Dublin Core Metadata for Electronic Journals

Ann Apps and Ross MacIntyre

MIMAS, Manchester Computing, University of Manchester,
Oxford Road, Manchester, M13 9PL, UK
ann.apps@man.ac.uk, ross.macintyre@man.ac.uk

Abstract. This paper describes the design of an electronic journals application where the article header information is held as Dublin Core metadata. Current best practice in the use of Dublin Core for bibliographic data description is indicated where this differs from pragmatic decisions made at the time when the application was designed. Using this working application as a case study to explore the specification of a metadata schema to describe bibliographic data has indicated that the use of Dublin Core metadata is viable within the academic journals publishing sector, albeit with the addition of some local, domain-specific extensions.
Keywords. Dublin Core, metadata, bibliographic citation, electronic journals.

1 Introduction

Metadata is a description of an information resource, and hence can be thought of as ‘data about data’. Within the context of the World Wide Web metadata may be used for information discovery, but metadata is also important in the context of cataloguing resources. Dublin Core Metadata [1] is an emerging standard for simple resource description and for provision of interoperability between metadata systems.

The article header and abstract information used by an electronic journals application when publishing academic journal articles on the World Wide Web is effectively metadata for those articles. This paper describes an implementation of an electronic journals application, for a publisher, where the article metadata is held as Dublin Core, and some of the problems associated with specifying the metadata schema to develop the application design.

2 Dublin Core Metadata

The Dublin Core Metadata Element Set has been endorsed by the Dublin Core Metadata Initiative after its development by the Dublin Core Working Groups. These international working groups are open membership email discussion groups, with occasional ‘face-to-face’ meetings at the International Dublin Core Metadata Workshop series. The Dublin Core standard for metadata specification is primarily concerned with the semantics of the metadata rather than the syntax used for its inclusion with an information resource. It is designed for simple resource description and to provide minimum interoperability between metadata systems, with a consequent potential for cross-domain metadata interchange. Dublin Core does not attempt to meet all the metadata requirements of all sectors, where Dublin Core would be enhanced to produce domain-specific metadata schemas for richer descriptions.

Basic Dublin Core (Version 1.1) is a fifteen metadata element set as developed by the Dublin Core Metadata Initiative: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. Detailed definitions of these elements are given on the Dublin Core Web Site [1]. All Dublin Core elements are optional and repeatable.

Qualified Dublin Core allows further refinement and specification of the element content using interoperability qualifiers to increase the semantic precision of metadata. The initial set of proposed qualifiers [2] has recently (April 2000) been approved by the Dublin Core Usage Committee, after many months of discussion by the element working groups. Qualified Dublin Core includes element refinement qualifiers such as ‘Alternative Title’ as well as the possibility of specifying encoding schemes for element values. The encoding schemes, which include controlled vocabularies and formal notations, may be international standards, for instance ISO639-2 for Language, or defined within Dublin Core.

2.1 Dublin Core in Practice

There have been problems associated with using Dublin Core in practical implementations. Firstly the basic element set may be insufficient to fully describe the resource within a particular domain. Secondly the timescale for Dublin Core metadata definition has meant that it has been necessary to anticipate the standard definitions before they have been fixed, sometimes incorrectly.

Basic Dublin Core Element Set. The basic Dublin Core element set does not attempt to meet all requirements of all resource domains. It is acknowledged that application designers will meet local functional requirements by the inclusion of additional elements and qualifiers within a local namespace. But it is expected that if local qualifiers are demonstrated to be of broader application they will attract wider deployment. Application designers will be encouraged to register their local qualifiers with the Dublin Core Metadata Registry [3] to leverage their usage, and hence interoperability, more widely.

Dublin Core Definition Timescale. Because of the nature of the method of defining Dublin Core through open-membership discussion groups, the length of time before versions have been baselined has been considerable. This is particularly the case with Qualified Dublin Core, whose first skeleton version has only recently been approved. This has caused a problem for application designers who have needed to produce an 'up-and-running' implementation quickly. Until now, using Dublin Core has been rather like trying to 'hit a moving target'. Inevitably some of the decisions made in these application designs have since been deprecated, but it is not always possible to change working applications easily.

