the requirement to pinpoint open access materials.

Table 2. Metadata Properties for SWAP[3]

Metadata Property/Attribute	Metadata Property VES	Value VES/SES
ScholarlyWork		
entity type	dc:type	eprint:EntityType (VES)
title	dc:title	
subject	dc:subject	
abstract	dcterms:abstract	
grant number	eprint:grantNumber	
has adaptation	eprint:hasAdaptation	
identifier (URI)	dc:identifier	dcterms:URI (SES)
creator	dc:creator	
funder	marcrel:FND	
supervisor	marcrel:THS	
affiliated institution	eprint:affiliatedInstitution	
is expressed as	eprint:isExpressedAs	
Expression		
entity type	dc:type	eprint:EntityType (VES)
title	dc:title	
description	dc:description	
date available	dcterms:available	dcterms:W3CDTF (SES)
status	eprint:status	eprint:Status (VES)
version number or string	eprint:version	
language	dc:language	dcterms:RFC3066 (VES)
genre/type	dc:type	eprint:Type (VES)
copyright holder	eprint:copyrightHolder	
has version	dcterms:hasVersion	
has translation	eprint:hasTranslation	
bibliographic citation	dcterms:bibliographicCitation	kev:ctx (SES)
references	dcterms:references	kev:ctx (SES)
identifier (URI)	dc:identifier	dcterms:URI (SES)
editor	marcrel:EDT	
is manifested as	eprint:isManifestedAs	
Manifestation		
entity type	dc:type	eprint:EntityType (VES)
format	dc:format	dcterms:IMT (VES)
date modified	dcterms:modified	dcterms:W3CDTF (SES)
publisher	dc:publisher	
is available as	eprint:isAvailableAs	
Copy		
entity type	dc:type	eprint:EntityType (VES)
date available	dcterms:available	dcterms:W3CDTF (SES)
access rights	dcterms:accessRights	eprint:AccessRights (VES)
licence	dcterms:licence	
is part of	dcterms:isPartOf	
identifier/locator (URI)	dc:identifier	dcterms:URI (SES)
Agent		
name	foaf:name	
family name	foaf:family_name	
given name	foaf:givenname	
type of agent	dc:type	eprint:EntityType (VES)
workplace homepage	foaf:workplaceHomepage	
mailbox	foaf:mbox	
homepage	foaf:homepage	
identifier (URI)	dc:identifier	dcterms:URI (SES)

```
http://purl.org/dc/dcmitype/Text
 ↳ http://purl.org/eprint/type/ScholarlyText
    ↳ http://purl.org/eprint/type/Book
    ↳ http://purl.org/eprint/type/BookItem
    ↳ http://purl.org/eprint/type/BookReview
    ↳ http://purl.org/eprint/type/ConferenceItem              Key
        ↳ http://purl.org/eprint/type/ConferencePaper        ↳ = sub-class
        ↳ http://purl.org/eprint/type/ConferencePoster
    ↱ http://purl.org/eprint/type/JournalItem
        ↳ http://purl.org/eprint/type/JournalArticle
    ↱ http://purl.org/eprint/type/NewsItem
    ↳ http://purl.org/eprint/type/Patent
    ↳ http://purl.org/eprint/type/Report
    ↳ http://purl.org/eprint/type/SubmittedJournalArticle
    ↳ http://purl.org/eprint/type/Thesis
    ↳ http://purl.org/eprint/type/WorkingPaper
```

Figure 5. The Eprints Type Vocabulary Encoding Scheme

Decisions about which metadata property was needed at each entity level were not always straightforward. In some cases decisions were taken which might be argued against, for example, the SWAP authors decided that copyrightHolder, and thus copyright ownership, is an attribute of the Expression. Publication was viewed as something that happened to the manifestation, the "format," rather than the intellectual expression, and thus publisher is captured at the manifestation level. Authorship, title, and subject seemed fairly clearly associated with the ScholarlyWork as a whole yet even these can be contentious—What happens to foreign titles of specific expressions? What if an additional author contributes to a particular expression? We live in an imperfect world and SWAP attempts to model for the most common cases while accepting that there will be gray areas and unanticipated situations. Defining what constitutes an expression and what a new ScholarlyWork is one area where it is difficult to mandate—community input and discussion are needed to best agree on approaches here.

