José Manuel Barrueco Cruz and Thomas Krichel

Subject description in the academic metadata format

1. Motivation

This paper follows our long standing concern to collect abstract and indexing (a&i) and make it freely available over the Internet Through the pioneering NetEc (http://netec.mcc.ac.uk) and RePEc (http://repec.org) projects, we have been in this business for over a decade, within the subject domain of economics. Our work has expanded over time in two ways. First, our methods of work are now gaining acceptance across other discipline-based community groups. Therefore, within the data that we hold we need a general tool to represent classification data. This paper addresses this concern. Second, our collection effort has moved beyond the description of document metadata that is central to the classic a&i collection. Our work sets out to document individual authors and academic institutions, and the relationship that they have with the documents. To encode this description, we need a specific metadata format. This format is the Academic Metadata Format (AMF). Thus our paper does not address the general representation of classification data, but how to do it within the constraints of AMF. The paper is organized as follows. First we write a little more about AMF in Section 2. Section 3 introduces subject classification schemes. Section 4 is the "main dish", the discussion of subject classification within AMF. Section 5 concludes.

2. AMF

The Academic Metadata Format, AMF, see http://amf.openlib.org, is an effort to set out a metadata set that will be useful to the academic self-documentation process that we have outlined in the previous section. The model underlying AMF is distinct from conventional bibliographic format because it emphases the separate description of persons—usually authors—and institutions—usually university departments. There are four basic record types that are called "text", "person", "organization" and "collection", respectively. These are represented by XML elements called "nouns elements". There are a numerous ways in which instances of noun elements relate. The XML elements that are describing the relationships are called verbs. Verbs only admit noun elements as child elements. The specific descriptive details of a noun instance are expressed in “adjective” child elements. Adjective elements usually admit no child elements. AMF is at the time of writing, documented at http://amf.openlib.org/doc/ebisu.hml. It is not definitive. At the moment, trials are in progress to determine its usefulness.

The main technical design constraint that went into the construction of AMF came from the design goal to provide drop-in functionality with the public metadata harvesting protocol developed by the Open Archives Initiative, see http://www.openarchives.org.

Therefore AMF must be encoded in XML and it must be constrained by means of an XML Schema instance, see http://www.w3.org/XML/Schema. Not all the constraints that are implied in the human description of AMF are controllable by XML Schema in its current version 1.0. Nevertheless, when we discuss subject classification information, we will require special attention to the fact that the AMF data should be able to validate with an XML Schema file.

3. Subject classification and classification schemes

In conventional bibliographic data subject classification information is essentially viewed as a data element in the bibliographic description with a controlled vocabulary. Usually, this vocabulary is a set of cryptic terms called classification codes. One reason to have the cryptic codes is to make it easy for a machine to identify the classification data and to associate to the classification code the membership in a conceptual classification group. A set of codes is often given a hierarchical structure. However, in the publicly available documentation on classification schemes the set of operations that one can associate with these hierarchical structures is limited. It is often not clear how the hierarchy operates. For example, it is not often made explicit how a document with a specific code relates to other documents with a less specific code in a parent category. As an illustration, consider the Journal of Economic Literature (JEL) classification system, form their web site at http://www.aeaweb.org/journal/elclasjn.html. Here we have

A - General Economics and Teaching

A00 - General

A1 - General Economics

A10 - General

A11 - Role of Economics; Role of Economists

And if we follow the hyperlink on the first line, we get to another page which says

A000 - General Economics and Teaching: General

A100 - General Economics: General

A110 - Role of Economics; Role of Economists

Never mind the addition of the zeros, the two views are quite different. The first view is implicitly a hierarchical one, the second appears to be more like a flat set of codes. These two different views also appear when we try to integrate the subject information into AMF. In the following, we investigate by which means AMF records can carry classification information. As an adjacent objective, we also investigate whether AMF can be used to encode subject information itself.

4. Representation of classification data in AMF

The discussion in this section is not limited to AMF, but could more generally be regarded as looking at the representation of subject data and subject schemes in XML. We investigate different approaches. For the sake of easy referral, we label each with the name of a country[1].

