A metadata framework to support scholarly communication*

Thomas Krichel
Palmer School of Library and Information Science
Long Island University, 720 Northern Boulevard, Brookville, New York 11548–1300 USA
http://openlib.org/home/krichel

Simeon M. Warner
MS B285, Los Alamos National Laboratory
Los Alamos, New Mexico 87545, USA
http://t8web.lanl.gov/people/simeon/

Abstract

In this paper, we consider the design of a new metadata format to advance scholarly communication over the Internet. This format is designed to be used within the Open Archives Initiative. It is based on work by the Dublin Core Metadata Initiative and others. We present a requirements analysis and then propose a conceptual framework for the metadata. We examine metadata quality control and assess the role of the Resource Description Framework.

1 Introduction

The Open Archive Initiative (OAI) has its origin in a meeting held in Santa Fe in 1999 to enhance the interoperability of eprint archives. The Santa Fe convention was written as a result of that meeting. It required that all archives following the convention would adopt a common minimal metadata for each full-text document called the Santa Fe metadata set. Since the Santa Fe meeting the scope of the Initiative has broadened considerably. It now promotes a general interoperability framework for digital libraries through the Open Archives Protocol (OAP). It follows that the common minimum metadata set should also have a broader scope. The Initiative thus adopted the unqualified Dublin Core (DC) metadata set expressed in an XML syntax as a common metadata format, to be implemented by all archives that implement the OAP. However, the OAI continues to emphasize that the protocol allows the parallel deployment of several metadata sets.

At the Ithaca meeting of the technical committee of the Open Archives Initiative on 7 and 8 September 2000, it was decided that the initiative should support work on a metadata format to be used by the eprint archiving community. This paper is a result of that decision. It is not an official statement of the OAI; it only expresses the view of its authors. It is submitted to the Dublin Core conference to stimulate feedback and discussion.

In the remainder of this paper, we will refer to this new metadata format as the Academic Metadata Format (AMF). AMF is specified in detail in [2]. This document discusses and justifies the overall design decisions taken in that document. In Section 2 we present an informal requirements analysis for AMF. In Section 3, we review existing work as a set of tools that helps us to fulfill our requirements. In Section 4, we present a conceptual model for AMF. In Section 5, we sketch the AMF syntax. In Section 6, we look at what can be done with AMF data. Section 7 concludes.

2 Requirements analysis

The creation of a metadata format to support scholarly communication was sponsored by the Open

*The work discussed here has received financial support by the Joint Information Systems Committee of the UK Higher Education Funding Councils through its Electronic Library Programme. It has benefited from comments by Tim D. Brody, Eberhard R. Hilf, Zhuoan Jiao, Ivan V. Kurmanov, Kathryn P. Read, Sophie C. Rigny, Thomas Severiens, and Shigeo Sugimoto. Thomas Krichel is especially grateful for the hospitality of the Institute for Economic Research at Hitotsubashi University, Tokyo, and of the Department of Economics at the University of Surrey, Guildford, where much of his contribution to this paper was made. The latest version of this document is available on the web at http://openlib.org/home/krichel/kanda.html.

Archives Initiative (OAI). More specifically the interested parties are those who attended the Santa Fe meeting in October 1999. This initial constituency comprises important initiatives in the author self-archiving movement. These initiatives aim to build an infrastructure for non-tollgated access to research documents on the Internet. This infrastructure is meant to be a basic free layer of primary research documents. It may also provide the foundation for other fee-based services that build secondary documents out of those primary research documents.

Formal archives like arXiv and formal archive collections like RePEc have faced the following two problems

1. Some have seen a small number of contributing authors.
2. Some have collected metadata that is poor and inconsistent.

In the meantime, there is a flourishing culture of informal archiving in many disciplines. This is either done through web sites managed by research groups or through the personal homepages of researchers. While informal archiving is welcome as a first step, there are important problems

1. Access to the papers is essentially limited to insiders who know about the informal archive.
2. Long-term availability may be uncertain.

Two approaches can be taken to strengthen the infrastructure of the free layer.

The first approach is to convince more authors to submit to formal archives. This will only be successful if the benefits of submission of a paper outweigh the cost of submission. The key element to maximizing the benefits of submission is to raise the visibility of the submitted documents. This can be achieved if the submitted documents appear in many high quality user services. The Open Archive Initiative facilitates inclusion in many user services. However the quality of these services will crucially depend on the quality of the metadata that is supplied to them. As soon as a user service does more than full-text indexing, it will need a large stock of high-quality metadata.