Usage Guidelines and the Description Set Profile

In the Scholarly Works Application profile, the usage guidelines are contained within the profile documentation itself (JISC, 2008, May 15) and it is here that we can document all of the terms, provide guidance, and give examples on how they should be used. The SWAP guidelines also

provide a link to the eprint terms document which defines all of the new "eprint" terms and vocabularies (JISC, 2008, May 14a). For the framework outlined above, the usage guidelines are optional and, indeed, how much or little additional guidance given about the profile is down to the profile authors and its user community needs. For other applications it might be that more complex usage guidelines demand separate documentation.

MAKING APPLICATION PROFILES MACHINE READABLE

So far, all of the profile elements discussed have been entirely human-readable. This is a worthy and necessary thing, but not enough for computers to do useful things with, without unnecessary programming effort. Metadata records are most often encoded in XML, with a formal schema to facilitate validation of records for "correctness." Yet there is no similar way of validating application profiles. Without this, an application cannot check that the profile is valid in terms of its use of the Dublin Core Abstract Model, or its use of metadata vocabularies and terms. With machine-readability, an application can make explicit decisions based on the profile upon which its metadata is based, for example it can know that a particular metadata property is a literal (a lexical string) or a nonliteral identifier for a separate resource, that only one such property should be stored, or that a date value will be encoded according to W3CDTF.

While the usage guidelines and the embedded links provide all of the constraints for the Scholarly Work Application Profile, this is not formally defined and cannot be meaningfully interpreted. This is the motivation for the new Dublin Core Description Set Profile work by Mikael Nilsson, described as "a formal representation of the constraints of a Dublin Core Application Profile" that can be used as

- as configuration for databases
- as configuration for metadata editing tools
- etc. (Nilsson, 2007c)

Figures 6 and 7 illustrate how the data from the existing Scholarly Works Application Profile documentation has been re-rendered, with embedded formatting. This is then transformed and rendered in an XML format, as illustrated in Nilsson (2007b). This XML format bridges the gap between the human-readable application profile documentation and the machine-readability of XML and, in the future, should allow applications to more efficiently construct metadata creation and storage procedures.

SYNTAX GUIDELINES AND DATA FORMATS

At the time of writing (January 2008), the DCMI is working on a revised XML format to support the serialization of DC description sets as described by the DCMI Abstract Model (DCAM). Since SWAP utilizes the richness of the Dublin Core Abstract Model, which cannot be expressed

Figure 6. The Scholarly Works Application Profile reexpressed

Figure 7. The underlying syntax used to create the description set profile

with current guidelines, it required a custom XML format. This was cre-ated by Pete Johnston and is known as Eprints DC XML. The format is based closely on the then drafts of the DCMI XML format. Figure 8 shows an example instance of the Eprints DC XML format. A W3C XML Schema and a RELAX NG Schema for Eprints DC XML are also available (JISC, 2008, May 14a).

"Dumbing Down"

It is a requirement of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) that for each "item" in a repository, the repository must support the dissemination of metadata records in the "oai_dc meta-data format." "oai_dc" is a format defined by the OAI-PMH specification to serialize Simple Dublin Core (DC). In the language of the DCMI Ab-stract Model, this equates to a description set comprising a single descrip-tion, containing statements that reference one of the fifteen properties of

```
<ep:descriptionSet
   xmlns:ep="http://purl.org/eprint/epdcx/2006-11-16/"
>
```

```
    <ep:description
      ep:resourceURI="http://eprints.gla.ac.uk/503/" >
      <ep:statement
        ep:propertyURI="&dc;type"
        ep:valueURI="&epentType;ScholarlyWork" />

<!-- more statements describing Scholarly Work -->

      <ep:statement
        ep:propertyURI="&eprint;isExpressedAs"
        ep:valueRef="expression1" />
    </ep:description>
```

Description of
Scholarly Work

```
    <ep:description
      ep:resourceId="expression1" >
      <ep:statement
        ep:propertyURI="&dc;type"
        ep:valueURI="&epentType;Expression" />

<!-- more statements describing Expression -->

      <ep:statement
        ep:propertyURI="&eprint;isManifestedAs"
        ep:valueRef="pdfmanifestation" />
    </ep:description>
```