In the Irish approach, there is no support for classification within AMF at all. Collectors of AMF data, however, are encouraged to use their own vocabulary and add it to the document metadata that they provide. Within every AMF noun, instances of foreign XML elements are allowed, as long as they are properly namespace qualified. Thus, is possible to introduce a namespace http://www.ams.org/msc/ with a prefix msc and then add an element <msc:msc2000>22A26<msc:msc2000> to the document metadata in AMF. Thus the AMF metadata embeds a non-AMF element that has the classification data. Unfortunately, this approach makes the maintenance of the XML data that contain the document metadata more complicated. In particular, if value checking is required within the AMF data using XML schema, the relevant schema file will have to be quoted every time the element is used. Thus, the previous example would write as

[1] Country names are not related to the contents of the proposed approach.

<msc:msc2000 xmlns = “http://www.ams.org/msc/” xsi:schemaLocation = ”http://www.ams.org/msc/ http://www.ams.org/msc2000.xsd”/>22A26<msc:msc2000>, which appears excessively bulky. A container architecture could be defined that contained several code values if they need to be quoted, but then verbosity is saved at the expense of simplicity.

In the English approach, classification data is part of a group data only. In this case, no special XML annotation is required. All the data can be provided by using standard AMF. Suppose an a group that gathers AMF data wants to use the JEL scheme. Then it is up to them to set up the classification as a series of AMF collection nouns. Such nouns can be nested. Thus we have <collection id=”auto:jel:a”><name>General Economics and Teaching</name><haspart><collection id=”auto:jel:a00”> <name> General</name> </haspart></collection> etc, where “auto” is the namespace of a fictitious group that collects AMF data. The scheme representation generates a set of handles, that are locally used within the group. The hierarchical structure of the classification scheme is fully evident in its representation. The representation of classification data can then be done within the document metadata as <ispartof><collection ref=”auto:jel:a00”/></ispartof>.

The common advantage to both approaches is that there needs to be no intervention and support work at the level of AMF itself as far as classification schemes are concerned. The Irish approach stresses a flat value space of allowable codes. It leaves the interpretation of the codes to an application that reads the AMF data to make sense on the code. The description of the classification scheme is made outside AMF data, in a XML Schema file. The English approach stresses the hierarchical structure of the classification system. It provides a much more rich description of the classification scheme within AMF data, rather than in an outside “metadata” scheme[2].

From the discussion of the first two approaches, it should become clear that there are two main issues with subject information. First, there is the issue of representation vs. control. Representation is the way the information is written down. If the way that the information is written is precise or involved, the cost of supplying subject classification data increases. On the other hand, the value of the subject information increase or at least the

[2] If we accept that metadata is data about data, then XML Schema qualifies as a metadata scheme, despite the fact that it is not frequently being referred to as such.

cost of its computerized recognition decreases. Control involves the checking of the validity of a written element of subject data. The Irish approach is strong on control, because it allows for controlling the data contents using a regular expression in the foreign XML Schema file. On the reverse it is weak on representation. The classification scheme is not properly represented by a list of codes. The second issue with the classification data within a metadata format is the external vs. internal representation of the scheme itself. It would be nice to have an internal representation. In that case the metadata format can be used to represent the classification scheme itself, rather than just the classification values. However, AMF itself has not been formulated to encode subject classification schemes and therefore it is not likely to make a good job out of it.

This idea inspires an approach that is in opposition to the English and Irish approaches. We will label it the French approach. Here, the view is what is really required is a representation of classification schemes themselves, before any inclusion of such schemes in AMF. Thus, a clearing house would be required that would take classification schemes that have been developed in the past and encode them in a specially designed XML format. Such a format would need careful design to support most, if not all the features currently available in classification schemes. Thus the issues of versioning over time as well as the hierarchical depth would need to be addressed. Each scheme could then be described in the scheme and registered with a clearing house. Authorities that collect AMF data can then use descriptive data from the clearing house to implement whatever scheme they wish to use among those that are offered. And at the level of AMF, all that is required is an adjective element <classification scheme=”jel1991”>, which would, together with the classification scheme server, enable the validation and processing of the content.