3 An Electronic Journals Application

The Manchester University Press (MUP) [4] electronic journals application hosted by MIMAS [5], at the University of Manchester, UK, holds the article metadata as enhanced Dublin Core using an XML [6] syntax. Journal article metadata is supplied to MIMAS in SGML format, using the Simplified SGML for Serial Headers (SSSH) [7] Document Type Definition (DTD) for the majority of the journals but a typesetter's proprietary DTD for one journal. During the data handling process this SGML metadata is transformed into the required 'Dublin Core in XML' format, and XML 'Table of Contents' files are generated, using specific OmniMark [8] programs written by MIMAS. The full article files are supplied as PDF and are provided in this format to the end-user. When an end-user requests viewing of the article metadata, including its abstract, the 'Dublin Core in XML' is transformed into HTML 'on-the-fly' using another OmniMark program written by MIMAS.

It was decided to hold the article header information in Dublin Core because it was recognised that this is essentially metadata for the articles. Thus exploration of using Dublin Core for bibliographic applications seemed appropriate. A common format for holding article metadata for all the journals was necessary because they are delivered to MIMAS by Manchester University Press in SGML using two different DTDs.

In general, the mapping between the article metadata elements and the Dublin Core Metadata Element Set was obvious. But decisions were needed how to capture: the journal article bibliographic citation information such as journal title, volume and issue numbers; the parts of the name of a Creator; the affiliation of a Creator; the location and size of the corresponding full article PDF file; the specific document Type; and the language of those elements which could be supplied in multiple languages. A pragmatic approach was taken to resolving these issues because timescale constraints obviated waiting for Dublin Core endorsed definitions. Details of how these were coded and of current best practice are described below.

Although all Dublin Core elements are optional, to implement a viable electronic journals application some elements must be mandatory. The application requires every article to have Title, Publisher, Date, Type, Format, Identifier, Source, Language, Relation, and Rights. This requirement is imposed by the XML DTD.

The design of this application was developed from work done at MIMAS on the *Nature* Digital Archive project [9], [10], and from previous work on the SuperJournal project [11]. Examples within the paper are from a sample article header published in the January 2000 issue of the `International Journal of Electrical Engineering Education' [12].

3.1 Simple Element Mappings

Some article metadata elements mapped obviously onto Dublin Core elements, as shown in Table 1.

Table 1. Mappings to Dublin Core elements.

Article Metadata	Dublin Core Element
Title	Title
Author	Creator
Keyword	Subject
Abstract	Description
Publisher name	Publisher
Cover date	Date
Article language	Language
Full article format	Format
Copyright	Rights

The Dublin Core elements Contributor and Coverage are not used. Multiple instances are allowed for Creator, and Subject, one for each keyword. Multiple instances are also provided for Title and Description to implement multiple language metadata.

4 Journal Article Citation Metadata

A major problem in using Dublin Core for journal article metadata is capturing the bibliographic citation information for an article within a journal issue. Since this electronic journals application was designed, this problem has been addressed by a Dublin Core Working Group who were specifically tasked with addressing this issue and recommending a solution. The decision made in this application has since been deprecated by the DC-Citation Working Group's [13] consensus, but the experience of designing the application was input to the working group of which MIMAS has active membership. Although a recommendation has been made by the DC-Citation Working Group it has not yet been endorsed by the Dublin Core Metadata Initiative and the requisite element qualifiers and encoding schemes have not yet been recommended by the relevant individual element working groups.

To capture a journal article citation for use within a bibliographic reference, the minimum requirement is the Journal Title, the Volume number and the Start Page number of the article within the printed journal. Generally bibliographic references also include the publication year. For article metadata within an electronic journals application it is also necessary to capture the number of the Issue containing the article. It is probably sensible to capture the ISSN number of the journal and the End Page of the article within the printed journal. There are some defined schemes for specifying this information, in particular the Serial Item and Contribution Identifier (SICI) [14].