Description of
Expression

```
    <ep:description
      ep:resourceId="pdfmanifestation" >
      <ep:statement
        ep:propertyURI="&dc;type"
        ep:valueURI="&epentType;Manifestation" />

<!-- more statements describing Manifestation -->

      <ep:statement
        ep:propertyURI="&eprint;isAvailableAs"
        ep:valueURI="http://eprints.gla.ac.uk/503/01/Eu_J._Hum_Gen.9(2)143_.pdf" />
      <ep:statement
        ep:propertyURI="&eprint;isAvailableAs"
        ep:valueURI="http://www.nature.com/ejhg/journal/v9/n2/pdf/5200590a.pdf" />
    </ep:description>
```

Description of
Manifestation

```
    <ep:description
      ep:resourceURI="http://eprints.gla.ac.uk/503/01/Eu_J._Hum_Gen.9(2)143_.pdf" >
      <ep:statement
        ep:propertyURI="&dc;type"
        ep:valueURI="&epentType;Copy" />

<!-- more statements describing Copy 1 -->

    </ep:description>
```

Description of
Copy 1

```
    <ep:description
      ep:resourceURI="http://www.nature.com/ejhg/journal/v9/n2/pdf/5200590a.pdf" >
      <ep:statement
        ep:propertyURI="&dc;type"
        ep:valueURI="&epentType;Copy" />

<!-- more statements describing Copy 2 -->

    </ep:description>
```

Description of
Copy 2

```
</ep:descriptionSet>
```

Figure 8. Eprints DC XML instance

the Dublin Core Metadata Element Set as a single value string, and cannot be enriched with vocabulary or syntax encoding schemes. The reason for this is that Simple DC offers a "lowest common denominator" for basic interoperability. Because of this requirement, the Scholarly Works Application needs to be expressible in oai_dc—a process known as "dumbing down" the richer metadata profile.

It is the intention of this article to look toward the future of metadata application profiles and their potential for richer functionality, and because of this the detailed aspects of the dumb-down algorithm will not be explained here.

Dumbing down is a naturally "lossy process"—loss of information content and clarity is inevitable in any mapping to Simple DC. For SWAP there were various possibilities for dumbing down, each equally valid. For example, a simple DC description for each entity could be generated, or a single description which attempts to capture as much information about all of the entities. For SWAP, the authors decided that a simple DC record for the ScholarlyWork and each Copy would be the most satisfactory solution.

SWAP thus provides a mapping from the Scholarly Works application profile to the Simple DC application profile which results in Simple DC records for both the ScholarlyWork entity and the Copies. For the ScholarlyWork our guidelines comply with those for the ePrints-UK Project, and make recommendations about where each metadata property from the full application profile should be mapped to in a simple DC record. Some information simply cannot be usefully mapped and must be left out.

The reasoning behind this mapping was that a description of the ScholarlyWork would provide a hook into the metadata about the work as a whole and traditional citation information. For the copy, a metadata record would provide an unambiguous identifier for the full text.

Engaging the Community

The hardest part of this process is engaging the community and encouraging uptake. The approach taken with this application profile is relatively complex and advocates a shift away from the view of metadata as a flat single-entity construct. This brings with it the payload of culture change and the chicken-and-egg situation: which comes first—the aggregator or the metadata to aggregate? This experience is not uncommon and many projects have taken a more simplified stance on metadata in order to achieve quicker results. Uptake at present has been slow, but awareness is growing about the need to share and expose more of the information that repositories already capture. SWAP may take a long time to impact fully on its intended community and may ultimately be superseded by newer profiles, yet its work in raising awareness of metadata issues and in encouraging a more functional approach to metadata has been and remains important.

There has been interest in the profile from the repositories community, and DCMI views it as an exemplar for their Singapore framework. Developers from EPrints, DSpace, and Fedora have been engaged in the process, and JISC continues to have an active involvement, in particular through its work with UKOLN on supporting metadata standards within the community.

The relatively new DC-Scholar Community also has a role to play in keeping up this engagement, through offering a place for open discussion and a new forum in which to undertake review and revision in the future.

Other Related Work

In the UK, JISC funds works on additional application profiles for time-based media, images, and geographic data, plus a scoping study on learning object metadata. The Images Application Profile (JISC, 2007, October 29) working group is currently developing their application profile for use with images across disciplines. They too are looking at FRBR as a potential model, and also at other relevant models. This is not the only community engaged in the drive to make better use of metadata. The Dublin Core Collections Applications profile is another DCMI exemplar in the area, and the DC-Education community is working hard on their educational profile, working alongside other initiatives to harmonize approaches to educational metadata. Resource Description and Access (RDA)—a new standard for bibliographic resources—is being developed at the moment and is looking to both FRBR and Dublin Core. MODS, the Metadata Object Description Schema from the Library of Congress, a relatively new format, also enables richer, more structured descriptions. All of these activities are looking beyond flat metadata.