The problem with this approach is that a central registry and description of classification systems does not exist. It will not be costly to maintain it because there is a limited amount of classification schemes that are worthy of inclusion. But the initial development gives raise to considerable conceptual and technical challenges. The conceptual challenges come from the representation of the scheme in XML. There will have to be a considerable investment in studying current practice to find a suitable format that can encode most classification schemes. The technical challenge lies in the implementation of the registration and control stages.

Thus if there is any support for classification schemes within AMF, and without the support of an external party, several approaches are possible. A Dutch approach would consist in allowing a classification "adjective" XML element, with a scheme attribute as previously discussed, but leaving the value of the attribute free. The argument in favour of this low-cost approach is that classification schemes are used by communities, and they have to agree on valid identifiers for the schemes. This Dutch approach was the one that was initially set up as the AMF format was created. At the time, it has simply been a matter of convenience. An alternative approach would be a Swiss approach, where the identifiers are registered within the AMF schema file as a set of allowable values, that are allowed to appear as an attribute to the classification element. In this case the name of the scheme is controlled, but the values are not. In the Swiss and the Dutch approaches, constraining the representation of classification codes and facilitating their use is not seen as a task for AMF. But the Swiss approach realizes that the usage of the classification schemes can be greatly enhanced if there is a centrally controlled vocabulary of scheme identifier. The implementation of such a controlled vocabulary within XML Schema for AMF is quite straightforward and does not merit further discussion here.

Alternatives to the Swiss and Dutch approach start with the idea that built-in support for classification schemes should be one of the cornerstones of AMF. Thus there needs to be an internal way in which we represent and control classification scheme data within the specification of AMF itself, rather than within AMF data, as the Irish and British and Dutch approaches do. Within that approach, our options are limited to the means provided by XML Schema. Although here could be a human readable set of additional guidelines, there is not much hope that anyone will follow them through, unless they are backed up by an XML Schema file. Overall, we have to lean towards an internal Irish approach. This is to use a flat-space representation of the value in an XML Schema file. Setting up the English approach of structured description with XML Schema may be possible, but it is cumbersome and of little practical value. Therefore, the Spanish approach takes a pragmatic view where allowable values appear as value definitions that are constrained by enumeration. The AMF XML Schema instance inserts files that represent classification scheme. For illustration, one such file that starts with <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault =" qualified" attributeFormDefault="qualified" targetNamespace="http://amf.openlib.org"> <xs:simpleType name = "msc2000Element"><xs:restriction base = "xs:string"> <xs:enumeration value = "00XX"> <xs:annotation> <xs:documentation> General </xs:documentation></xs:annotation></xs:enumeration>, then there follow a lot more enumeration types, before the type declaration is closed and a list type is defined </xs:simpleType><xs:simpleType name="msc2000"><xs:list itemType = "msc2000Element" /></xs:simpleType></xs:schema>. There are obvious problems with this approach. First, this is not a hierarchical representation the scheme. Second, the scheme will have to be introduced using the xsi:type attribute, which makes the way AMF is written out dependent on XML Schema. Third, working with a list type assumes that values of classification codes do not contain spaces, because a space introduces a new list element. Despite these problems, we can see this approach as an enhancement of the capabilities of AMF as such. Therefore we are currently investigating how many schemes we can cover with this approach.

5. Conclusions

Subject classification representation using XML is still in its infancy. We find that there are good for this sorry state of affairs. The whole encoding issues is problematic. It will require more work before we can see XML-based tools to store and manipulate subject classification data in a situation where several schemes may be deployed. It is clear to us that if such XML encoding can not be implemented in a coherent way, the supply of subject data and the deployment of classification schemes will be limited. In this paper, we have made a pioneering contribution to work aimed at enhancing classification information in the area of the academic document data. We hope that we will be able to use that data to provide better representation of peer groups of academics. We also hope that the data will be used to aid information retrieval in a more conventional sense.