4.1 Journal Article Citation within MUP E-Journals

Within the Manchester University Press E-Journals application, the journal article citation information is held within Source, using structured values within the local namespace, ie. according to an 'internal' scheme, 'MUP.JNLCIT'. These structured values capture separately the Journal Title (JTL), the Volume (VID), the Issue number (IID), the Start Page (PPF) and the End Page (PPL). Holding this information as a structured value simplifies subsequent parsing and processing by other applications, such as the program which displays the article metadata to the end-user. For example:

```
<MUP.Source SCHEME="MUP.JNLCIT">
    <MUP.JTL>International Journal of Electrical Engineering
             Education</MUP.JTL>
    <MUP.VID>37</MUP.VID>
    <MUP.IID>1</MUP.IID>
    <MUP.PPF>26</MUP.PPF>
    <MUP.PPL>37</MUP.PPL>
</MUP.Source>
```

Further citation information is held within Identifier and within Relation. Identifier holds the SICI (Version 1) for the journal issue containing the article, though ideally this SICI should be to article level. A second instance of Identifier contains an internal identifier used for subsequent processing, made up of the local abbreviation for the journal ('IJEEE' in the example below), the Volume number, the Issue number, and the Article number within the issue. An instance of Relation details the ISSN of the journal. For example:

```
<DC.Identifier SCHEME="SICI">0020-7209(20000101)37:1</DC.Identifier>
<DC.Identifier SCHEME="MUP.ARTID">IJEEEV37I1A3</DC.Identifier >
<DC.Relation SCHEME="ISSN"
    RELATION="IsPartOf">0020-7209</DC.Relation>
```

Note that within all examples, the local namespace is 'MUP'.

4.2 DC-Citation Recommendation

The final recommendation of the DC-Citation Working Group was to hold the journal article bibliographic citation information, including page range, within Identifier, using either a structured value or a prescribed syntax within the text string content of the Identifier value. Using structured values would make subsequent parsing and processing of the Identifier simpler, but these have not yet been ratified by the Dublin Core Metadata Initiative. If agreed by the Identifier Working Group, this use of Identifier would be qualified with a 'Citation' qualifier. Because all Dublin Core elements are repeatable, it would be possible to have additional instances of Identifier containing the journal article identifier encoded according to other schemes such as SICI or Digital Object Identifier (DOI) [15]. Relation will hold the citation information for the next level above. So for an article the Relation 'IsPartOf' would indicate the bibliographic citation for the journal issue.

Using this recommendation and a structured value 'DCCITE', the above example could be encoded as:

```
<DC.Identifier.Citation SCHEME="DCCITE">
    <JournalTitleFull>International Journal of Electrical Engineering
                      Education</JournalTitleFull>
    <JournalTitleAbbreviated>IJEEE</JournalTitleAbbreviated>
    <JournalChronology>January 2000</JournalChronology>
    <JournalVolume>37</JournalVolume>
    <JournalIssueNumber>1</JournalIssueNumber>
    <JournalPages>26-37</JournalPages>
</DC.Identifier.Citation>
```

'JournalChronology' is included in this recommended journal article citation metadata to overcome issues surrounding the use and semantics of DC.Date to indicate the journal cover date, which is essentially an artificial date but necessary for journal cataloguing and discovery. An additional advantage to including JournalChronology within the citation is the ability to capture dates such as 'Spring 2000' which appear on some journal issues but are difficult to encode.

If a structured value were not used this example would become:

```
<DC.Identifier.Citation SCHEME="DCCITE">
    JournalTitleFull: International Journal of Electrical Engineering
                      Education;
    JournalTitleAbbreviated: IJEEE;
    JournalChronology: January 2000;
    JournalVolume: 37;
    JournalIssueNumber: 1;
    JournalPages: 26-37
</DC.Identifier.Citation>
```

An article could additionally be identified using another encoding scheme such as SICI or DOI. Its containing issue could be indicated using DC.Relation 'IsPartOf'. For example:

```
<DC.Identifier SCHEME="DOI">10.1060/IJEEE.2000.003</DC.Identifier>
<DC.Relation.IsPartOf SCHEME="SICI">
    0020-7209(20000101)37:1</DC.Relation.IsPartOf>
```

There was much discussion within the DC-Citation Working Group before this recommendation was made following the 7th International Dublin Core Workshop in Frankfurt, Germany in October 1999. As well as the above coding method in the MUP E-Journals application using Source, it was suggested that the journal article citation information should be in Relation using the 'IsPartOf' qualifier.

