Control of records within the Academic Metadata Format
Thomas Kriche and Simeon Warner
2000–11–9

The latest version of this document is available on the web at http://openlib.org/home/krichel/meguro.html. It has benefited from comments by Ivan V. Kurmanov.

1 Introduction
In Kriche and Warner (2000), there are some fundamental design decisions for a new metadata format to describe academia. We refer to the format as the Academic Metadata Format (AMF) in the following. In this sequel paper, we address issue of control of the metadata records and the extensibility of the format. We introduce and motivate control in Section 2. In Section 3, we introduce forms of control. In Section 4 we discuss extensibility.

2 Why control metadata?
It is clear that a comprehensive collection of metadata about academic works that are freely available can only come from the providers of the documents. These form a large group of people. Some will provide data directly. Others will do it via some intermediary. This may be the department they work in, the library of the institution they work in, or some other third-party. If access to the documents themselves is limited through a toll gate, then the keepers of the toll gate may be expected to provide metadata. In principle, they should have good incentives to provide such data since it advertises the contents that they hold. In the past, the “publishers” have left the job of creating catalogs of their output to other intermediaries such as Online Computer Library Center (OCLC), for books and the Institute for Scientific Information (ISI) for articles. There are moves by the publishing industry to move towards a disintermediated future with the DOI and Cross-Ref initiatives. However, it is unlikely that we will ever reach a stage where there is one single organization that would control the supply of metadata about academic work. Therefore we are facing a situation of distributed metadata provision. Under these circumstances, we need to harvest the metadata that is provided. The harvesting may be based on the Open Archives Initiative Protocol. It may also use other protocols that act as an intermediary between the provider and the Open Archive compliant machine interface. Whatever solution is adopted, the quality of the harvested metadata is likely to be a major concern.

The lack of training of providers is a significant obstacle that any free infrastructure for metadata provision will face. A non-free metadata provision may rely on specially skilled people. A free infrastructure will not be able to rely on these skills and the quality of the data must be expected to vary. Careful design of the metadata format and constructive use of computer techniques can be used to maximize the quality of the data. An important aspect
of this process is to remove bad data before it reaches the user service. This is an instance of metadata control. More generally, metadata control should be understood as any act that lies in between the metadata at the arrival of the service provider and its presentation in a user service. We will develop forms of control in the next section.

3 Forms of 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 group, a person or an institution.

3.1 Syntactic control

By syntactic control we mean the control of the syntax of the metadata. The base syntax will be XML. XML Schema will be used to express additional syntax rules while also expressing some semantics. The data types used will be the simple types of XML Schema or complex types built from them. Syntactic control is the basic control form. Only those records that have a correct syntax will be submitted to the other forms of control.

3.2 Relational control

Relational validity concerns the validity of identifiers that the scheme proposes. For example, if a document belongs to a certain group, then relational validity is concerns the existence of the group. Similarly if a document points to a personal record as one of the authors, the relational validity will have to check that this information is valid. For both examples, the record is valid if the identifier can be resolved and the related metadata or document retrieved.

Since the Open Archive record can carry several metadata formats, it is an open question what should be done if there no record in AMF, but only some Dublin Core metadata, say. We will say that as far as we are concerned the record is not relationally valid. However that is simply a matter of internal vocabulary. An user service is free to use any metadata it sees fit.

3.3 Retrieval control

By retrieval control, we refer to the validity of the retrieval of the elements that the metadata points to. For offline documents this is very difficult to achieve. For online documents, it is easy to check with URL checker as long as the document has no access restrictions and the retrieval metadata is complete. It will be desirable to build a checker to control 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.

3.4 Verity control

By verity control we mean that statements implied in the metadata are true. One way to prepare for verity control is to make the metadata fairly precise on things that are important from on information retrieval point of view. For example the AMF should be fairly precise on access to full-text elements. This may help to spot instances where the full-text that is offered is only another intermediate web page.
There is no way that this type of control can be realized directly within the AMF framework. It is nevertheless to think on 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.

3.5 Identity control

By identity control we mean verification that any item is described by only one metadata record only. A one-to-one correspondence between item and description of the item. We propose that the metadata format should not concern itself with identity control. The reason for that is that identity control involves a lot of human effort. Any provision for such control thus depends the social scenario of its implementation. But the scenario of implementation is not yet known and can not be decreed at this early stage in public domain bibliographic database development. The problem of multiple descriptions is more or less severe, depending on the item. For persons and institutions the control a lack of identity control is quite severe. These items only exist once in reality. Using an identification for these items really only makes sense if there is a one-to-one item to record correspondence.

On the other hand, for document and group of documents a lack of identity control is less problematic. Many documents exists in a number of slightly different versions. In this environment, the issue of sameness becomes very difficult to decide upon. For groups it may quite often happen that different contributors maintain different group descriptions for groups 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 a range years, and the publisher for another range of years.

3.6 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 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. The AMF is tailored for the description of academic papers and this will hopefully limit the supply of inappropriate material.

4 Extensibility

In this section we consider the extensibility of the scheme across time and domains.

4.1 Protocol versions

To reduce interoperability problems across we suggest that the AMF should have a version number, and each record should include the appropriate protocol version number. However, it should be noted that the format may be enhanced, i.e. some new elements and element types may be added, without there be a change in the version number, because enhancements do not invalidate currently available records. It is therefore expected that the scheme be enhanced over time. Only when a change is made that invalidates currently available records there will be a change in version number.
4.2 Local extensions

Extensibility allows for flexibility to accommodate “local” groups. However, it also limits interoperability since the data contributed by the local group can no longer be fully understood when using the general schema. The syntactical validating software should be accommodating for that. It should use XML namespaces to get to a specification file, and failing that, it should use a built-in address to get a root specification file. It is up to the maintainers of the root specification file to make this specification so complete that no local additions will be needed by anyone.

Some complication of the AMF may arise in the use of identifiers. Imagine for example the use of classification data. Imagine that the archive RePEc has define a group RePEc:jel1984:a1 to contain the papers that have the classification number a1 in the Journal of Economic Literature classification scheme. Then we can write that a paper has that classification as

\[ \text{<member>oai:RePEc:jel1984:a1</member>} \]

A more intuitive way to express the same semantics would be

\[ \text{<jel1984>a1</jel1984>} \]

An aliasing mechanism may be used to allow for such aliasing. The scheme should be designed assuming that such aliasing can be provided. In this particular instance the alias concerns both the element name and its value.

References