24 Hour Museum Metasearch project: schemas

System Simulation Ltd

Version 1.0

24 Hour Museum Metasearch project: schemas

Introduction

The Metasearch project will use the Open Archives Initiative (OAI) protocol to harvest data from participating institutions. This document describes the schemas these institutions should use, when returning data in response to OAI requests.

It has been agreed that the Metasearch project requires searchable data which is organised in a manner that diverges from standard Dublin Core Simple metadata standards. The term "DC Culture" has been coined to describe this form of metadata. In addition, the project requires useful summaries about each object.

Overview

We have decided to separate these requirements, and design two XML schemas for use with the OAI protocol: one for metadata, and a second one for object summaries. This means that the OAI harvester will have to ask for two sets of metadata rather than one. This should not be a complication for users, since the OAI protocol is designed to support exactly this situation.

We feel that this approach has two advantages.

First, it clearly separates the "searchable" metadata from the "descriptive" summary. This allows respondents to provide, for example, a description for online display which is different from (e.g. shorter than) the searchable description they provide in their DC Culture metadata.

Second, we feel it will prove more flexible in the longer term. The requirements for summaries might change in future, or more detail about each object might need to be delivered via OAI. If this happened, the "summary" schema could be amended, or an additional one introduced, without any affect on the metadata already delivered, or on the indexes which will have been set up on the basis of this metadata.

OAI schemas

The two schemas defined for this project are:

- DC Culture: an adaptation of Dublin Core Simple for cultural metadata
- Metasearch summary: a set of summary information suitable for display on the 24 Hour Museum website

DC Culture

The intention with DC Culture is to work, as far as possible, within the framework already established by the Dublin Core Metadata Initiative (DCMI)[1]. This requires us to "invent" only those aspects which are essential for our purposes, and to re-use DC Simple concepts wherever possible.

High-level access points

We assert that the culture section requires the CIMI/Aquarelle High-Level Access Points "who", "what", "where" and "when". The DC Culture schema adopts these concepts and in the rest of the document we will refer to them as "DC Culture High-Level Elements".

DC simple access points

Our analysis of the existing DC Simple access points suggests that all of them except Subject, Coverage and Date are specialised instances of one of these high-level access points, as follows:

DC Culture High-Level element	**corresponding DC Simple element(s)**
who	creator, publisher, contributor, rights, subject
what	title, subject, description, type, format, identifier, source, language, relation, subject
where	subject, coverage
when	subject, coverage, date

This means that Coverage and Subject both map to more than one DC Culture high-level element. Every item of data which would be mapped to 'subject' in a DC Simple setting can be mapped instead to one of the DC Culture high-level access points. In addition, Date has a confusing relationship to our "when" concept (i.e. it is effectively the same thing). As a result, we have decided to leave these three concepts out of our metadata framework, so as to avoid semantic confusion. This means that the table above can be simplified to this form:

[1] http://www.dublincore.org/

DC Culture high-level element	corresponding DC Simple element(s)
who	creator publisher contributor rights
what	title description type format identifier source language relation
where	-
when	-

Data mapping guidelines

The advice we give to data providers is to use a [more specific] DC concept from the table above where it is appropriate, and to use the generic DC Culture high-level access point where there is no directly equivalent DC concept.

Data providers should adjust their mappings as follows when data would have been mapped to one of the three DC Simple access points which we are leaving out:

omitted DC Simple element	DC Culture element(s) to use instead
Subject	who, what, when, where as appropriate
Coverage	when or where as appropriate
Date	when

Data recording guidelines

In following our strategy of sticking to Dublin Core guidelines, we are faced with something of a quandary when it comes to advising on the form in which data should be presented. The DC Simple schema makes no attempt to define formal syntax rules for any element type: they are all just "strings". We have continued this practice when defining the format of the DC Culture high-level access points.

Sample entry

This entry validates against the DC Culture schema:

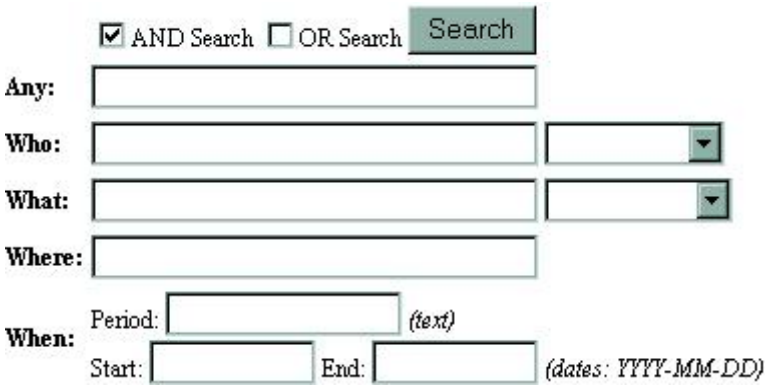
```
<?xml version="1.0"?>

<dc-culture
 xmlns="URI:DC.Culture/XMLSchema/1.0"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
<who xml:lang="en">Light, Richard</who>
<what xml:lang="en">mug</what>
<dc:title xml:lang="en">Siena mug</dc:title>
<where xml:lang="en">U.K.; West Sussex; Burgess Hill</where>
<when>2003-03-11</when>
</dc-culture>
```