Although these recommendations have been made by the DC-Citation Working Group, the indicated qualifiers and encoding schemes must be recommended by the relevant Dublin Core Element Working Groups and ratified by the Dublin Core Metadata Initiative. Currently the only approved encoding scheme for Identifier is 'URI', which itself is not yet a clearly defined standard, and it has no approved element qualifiers.

5 Author Name and Affiliation

Although 'Author' maps obviously onto DC.Creator, there are currently no approved qualifiers or encoding schemes for DC.Creator. The element value is simply a Creator's name as a free-text string. Within an electronic journals application it is desirable to split the author's name into constituent parts. For instance, indexing on authors' surnames could provide useful functionality to end-users. Within the MUP E-Journals application, the author names are encoded using a local structured value, 'MUP.AU' which captures separately an author's family name, first names and an optional suffix.

A further problem is to capture an author's affiliation. For an article published in an academic journal this affiliation indicates the author's institution at the time when the article was published, which is not necessarily the author's current address. It would have been possible to extend the above creator structured value, 'MUP.AU', to additionally capture this affiliation. But if more than one author has the same affiliation, this would be repeated for each author. Some authors have more than one affiliation if they are associated with more than one institution. As well as not wishing to repeat addresses within the information displayed to the end-user, it seemed better to replicate the information supplied in the original typeset SGML where an address is defined once with pointers to it from the author names. Thus a further local scheme, 'MUP.AFFS' is defined to capture addresses including an identifier attribute. In addition, Creator has a local attribute pointing to the relevant address identifiers, which may be a comma-separated list.

Using as an example the same article as in the previous examples, the author details are encoded as:

```
<MUP.Creator SCHEME="MUP.AU" IDS="AFF1">
    <MUP.FNMS>Bill</MUP.FNMS><MUP.SNM>Olivier</MUP.SNM>
</MUP.Creator>
<MUP.Creator SCHEME="MUP.AU" IDS="AFF1">
    <MUP.FNMS>Oleg</MUP.FNMS><MUP.SNM>Liber</MUP.SNM>
</MUP.Creator>
<MUP.Creator SCHEME="MUP.AU" IDS="AFF2">
    <MUP.FNMS>Paul</MUP.FNMS><MUP.SNM>Lefrere</MUP.SNM>
</MUP.Creator>
<MUP.Creator SCHEME="MUP.AFFS">
    <MUP.AFF ID="AFF1">University of Wales</MUP.AFF>
    <MUP.AFF ID="AFF2">The Open University</MUP.AFF>
</MUP.Creator>
```

This specification scheme for author information does not allow for any grouping of authors. Author grouping is used in article header metadata by some publishers to indicate those authors whose affiliation is the same. It may also be used to indicate significant groupings of those contributing to an article. The journals included in this application do not make use of author grouping, although the SGML DTDs used for data supply would allow it, so this was not a requirement within the described application. But author grouping could have been employed as an alternative solution to the problem of avoiding address repetition. Dublin Core does not include any notion of grouping of the elements,

which are all optional and repeatable, but it would be possible to impose grouping by the syntax used in an actual implementation.

