Gordon Dunsire

Integrating Dublin Core/RDF records with MARC21 via the OCLC Connexion service at the Centre for Digital Library Research

Gordon Dunsire

This paper discusses the use of OCLC's Connexion service (formerly CORC) by the Centre for Digital Library Research at the University of Strathclyde, Glasgow, Scotland. The Centre has completed, and is currently engaged in, a number of research projects involving the application of Dublin Core and MARC21 in creating metadata for digital resources; these include projects on the Glasgow Digital Library, East Dunbartonshire local history, and a pilot Scottish Cultural Portal. The Connexion service provides a MARC21-DC converter, and the Centre has been investigating its incorporation in workflows for creating and maintaining digital libraries. This has included the use of databases to store metadata, with subsequent output to Dublin Core and conversion to MARC21.

The Centre for Digital Library Research

The Centre for Digital Library Research (CDLR)[1] is based at Strathclyde University in Glasgow, Scotland. CDLR was formed in 1999, and »... seeks to combine theory with practice in innovative ways ...«

The Centre is engaged in a range of research and development projects. Several of these have a primary focus on metadata issues, and include the High Level Thesaurus (HILT)[2], Harvesting Institutional Resources in Scotland Testbed (HaIRST)[3], COPAC-Clumps Interoperability (CC-interop)[4], and Scottish Portals for Education, Information and Research (SPEIR)[5] projects.

results of previous projects as test-beds and data sources for research

CDLR also manages services developed from previous projects, many of which are used as test-beds and data sources for research. These include BUBL[6], Co-operative Information Retrieval Network for Scotland (CAIRNS)[7], Scottish Collections Network (SCONE)[8], Research Collections Online (RCO)[9], and the pilot for the Scottish Cultural Portal (under construction).

Metadata for digital collections

CDLR is involved in two particular projects that have raised a number of issues in the area of management of the creation and transformation of metadata for electronic information resources.

Creating a permanent digital archive of local materials[10] was a project funded by the Scottish Library and Information Council (SLIC) »... to establish viable routines and procedures for local ›content creation‹ in small and medium-sized library authorities in Scotland.« The project involved the digitisation of 100 photographic and print items from the local studies collections of the William Patrick Library in Kirkintilloch, East Dunbartonshire, and the creation of metadata for administrative and retrieval purposes. A specific requirement of the project was to consider »...the use of established library procedures for the creation and distribution of metadata, adherence to the UK government's eGif guidelines and best practice in the cataloguing and indexing of web content.« The e-Government Interoperability Framework (eGIF)[11] requires the use of the Dublin Core (DC) metadata format for web content created by government agencies, including public libraries. At the same time, MARC21 is the metadata format established for many libraries, so the project had to investigate methods and tools for casting metadata in both formats.

The Glasgow Digital Library (GDL)[12] aims to become »... a virtual co-library of the majority of public institutions in Glasgow ...«. CDLR is managing the metadata for a number of its collections of digitised information resources: ASPECT: Access to Scottish Parliamentary election materials 1999[13]; Springburn Virtual Museum[14]; Red Clydeside[15]; Victorian times[16]; and Voyage of the Scotia[17]. Much of this material also requires conformance to eGIF as well as being cross-searchable with other library catalogues in the Z39.50 and MARC-based CAIRNS service.

Both of these areas of research investigated the use of the transformation tool in the OCLC Connexion service[18] for creating MARC21 metadata from DC and vice versa. In addition, the various strands of the GDL resulted in the development of methodologies and tools to support its policy on metadata standards[19], which seeks simplicity and flexibility in conjunction with conformance to other standards.

The general methodology used by the GDL identifies four distinct processes in the creation of metadata:

— Determination of descriptive and administrative content for resource identification.
— Determination of headings for resource retrieval.
— Creation of records in a structured repository to store descriptive content and headings.
— Casting of repository records to specific metadata formats for export to retrieval systems.

Each of these processes must resolve specific issues that may arise when creating metadata for collections of electronic resources.

Descriptive metadata

Descriptive and administrative information is required at both the item and collection levels. Individual resources need item-specific descriptors such as image caption, physical characteristics, and electronic location. Some data may be common to every resource, such as conditions governing reproduction and use, and statements of intellectual property right. Similar considerations apply to the determination of headings for retrieval, with a single metadata record containing some headings unique to the resource, and others common to all resources in the collection. This is often the case with collections of resources that have been digitised because they have a common subject or creator, for example photographs or personal archives.