The second approach is to help improve the dissemination infrastructure of the homepage publication movement. This necessitates improvement in the quality of the metadata available through such pages by supporting simple, author-supplied metadata. These data can then be fed into the Open Archives system using an intermediate provider. For this or a similar approach to work, there must be a basic level of metadata that is simple and intuitive for a non-specialist.

Both approaches are supported by the creation of metadata collections that are

- large
- simple to compose
- high-quality

We will address these requirements in turn.

The requirement to collect large collections of metadata suggests that the format should be able to address any kind of research report, be it in the past or at present. Thus the format should not be dependent on the physical support, and it should not be populated only with data on freely available documents. It appears unlikely that the format will attract a lot of document descriptions without the collaboration of the established scientific, technical and medical (STM) publishers. It will therefore be important to accommodate the wishes of that clientele. The metadata format must meet their dissemination requirements.

The requirement that the metadata should be simple to compose calls for a plain vocabulary of field names. The syntax should also be quite simple and self-explanatory. Although metadata creators may use forms to create metadata, we should not expect them to do that. These are some of the qualities frequently associated with XML.

The requirement of high-quality is more elusive. It certainly implies that there should be a significant amount of sub-tagging of information. For example, the format should allow for different author names to be unambiguously separated and there should be ways to associate institutional affiliations with each author. Metadata should also be able to include citation information and classification data.

The requirement of high quality also suggests that the metadata format should be able to express additional statements about the primary records. This will allow third parties to refine the data that is found in the primary document data. This is best seen through an example. Imagine that there is an author who has published technical documents in various archives that, for example, are managed by university libraries. There should be some mechanism by which a person can say that she has been an author of these documents and where her current homepage is. This will allow user services to directly link to an author's homepage where the user may find important, recent information about the author. For another example, if there are sets of non-peer reviewed papers available, there should be a possibility for a publisher to state that they hold peer-reviewed versions of these papers. Note that both the author and the publisher may use intermediate digital libraries to accomplish the provision of the additional information.

This naturally leads to a fourth requirement for the metadata collection, the requirement that it should be comprehensive. The descriptive effort should go beyond a catalog of works produced by academics. Instead, it should attempt to document the whole scholarly communication process. Naturally this comprises the documents produced, but it also includes data about the authors of the research document, and the institutions that support the research process. Thus the metadata that we require is process-focussed, rather than resource-focussed.

To build the metadata collection, we need a metadata format. The metadata format is the building block of the collection. The format is discussed in the next three sections.

3 Review of available work

A metadata format can be thought of as consisting of two components. These are

1. a vocabulary of terms and their significance
2. a syntax that links these terms together

As far as the vocabulary of terms is concerned there are five "standard" vocabularies that appear relevant

1. Dublin Core Metadata Element Set, [3]
2. Dublin Core Qualifiers, [5]
3. DCMI Type Vocabulary, [4]
4. OpenURL syntax, [9]
5. vCard, [6], see also [7] ,

These vocabularies offer sets of concepts that have been given an identifier in the vocabulary. For example, the Dublin Core metadata defines a concept "a date associate with an event in the life cycle of the resource" and associates the identifier `date` with it.

It is useful to use the same concepts and identifiers as the standard vocabulary for two reasons. First, some standard vocabularies—in particular those sponsored by the Dublin Core Metadata Initiative—are the fruit of a long and painful consensus-searching process among a group of metadata specialists from a variety of domain backgrounds. Second, adopting the standard concepts and identifiers will make the format easier to understand and is likely to aid the acceptance of a new format. It is not strictly necessary to use the same identifier with the same concept as the standard vocabulary does. The case for doing that has to be considered on a case-by-case basis. For example, in the world of academic documents, it is more common to use the term "author" for the person who is primarily responsible for a research paper, rather than the term "creator" as suggested by the Dublin Core Vocabulary. Since AMF is aimed at non-specialists it is better to keep the concept identifier that is more familiar to the members of the provider community. As such, the choice of "creator" appears suboptimal. If a conversion to a plain Dublin Core is required, the changeover can be made by a computer program.

As far as the syntax is concerned, we are in many ways constrained by Open Archives Initiative adopting of XML as a base format. We could use the Resource Description Framework (RDF) set out in [8] . However, we are not alone in thinking that the syntax prescribed in this document is convoluted. Also, in RDF resources have to be identified, which is problematic in the scenario that we are operating in. In a decentralized collection, the same resource may be described several times. Without a coordinated effort between metadata collectors, all we can identify in such a collection are the descriptions of resources, but not the resources themselves. This implies that RDF is not suitable for a collaborative collection with no coordination. However, RDF will still be very useful beyond AMF in the further development of AMF-based collections. We will come back to this issue later.