Searching on DC Culture metadata

The idea behind our design is that the same set of metadata can be used to support searching at any or all of three levels:

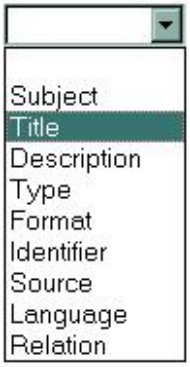
- generic (all metadata, i.e. equivalent to a free text search)
- high-level (a specified DC Culture high-level access point and the corresponding DC Simple access points)
- specific (a specified DC Simple access point or a DC Culture high-level access point without corresponding DC Simple access points)

This diagram shows how a web user interface might offer these three levels of searching as user-selectable options:



Combo box for Who:

Combo box for What:



Let's take the case of a Who search:

1. 'Combo=blank' retrieves metadata in <who>, <creator>, <publisher>, <contributor> & <rights>

2. 'Combo=Subject' retrieves metadata in <who> only

3. 'Combo=Creator' retrieves metadata in <creator> only

Formal identification of DC Culture schema

The namespace URI of this schema is currently:

"http://www.minervaeurope.org/DC.Culture/XMLSchema/1.0"

Metasearch summary

This schema is rather more ad hoc than DC Culture. For a start, it is intended for use within one specific project, whereas we can see DC Culture being applied more widely. As a result, we can adjust this schema with less concern about potential knock-on effects for other users.

Summary concepts

The schema currently defines the following concepts:

- title: a unique title, or descriptive name for the item, suitable as a heading
- description: a free-text description of the item
- link: a link to further details, e.g. a full record on the museum's own web site
- image: a link to an image
- credit-line: "An acknowledgement of donations of funds or objects on a display label" [SPECTRUM definition]

"link" and "image" both have two subelements:

- url: the URL of the resource being pointed to
- caption: a descriptive caption which can act as the text on which the user clicks to activate the link

Both "link" and "image" are optional and repeatable, allowing flexibility in the number of type of links that are offered by contributors.

Formal identification of 24 Hour Metasearch Summary schema

The namespace URI of this schema is currently:

"http://www.24hourmuseum.org.uk/Metasearch/Summary/XMLSchema/1.0"

Resolution of issues in the context of the 24HourMuseum Metasearch project

Schema URIs

Each schema needs to have an authoritative, unchanging, URI. This should be based on a domain name or other prefix which is meaningful in the cultural arena. (Note that once a URI is agreed, we are stuck with it. An example of this occurs in DC Simple, where the web site and the DC schemas have moved to dublincore.org, but the URI for the DC elements is still:

xmlns=http://purl.org/dc/elements/1.1/

which is where they used to be hosted.)

The actual namespace and the site where the schemas are hosted do not need to have the same prefix, but there are clear advantages to their doing so.

- Resolved to use www.minervaeurope.org and www.24hourmuseum.org.uk respectively.

Data recording conventions

As noted above, the DCMI is rather supine about defining syntax and vocabulary conventions for the DC Simple access points. We need to mandate a format for data which is precise enough to allow effective cross-searching of the data that is returned.

One good starting-point would be simply to adopt CIMI's advice on the use of Dublin Core:

http://www.cimi.org/old_site/documents/meta_bestprac_v1_1_210400.pdf

We have also thought long and hard about how to record 'where' and 'when' information. In theory, this should be easy, since DCMI have issued specific guidelines on recording co-ordinate and date information:

http://dublincore.org/documents/2000/07/28/dcmi-point/
http://dublincore.org/documents/2000/07/28/dcmi-period/

However, we share CIMI's reservations about the impact of DCSV encoding[2] on interoperability, and without DCSV we have problems in recording a range of dates, or a co-ordinate system, within a single text field. In addition, the DCMI-Point proposal takes no account of the situation where a place is defined by a sequence of place keywords (e.g. "U.K.; West Sussex; Burgess Hill"). The XML-coding proposals in the DCMI documents referred to above are purely suggestive (as well as being nearly three years old), so we would hesitate to adopt them.