6 Other Issues

6.1 Full Article Format and File Size

It is necessary to capture in some way the format of the full text article to which the article metadata refers. Although within this application all articles are in PDF format, other electronic journal applications may offer full text articles in a choice of formats, such as HTML in addition to PDF. The full article format is captured within DC.Format. It is additionally required to hold the size of the PDF file so that an end-user could be informed for download. The internal path and name of the file is useful to the application, though the path is possibly deducible using internal naming conventions. Full article file path and size are encoded using a local qualifier to DC.Relation, 'IsAbstractOf', with the PDF file size as an additional attribute. For example:

```
<DC.Format SCHEME="IMT">application/pdf</DC.Format>
<MUP.Relation SCHEME= "MUP.PDF" RELATION="IsAbstractOf"
    PDFSIZE= "99">IJEEE/V37I1/370026.pdf</MUP.Relation>
```

Using the now approved Dublin Core Format element qualifiers, and Identifier with a local encoding scheme, this would be better coded as:

```
<DC.Format.Medium SCHEME="IMT">application/pdf</DC.Format.Medium>
<DC.Format.Extent>99</DC.Format.Extent>
<MUP.Identifier SCHEME="MUP.PDF">
    IJEEE/V37I1/370026.pdf</MUP.Identifier>
```

6.2 Journal Article Type

The currently defined encoding scheme for DC.Type, the DCT1 Type Vocabulary, does not provide any means of indicating that the metadata is for a journal article. The 'DCT1' encoding scheme is at a higher level of abstraction. At present, its approved terms are: Interactive Resource; Dataset; Event; Image; Sound; Service; Software; Collection; Text. Using this scheme the article would simply be indicated as text:

```
<DC.Type SCHEME="DCT1">Text</DC.Type>
```

In addition to research articles an electronic journals application will contain 'Table of Contents' files, and could contain other types of document. Thus a local encoding scheme, 'MUP.TYPE' is used:

```
<MUP.Type SCHEME="MUP.TYPE">Research Article</MUP.Type>
```

A list of document types necessary for an electronic journals application may include: Announcement; Book Review; Corrigendum; Critique; Editorial; Erratum; Discussion Forum; Invited Commentary; Letter to the Editor; Obituary; Personal View; Research Note; Research Article; Review Article; Short Communication; Special Report; Table of Contents; Technical Report. Possibly in the future a list of types suitable for the academic journals publishing sector will be registered as a domain-specific controlled vocabulary with Dublin Core.

6.3 Language

Although all of the articles within this electronic journals application are written in English, one of the Manchester University Press journals has article titles and abstracts additionally in French, German and Spanish. These have to be captured within the article metadata, and are displayed to the end-user when viewing article information. Thus Title and Description have a language attribute and may have multiple instances. For example:

```
<DC.Title LANGUAGE="en">Specifications and standards for learning
    technologies: the IMS project</DC.Title>
```

```
<DC.Title LANGUAGE="fr">Sp&eacute;cifications et normes pour
    technologies de formation: le projet IMS</DC.Title>
<DC.Title LANGUAGE="de">Spezifikationen und Normen f&uuml;r
    Lerntechnologien: das IMS Projekt</DC.Title>
<DC.Title LANGUAGE="es">Especificaciones y est&aacute;ndares para las
    tecnolog&iacute;as de la ense&ntilde;anza: el proyecto
    IMS</DC.Title>

<DC.Description LANGUAGE="en">
    <P>The Instructional Management System (IMS) project is
        developing a set of specifications for learning
        technologies…</P>
</DC.Description>
<DC.Description LANGUAGE="fr">
    <P>Le projet IMS d&eacute;veloppe un ensemble de
        sp&eacute;cifications pour les technologies de formation…</P>
</DC.Description>
<DC.Description LANGUAGE="de">
    <P>Das IMS Projekt entwickelt einen Staz Spezifikationen f&uuml;r
    Lerntechnologien…</P>
</DC.Description>
<DC.Description LANGUAGE="es">
    <P>El proyecto IMS est&aacute; desarrollando un conjunto de
    especificaciones para las tecnolog&iacute;as …</P>
</DC.Description>
```

Non-keyboard characters are held as SGML character entities within the article metadata, for example 'è' is encoded as '`&eacute;`' and 'ü' as '`&uuml;`'. Most of the characters used in European languages are displayed correctly by Web browsers when included in HTML in this SGML character entity encoding. Other characters are converted to a displayable encoding when converted to HTML for end-user display of an article's metadata by using OmniMark, which is an SGML-aware language. For example, '£' is translated to '`&#163;`' and 'α' is translated to '`<FONT FACE="symbol">&#097;</FONT>`'.