4 A conceptual framework

From the requirements analysis, it appears that there is a need to go beyond the bibliographic tradition of scholarly communications metadata that concentrates on the description of resources. Instead, we have come to believe that a sufficiently comprehensive description of the scholarly communication process can be produced within a conceptual framework that comprises four classes of entities, which we shall refer to as entity classes.

1. resources
2. groups of resources
3. people
4. institutions

AMF should allow the specification of properties for instances of the entity classes, and the specification of relations between them. Accordingly it seems natural to set up four record formats, one for each of the entity classes. However, on closer scrutiny of each of the types and the properties one may wish to give to its instances, it turns out that this one-to-one correspondence between entity class and record format is not the optimal way forward. To see that, it is instructive to look at each of the entity classes in turn.

4.1 Resources

Looking at the DCMI type vocabulary it appears that the resources that are most important in the academic world are of the type "text". For immediate use the creation of a text metadata type is sufficient. Some existing scholarly communication initiatives also catalog non-textual resources. For example, RePEc has an extensive catalog of computational software routines that are used mainly by econometricians. Thus other resource types will have to be supported in the future.

Academic texts are produced under various publication forms as journal article, conference papers, survey articles etc. However, the fundamental properties, such as authors, title, abstract etc appear to be the same across all these forms. We can also refer to these as the primary properties of a text. Other properties concern aspects of the interface between the text and the surrounding technical and/or social reality that make for the difference between these forms. We can call these secondary properties of a text. For example the name of the journal where a journal article is to be found is an aspect of a technical interface to the text. For another example, the name of the conference where a conference paper was presented is an aspect of a social interface to the text. The same text could also be accessible in a different form. It would have the same intellectual contents and identical fundamental properties. Therefore it makes sense to qualify all forms of texts with the same properties. This is considerably simpler than setting up a taxonomy of publication forms, and create a specific set of properties for each of these forms. Having one text type also allows the creation of metadata records for a text instance before its form of publication is known.

4.2 Collections of resources

As we have seen in the previous subsection, many of the secondary properties of textual resources are lost by aggregation when considering one textual metadata type. However in most cases, these properties in fact do not apply to one instance of textual material, but for a whole collections of texts. This is the case of a text appearing in a journal, or for a text being given as a paper at a conference. These properties are therefore better left to be described through grouping of resources. At its most general level, a group can be thought of as a statement made about several resources.

Collections may be nested. An example for such nesting is the representation of classification schemes. Classification schemes have no representation as such in AMF. A classification scheme is represented indirectly by the texts that belong to it. A classification of a text is its membership of a collection that holds all the texts that have the same classification code. Since classification schemes are usually hierarchical in nature, nested collections are used.

4.3 People and Institutions

The third fundamental entity class covered by AMF are physical persons. It is useful to be able to separate personal data out of the data on resources. This will allow a better representation of author data, with is one of the crucial components of research document metadata. Many of the practical difficulties of constructing a simple metadata format for resources come from attempts to overload the document information with personal metadata. A better information infrastructure of personal data will also allow innovative personal logging facilities for the providers of documents. More generally, it will allow for better quantitative assessment of academic research.

Finally, the last fundamental entity class are the institutions who support the creation and distribution of scholarly data. These include commercial publishers, academic institutions and scholarly societies. These are the most important actors that we need to get involved in the creation of the dataset. It is therefore important that the data collected allows these actors to self-document.

The difference between a physical person and an institution is clear enough. An institution is a group of one or more physical persons acting together. However, looking at the properties one wishes to associate with both physical persons and institutions, it turns out that these are almost the same. It is therefore doubtful if two separate metadata formats are required for both entity classes. The difficulty of creating one single entity class is essentially one of naming. There is no commonly-understood and unambiguous English language term by which to address the union of the two entity classes.

5 Nouns, adjectives and verbs

If ease of data composition were the only design goal for a metadata format, then this format should look like natural language. It is of course not useful to compose a dataset using natural language because of the difficulties that a computer would have parsing the composed data. However, we have adopted some elements of natural language grammar in the design of AMF. When setting out AMF in XML, we call the elements that represent the fundamental entity classes "nouns".

From the previous section, we have found three nouns. In AMF, they are called "text", "collection", and "p/o". An instance of the "p/o" noun can either be labeled as "person" or "organization", but in either case the record allows for identical properties, thus, as far as the overall format is concerned, they identify the same noun. As far as syntax is concerned, a instance of a noun is represented as an XML element.

We make use of two other terms from natural language grammar for AMF. Nouns elements can have two types of child elements.