- Resolved to follow the DCMI recommendations in the first instance.

[2] http://dublincore.org/documents/dcmi-dcsv/
[3] Guide to Best Practice, Dublin Core Version 1.1 (http://www.cimi.org/old_site/documents/meta_bestprac_v1_1_210400.pdf and http://www.cimi.org/public_docs/TBK_DCQ_Commentary_v5_0700.pdf)

Omission of 'subject'

We are happy with our proposal to drop the 'subject' access point from the DC Culture schema. This is because we feel confident that every item of data which would be mapped to 'subject' in a DC Simple setting can be mapped instead to one of the DC Culture high-level access points. However, this assertion needs to be case-tested by the participants. If it transpires that 'subject' is actually required, it may be necessary to reinstate it.

- Resolved as proposed.

'who' concepts

Although we have stopped at defining a single access point called Who, it should be noted that there is wide agreement within the cultural sector that more precise concepts could be attempted. There is a clear distinction between an individual and groupings of people, and between different types of grouping. SPECTRUM has Person and Organisation, but also defines the concept People (as in "a people"). The CRM has Person and Group (with sub-concept Legal Body).

We feel that analysing Who into sub-concepts would be counter to the principle of establishing high-level access points. There is also the potential for problems where contributors might have data in one field which could be either, say, Person or Organisation. This would mean that automatic mapping in response to OAI requests was not possible.

- Resolved as proposed. This is perhaps an issue to re-visit once the basic concept of DC Culture has been tested, and we are looking at how it might be refined.

Summary structure

We are very conscious that the summary structure is untested, and expect that it will need to be revised before it can be put into active service. For example, the "description" element currently has no sub-structure and is not repeatable.

- We suggest that the next step should be to ask each contributor to work out what summary information they want to provide, and then to see how that information fits into the current framework. We are happy to amend the summary structure as required: this won't be a big job now that the basic framework has been established.

Richard Light, Damien Dudouit
11th March 2003

Appendix A. Comparison of DC Simple with cultural frameworks

This section is taken from a discussion paper which was prepared by Richard Light as the first stage of this work. It outlines the arguments which led to the design we are now proposing.

Relationship between DC Culture and DC Simple

I'm not sure that I can see any useful relationship between Dublin Core and the proposed DC Culture elements.

For a start, David's mapping between the two frameworks (reproduced below) demonstrates that there isn't a clear "specialisation" of generic DC concepts as we move into this specialised domain. You would expect the concepts in a domain specialisation to be more specific than those in DC Simple, but DC Culture's are not – by design. Part of the motivation for the current DC Culture design is to start as generically as possible, i.e. to be even more high-level than DC itself.

DC Culture element	corresponding DC Simple element(s)
who	Creator, Publisher, Contributor, Rights, Coverage
what	Title, Subject, Description, Type, Format, Identifier, Source, Language, Relation, Coverage
where	Coverage
when	Coverage, Date

Conversely, one DC Culture element ("where") is actually more precise in its meaning than the "corresponding" DC Simple element. "Coverage" maps to all four DC Culture elements.

Therefore, in my view, any attempt to formalise these correspondences would result in a semantic jumble. CIMI's experience as recounted in its Guide to Best Practice[3] suggests that the syntax of qualified DC can lead to a loss of interoperability. My analysis of the situation suggests that you couldn't sensibly apply DC qualification to the DC Culture elements, even if you wanted to!

Another problem with DC Simple is that it provides no proper support for "grouped" information. Examples are dates, and the use of co-ordinates to describe places. The DCMI Point Encoding Scheme[4] can be adopted for the latter.

[4] DCMI Point Encoding Scheme (http://dublincore.org/documents/2000/07/28/dcmi-point/)

The CIDOC Conceptual Reference Model (CRM)

The CIDOC CRM provides the following concepts which match the DC Culture elements:

DC Culture element	CRM concept (and sub-concepts)
who	Actor Person Group Legal body
what	Physical Object Biological object [Person] Man-made object Iconographic object
where	Place Physical feature Man-made feature Site
when	Time span Temporal entity Condition state Period Event Beginning of existence Production Conceptual creation Formation Birth End of existence Destruction Dissolution Death Activity [various collection management concepts come here]

Note that there isn't a clean separation in this framework, any more than there is in DC Simple. For example, Person appears under "what" as well as under "who". However, this more abstract treatment of cultural concepts may be useful when deciding how to assign existing data to the DC Culture high-level access points.

This comparison ignores any associations between these entities: a point which was made in the CIMI study of DC, where roles played by people were cited as an example of the type of qualifier which DC did not easily allow. As a general issue, we need to decide how "rich" to make the metadata which we are exporting for this project.