7 Conclusion

The experience of designing and implementing this electronic journals application has provided a case study to explore the viability of using Dublin Core for journal article metadata. It has indicated that it is possible to use Dublin Core within the academic journals publishing sector with the addition of some local encoding schemes, in particular structured values. It was not necessary to include new local metadata elements. Any future new design of an electronic journals application would take into account more recent decisions about Qualified Dublin Core, in particular the recommendations made by the Dublin Core Citation Working Group.

There are initiatives to progress Dublin Core into an international standard for metadata for simple resource description and minimum interoperability. MIMAS is a member of the European CEN/ISSS Workshop on Metadata for Multimedia Information – Dublin Core (MMI-DC) [16] which seeks to ratify the use of Dublin Core metadata as a standard within Europe as well as providing guidelines on its use and an Observatory on European projects which are using Dublin Core.

Using a standard metadata description such as Dublin Core for an application will assist in future interoperability with other applications. An application may wish to provide cross-domain searching, for instance across both research datasets and the corresponding research literature. Or it may wish to display only a subset of the information, for instance on a mobile phone screen.

Up until now, the main problem of using Dublin Core metadata to implement a production quality application has been the lack of a stable definition. But Basic Dublin Core Version 1.1 is now fixed, and the first version of Qualified Dublin Core has recently been announced to Dublin Core Working Group members. The Dublin Core Metadata Initiative intend to set up a Dublin Core Metadata Registry for the registration of local, domain-specific elements, qualifiers and encoding schemes in addition to

those endorsed by Dublin Core. This will allow sharing of interoperable metadata items and leverage the approval of further qualifiers and encoding schemes by the Dublin Core Metadata Initiative for general use. So it appears that Dublin Core metadata is now mature enough, and acceptable as a standard, for its use to be recommended wherever metadata is required.

References

1. Dublin Core Metadata web site. http://purl.org/dc/
2. Weibel, S.: Approval of Initial Dublin Core Interoperability Qualifiers. Email message to DC-General Working Group (2000) http://www.mailbase.ac.uk/lists/dc-general/2000-04/0010.html
3. Dublin Core Metadata Registry Working Group. http://www.mailbase.ac.uk/lists/dc-registry/
4. Manchester University Press web site. http://www.man.ac.uk/mup/
5. Electronic Publishing at MIMAS web site. http://epub.mimas.ac.uk/
6. XML. http://www.w3.org/XML
7. SSSH, Simplified SGML for Serial Headers, DTD. http://www.oasis-open.org/cover/gen-apps.html#sssh
8. OmniMark Technologies web site. http://www.omnimark.com
9. MacIntyre, R., Tanner, S.: Nature - a Prototype Digital Archive. International Journal on Digital Libraries (2000), Springer-Verlag, (Scheduled for 2000(2))
10. Apps, A., MacIntyre, R.: Metadata for the Nature Digital Archive. http://epub.mimas.ac.uk/natpaper.html
11. Apps, A.: SuperJournal Metadata Specification. SuperJournal Project Report, http://www.superjournal.ac.uk/sj/sjmc141.htm
12. Olivier, B., Liber, O., Lefrere, P.: Specifications and standards for learning technologies: the IMS Project. International Journal of Electrical Engineering Education 37(1) (2000) 26-37 doi://10.1060/IJEEE.2000.003 (http://dx.doi.org/10.1060/IJEEE.2000.003)
13. Dublin Core Bibliographic Citations and Versions Working Group. http://purl.org/dc/groups/citation.htm
14. SICI, Serial Item and Contribution Identifier. http://sunsite.berkeley.edu/SICI/
15. DOI, Digital Object Identifier. http://www.doi.org
16. CEN/ISSS Workshop on Metadata for Multimedia Information (MMI-DC). http://www.cenorm.be/isss/Workshop/MMI-DC