A first type are called "adjectives". Some of them admit other adjectives as children, but most of them admit no children. Just as in natural language, adjectives are used to qualify a noun.

A second type are called "verbs". A verb must have at least one noun as child element. Just as in natural language, verbs are used to relate two nouns.

An example is useful here.

```
<text>
 <title>A metadata framework to support
        scholarly communication</title>
 <hasauthor>
  <person>
    <name>Thomas Krichel</name>
  </person>
 </hasauthor>
 <hasauthor>
  <person>
    <name>Simeon M. Warner</name>
  </person>
 </hasauthor>
</text>
```

This example uses the `text` and `person` nouns, the `hasauthor` verb and the adjectives `title` and `name`. The word "title" is taken from the Dublin Core metadata set and the word "name" from vCard. The record is simple to compose, and its meaning is quite evident to any English-speaking person. Therefore this syntax satisfies the requirement of simplicity.

Richness of description will not be achieved by enriching nouns with ever more adjectives. This would run counter the requirement of simplicity. Instead, we make use of verbs that relate nouns. In that sense we create a "relational" metadata format. The use of relational features is further enhanced through the use of identifiers. Each noun may be given an identifier. It may also refer to an identified noun as an alternative description. Thus the previous example—leaving out Simeon for brevity—may be written as

```
<text id="kanda">
 <title>A metadata framework to support
        scholarly communication</title>
 <hasauthor>
  <person ref="thomas_krichel"/>
 </hasauthor>
</text>
<person id="thomas_krichel">
 <name>Thomas Krichel</name>
 <isauthorof>
  <text ref="kanda"/>
 </isauthorof>
</person>
```

The interesting feature of this example is that the two records may be maintained by different persons, in different files. The record about the person may be maintained by the person herself. RePEc already uses personal records just like those in the example, that are created and maintained by registrants, see [1]. The `id` and `ref` constructs may be used in many other circumstances. For example they can be used to add records to classify identified documents in subject classification schemes. Again, this allows for labor sharing in the overall documentation process.

6 Beyond AMF

In this final section, we wish to look ahead to the time when there will be many collections of AMF data available. Some will be large, some will be small. There will be a large number of people involved in the production and management of AMF data. Some will provide data directly, others will do it via some intermediary. This intermediary may be the department they work in, the library of the institution they work in, or some other third party. Whatever solution is adopted, the quality of the harvested metadata is likely to be a major concern. Careful design of the metadata format and constructive use of computer techniques can be used to maximize the quality of the data.

6.1 Quality control

We are not aware of a commonly accepted list of forms of control that are used with the handling of metadata. We therefore introduce our own vocabulary for the control issues as we see them arising. We will use the term item to refer to either a document, a collection, a person or an institution. However, much of what is set out in the remainder of this Section can be used as a conceptual framework to discuss the control of records in other metadata formats.

6.1.1 Syntactic control

By syntactic control we mean the control of the syntax of the metadata, to ensure that it is computer parseable. The base syntax of the AMF is XML. An XML Schema is used to express additional syntax rules while also expressing some semantics. Syntactic control is the basic control form. Only those records that have a correct syntax will be submitted to the other forms of control.

6.1.2 Retrieval control

By retrieval control, we refer to the ability to retrieve the elements that the metadata points to. Since resources are the only entity class that can be retrieved, it is only applicable to instances of that class. For offline documents such control is very difficult to achieve. For online documents, it is easy to check with a URL checker as long as the document has no access restrictions and there are URLs for the full-text. AMF is very careful about distinguishing parts of the full-text from intermediate web pages. It is hoped that this distinction will be understood and adhered to. Still, it will be desirable to build a checker that can verify that a URL link goes to the full text of a paper rather than to a bibliographic page that itself links to the full text.

6.1.3 Identity control

By identity control we mean the verification that any item is described by one metadata record only. A one-to-one correspondence between item and description of the item. We propose that AMF should not concern itself with identity control. The reason for this is that identity control involves a lot of human effort. Any provision for such control thus depends the social scenario of its implementation which is not yet known.

The problem of multiple descriptions is more or less severe, depending on the item. For persons and institutions a lack of identity control is quite severe. These items exist only once in reality. Using an identification for these items makes sense only if there is a one-to-one item to record correspondence.

On the other hand, for resources and collections of resource a lack of identity control is less problematic. Many texts exists in a number of slightly different versions. In this environment, the issue of sameness becomes very difficult to decide upon. For collections it may quite often happen that different contributors maintain different collection descriptions for collections that are essentially the same. That would for example be the case for a journal that may be maintained by its publisher and by a library for example. The library could maintain data for a range of years, and the publisher for another range of years.