Other relevant work also exists in the area of music. In particular the Music Ontology Specification draws on FRBR, FOAF, and various other ontologies and "provides main concepts and properties for describing music (i.e., artists, albums, tracks, but also performances, arrangements, etc.) on the Semantic Web" (Giasson & Raimond, 2008). This RDF-based schema contains the FRBR-esque terms MusicalWork, MusicalExpression, MusicalManifestation, and MusicalItem. Additionally, the Variations3 (Variations3, 2008) project at Indiana University has recently compared their existing metadata model against FRBR, mapping their own model and requirements to the FRBR entities, relationships, and attributes.

The Open Archives Initiative Object Reuse and Exchange project (OAI-ORE) (Open Archives Initiative, n.d.) is working to make the sharing and exchange of scholarly information more achievable, with more richness, while doing so within the existing Web architecture. There is a similarity here between the aims of the Scholarly Works Application Profile and this project, as both recognize the need to make resources available via the Web using preexisting standards and formats, for encouraging adoption

of standard mechanisms for sharing, describing, and reusing objects, and doing this in a way that allows traditional documents and semantic data to coexist and benefit each other. Developments in the semantic Web are key to the success of scholarly information exchange, and both Dublin Core and OAI-ORE are well aware of this.

CONCLUSIONS

The approach taken in the Scholarly Works Application Profile is guided by the functional requirements and the primary use case of richer, more functional, metadata. In practice, it should support navigation between different versions of a particular eprint and the more ready discovery and identification of appropriate and, hopefully, open access full texts. It represents a relatively new, largely untested, approach to metadata, taking as it does the FRBR model and applying it to scholarly works. It fits well with the current work of the DCMI, in particular by utilizing the strengths of the DCMI Abstract Model, allowing the grouping of descriptions of multiple entities into a single description set.

Concerns about its complexity are valid, but may prove unfounded since much of the metadata within the profile is already being captured by repositories; it just cannot be effectively exposed to aggregators and other systems using simple Dublin Core. Capturing the metadata from users should not need them to be exposed to additional complexity and if automated extraction techniques and more consistent creation practices develop, this process should become easier. What the profile ought to do, also, is draw attention to the need for a model-based approach, and highlight the value of knowing just what it is that you are trying to describe.

DCMI is very committed to the notion of application profiles and is developing a framework that should further support profile authors, application developers and, ultimately, the creators and consumers of metadata. As part of its Usage Board work, SWAP will be formally reviewed and, hopefully, ratified. These activities work together to promote both on-the-ground use of SWAP and better practices in developing application profiles more geared to requirements. Ultimately, though it is only through practical application of such profiles that we can really begin to see these benefits manifest.

NOTES

1. See the full use case here: http://www.ukoln.ac.uk/repositories/digirep/index/EPrint sAP_use_case_1.
2. "The FOAF (Friend of a Friend) project is a community driven effort to define an RDF vocabulary for expressing metadata about people, and their interests, relationships and activities … FOAF facilitates the creation of the Semantic Web equivalent of the archetypal personal homepage" (Dodds, 2004).
3. Key to namespaces used in the table:
    - dc: <http://purl.org/dc/elements/1.1/>
    - dcterms: <http://purl.org/dc/terms/>

- marcrel: <http://www.loc.gov/loc.terms/relators/>
- foaf: <http://xmlns.com/foaf/0.1/>
- eprint: <http://purl.org/eprint/terms/>

References

Budapest Open Access Initiative. (n.d.).Budapest Open Access Initiative. Retrieved October 27, 2008, from http://www.soros.org/openaccess/

DCMI. (2007, October 1). Scholarly Communications Community. Retrieved October 27, 2008, from http://dublincore.org/groups/scholar/

DCMI. (2008, January 14). Dublin Core Metadata Element Set, Version 1.1. DCMI Recommendation. Retrieved October 10, 2008, from http://dublincore.org/documents/dces

DCMI Usage Board. (2006). DCMI Metadata Terms, DCMI Recommendation. Retrieved October 10, 2008, from http://dublincore.org/documents/dcmi-terms

DCMI Usage Board. (2006, August). DCMI Type Vocabulary, DCMI Recommendation. Retrieved October 27, 2008, from http://dublincore.org/documents/dcmi-type-vocabulary/