6.1.4 Referential control

Referential validity concerns the resolvability of identifiers referenced in a collection of AMF records. For example, if a text belongs to a certain collection, then referential validity concerns the existence of a descriptive record for the collection. Similarly if a text points to a 'person' record as one of the authors, referential validity concerns the existence of the 'person' record with the referenced identifier. In both examples, the strongest validity test would require checking not only that the identifier can be resolved to an AMF record, but also that the AMF record is itself valid (this process might proceed many levels down). Most likely, such resolution processes will involve the OAI protocols.

6.1.5 Verity control

By verity we mean that statements implied in the metadata are true. There is no way that this type of control can be realized directly within the AMF framework. It is nevertheless useful to think about how to implement verity control. Without this type of control the metadata may have no value at all. The example of the html `<meta>` tag is particularly telling. Search engine constructors have found that there are so many misleading values in `<meta>` tags that many do not use them at all.

6.1.6 Accession control

By accession control we mean the ability to control the collection of records such that records collected fit in with the aim of the collection. If the metadata collection is small, there is no problem with accession control, because the deployment and use of the collection can be held within a small community. However as the collection increases, there will be more interest from outside to use the collection to advertise inappropriate contents. Since AMF is tailored for the description of academic papers and this will hopefully limit the supply of inappropriate material.

6.2 Use of RDF

It should be clear that a pile of AMF records, scattered across the Internet, will not be sufficient to create a comprehensive description of academic reality. AMF should really be thought of as providing an entry-level collection of primary descriptive data. The quality control issues can not be solved with AMF, other tools are required.

In particular, it appears that RDF has properties that allow for a framework in which some of the quality control issues can be addressed. The primary AMF data can be trivially be converted to an RDF syntax. RDF then allows formal statements to be made about AMF content. For example, it would be possible for a rating agency to rate contents represented by AMF records. This will open avenues to implement accession control. One can imagine a variety of rating agencies who perform different ratings, from simple subject focusing to elaborate peer-review. In the same way RDF also permits verity control. RDF statements can be used to place errata in the AMF/RDF dataset. When it comes to identity control, RDF statements can be used to indicate that one metadata provider thinks that two AMF records are in fact describing the same entity.

Given the power of RDF, it likely that sets of AMF metadata will be converted into a different, RDF-based format in the future. This RDF-based format should directly dumb-down to DC and AMF. In addition, it will have more facilities to express matters relating to metadata quality control.

However, while RDF statements are an elegant technical device, they are not a substitute for cooperative human action to fix mistakes and enhance the dataset. However, the social structures that are needed for this action are not in place yet. All we can see at this point is that it appears likely that different communities will adopt different solutions.

7 Conclusions

The absence of a simple yet comprehensive metadata format for scholarly communication has limited the development of scholarly communication over the Internet. In this paper, we have described the design of a simple metadata format that could fill this gap.

Of course, the proof in the pudding is in the eating. Two leading author self archiving initiatives, arXiv and RePEc, are testing AMF as a potential future metadata standard. At the time of writing the AMF is still in a development phase.

References

[1] J. M. Barrueco Cruz, M. Klink, and T. Krichel. Personal data in a large digital library. presented at ECDL2000, available at http://openlib.org/home/krichel/papers/phoenix.html, 2000.

[2] T. D. Brody, Z. Jiao, T. Krichel, and S. M. Warner. Syntax and vocabulary of the academic metadata format. available at http://amf.openlib.org/doc/ebisu.html, 2001.

[3] DCMI. Dublin core metadata element set, version 1.1: Reference description. available at http://www.dublincore.org/documents/dces/, 1999.

[4] DCMI. Dcmi type vocabulary. available at http://www.dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/, 2000.

[5] DCMI. Dublin core qualifiers. available at http://www.dublincore.org/documents/dcmes-qualifiers/, 2000.

[6] T. Howes, M. Smith, and F. Dawson. A mime content-type for directory information. RFC 2425, available at http://www.imc.org/rfc2425, 1998.

[7] R. Ianella. Representing vcard objects in rdf/xml. available at http://www.w3.org/TR/vcard-rdf, 2001.

[8] O. Lassila and R. R. Swick. Resource description framework (rdf) model and syntax specification. W3C Recommendation, http://www.w3.org/TR/REC-rdf-syntax/, 1998.

[9] H. Van de Sompel, P. Hochstenbach, and O. Beit-Arie. Openurl syntax description. available at http://www.sfxit.com/openurl/openurl.html, 2000.