DEFF, Denmark's Electronic Research Library. (2006, May 18). *DDF-MXD: Danish Research Database: Metadata Exchange Format for Documents, Version 1.1.0.* Retrieved October 27, 2008, from http://mx.forskningsdatabasen.dk/mxd/1.1.0/DDF_MXD_v1.1.0.pdf

Dodds, L. (2004, February 4). An Introduction to FOAF. *XML.com.* Retrieved October 27, 2008, from http://www.xml.com/pub/a/2004/02/04/foaf.html

Giasson, F., & Raimond, Y. (2008). Music Ontology Specification, revision 1.12. Retrieved October 27, 2008, from http://dublincore.org/architecturewiki/DescriptionSetProfile

IFLA. (1998). *Functional Requirements for Bibliographic Records,* UBCIM Publications, IFLA Section on Cataloguing. Munich: K. G. Saur. Retrieved October 27, 2008, from http://www.ifla.org/VII/s13/frbr/frbr.pdf

Intute. (n.d.). Projects. Retrieved October 27, 2008, from http://www.intute.ac.uk/projects.html

JISC. (2007, October 29). Images Application Profile. Retrieved October 27, 2008, from http://www.ukoln.ac.uk/repositories/digirep/index/Images_Application_Profile

JISC. (2008, May 14a). Eprints Terms. Retrieved October 27, 2008, from http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Terms#

JISC. (2008, May 14b). Functional Requirements. Retrieved October 27, 2008, from http://www.ukoln.ac.uk/repositories/digirep/index/Functional_Requirements

JISC. (2008, May 15). Scholarly Works application Profile. Retrieved October 27, 2008, from http://www.ukoln.ac.uk/repositories/digirep/index/Scholarly_Works_Application_Profile

JISC. (2008, October 16). Repositories Research. Retrieved October 27, 2008, from http://www.ukoln.ac.uk/repositories/digirep/

MacLeod, R. (2007). *PerX Final Report.* JISC. Retrieved October 27, 2008 from http://www.icbl.hw.ac.uk/perx/docs/PerX_FinalReport.doc

Nilsson, M. (2007a). *DCMI Description Set Profile.* Dublin Core Metadata Initiative. Retrieved October 27, 2008, from http://dublincore.org/architecturewiki/DescriptionSetProfile

Nilsson, M. (2007b). *The Singapore Framework for Dublin Core Application Profiles.* Presentation at Dublin Core Singapore, September 2007. Retrieved October 27, 2008, from http://dublincore.org/architecturewiki/SingaporeFramework?action=AttachFile&do=get&target=DSP.pdf

Nilsson, M. (2007c). Formalizing Dublin Core applicaion profiles, description set profiles and graph constraints. Retrieved December 14, 2007, from http://www.mtsr.ionio.gr/proceedings/nilsson.pdf

Open Archives Initiative. (n.d.). Object Reuse and Exchange. Retrieved October 27, 2008, from http://www.openarchives.org/ore/

Powell, A., Day, M., & Cliff, P. (2005). *Using Simple Dublin Core to Describe eprints.* Version 1.2. ePrints UK. Retrieved October 27, 2008, from http://eprints-uk.rdn.ac.uk/project/docs/simpledc-guidelines/

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2007). DCMI Abstract Model. Retrieved October 27, 2008, from http://dublincore.org/documents/abstract-model/

Prospero. (2006, July 27). Preparatory Phase. Retrieved October 27, 2008, from http://edina
.ac.uk/projects/prospero/index.html
Suber, P. (2007). Budapest Open Access Initiative: Frequently Asked Questions. Retrieved
October 27, 2008, from http://www.earlham.edu/~peters/fos/boaifaq.htm#openaccess
Variations3. (2008). Variations3. Retrieved October 27, 2008, from http://www.dlib.indiana
.edu/projects/variations3/

Julie Allinson is digital library manager at the University of York where she is managing a project to create a multimedia repository for the university's digital research resources. Julie also manages SWORD (Simple Web-service Offering Repository Deposit), a project to specify a common deposit standard. She is cochair of the DCMI Scholarly Communications Community, a member of the Object Reuse and Exchange project Liaison Group, JISC's Common Repository Interfaces Group and the DCMI Usage Board.

Previously Julie was repositories research officer at UKOLN, University of Bath and Content Coordinator for the Intute Arts and Humanities service. She has also worked as a librarian and archivist for the universities of Liverpool and Nottingham.