Item-level description for digitised images and manuscripts can be difficult. Pictorial and archival originals may have only a short caption or brief notes, unlike printed text materials which often contain adequate metadata such as a title page and colophon. In these circumstances, the best way of determining descriptive metadata may be to use local expertise. Such expertise may be found in a local history librarian or archivist, but this will not be the case if the original resources have not been managed in a professional environment. The local expert may well be someone who has no knowledge of metadata or modern retrieval systems. If the determination and recording of item-level descriptions is to be effective and efficient in such a case, a simple and widely available data capture tool is required. MS Access is one such tool that has been used successfully by the GDL.

Descriptive elements at the collection level are often determined during the planning of digitisation projects. For example, the scope of the collection may be based on a single subject such as the Voyage of the Scotia, requiring the same subject heading to be present in every record. Statements of intellectual property rights common to each item may be a result of policies arising from the digitisation process itself, as is the case with the project on digital archiving of local materials.

This suggests a two-step approach to creating descriptive metadata. All item-specific elements can be determined and recorded in one step, with the common elements added subsequently, or vice versa, with item-specific data being added to a record pre-loaded with the common data. CDLR has identified several ways of achieving this. For the item-then-collection approach, common elements can be added to MS Access records using table-updating tools or Visual Basic scripts. Similar tools are available or can be readily developed for other SQL compatible databases. Repositories often cater for the collection-then-item approach for specific metadata formats. For example most modern library management systems offer MARC templates or work forms for pre-loading common content. Specifying default content for fields in new records is another method that can be applied in general database management systems.

toolkit to support workflows

CDLR has tested these different techniques, and developed a toolkit of scripts for processing MS Access, MS SQL Server, and XML metadata repositories. This toolkit is used to support the differing metadata creation workflows required by the GDL and other services that CDLR manages.

Headings

The GDL intends the metadata for different digital collections to be cross-searchable. This requires consistency in the content of metadata headings, so that recall is sufficiently precise, and the enquirer is not burdened with unnecessary variation in the naming of the same person, organization or subject.

Established library procedures achieve this by using authority files for the names of persons and corporate bodies, and subject topics. The use of such precoordinate headings is made more effective and efficient if authority files can be integrated with the repository for descriptive metadata. Currently, however, such integration is usually only available in mature library management systems, and is often impossible to achieve when item-level records are being created in local, non-library repositories. This suggests that authoritative headings might best be determined after descriptive records have been completed, by transferring such records from the local repository to one which offers integrated authority control.

more maintenance required for digital resources

Authority files for digital resources often require significantly more maintenance than those for print-based resources. In particular, unique materials such as photographs, printed ephemera, and manuscripts may not have been bibliographically recorded before digitisation, creating many new headings for persons, organizations, and places. Original digital resources such as organizational websites similarly require new headings to be established if the organization has not previously published anything in print.

OCLC Connexion

CDLR is a user of the Connexion service, formerly known as CORC, available from OCLC. The subscription to this service is funded by SLIC in order to support the Scottish Portals Initiative (SPI) in its aim to promote »collaborative e-resource cataloguing«. The CDLR account is, with permission of OCLC, available

for use by smaller libraries on a consortium basis. CDLR is also collaborating with OCLC in several other initiatives, including the use of the Dewey Decimal Classification in the HILT project.

development of mechanisms for collaborative authority heading maintenance

In order to foster a common approach to indexing and terminology standards, the CDLR is developing mechanisms for collaborative authority heading maintenance in partnership with the National Library of Scotland and Strathclyde University Library. The principal mechanism is participation in NACO[20] and SACO[21], the name and subject components of the Program for Cooperative Cataloging of the Library of Congress. Interoperability with NACO and SACO is incorporated into the Connexion service.

Connexion offers several tools that can be used to resolve some of the issues identified in descriptive content and heading creation. Common content can be defined in »constant data« records for incorporation in new record templates or addition to existing records, in both MARC21 and DC format. Headings can be automatically checked against the Library of Congress Name Authority File (LCNAF) and Library of Congress Subject Headings (LCSH). There is also a facility to create DC records by harvesting the meta tags of HTML documents, using a specific URL or by following hyperlinks from another document. Connexion itself is a web-based service, and can be accessed from any location with a suitable browser.

Workflows for metadata creation

The tools and facilities of Connexion, together with those developed by CDLR, form a very flexible toolkit for supporting the creation and maintenance of metadata in many different scenarios. Three examples are:

1. Creation of eGIF compliant metadata for HTML resources in a local information retrieval service.

1.1. Item-specific descriptive and heading elements are created by a local expert using MS Access forms.
1.2. Collection-level metadata are added to every record automatically via a query.
1.3. Metadata are exported from the MS Access database in DC format using a Visual Basic script.
1.4. The DC tags are embedded in the digital resource.

2. Creation of metadata for resource discovery in global and local systems.

2.1. Collection-level descriptive content and headings are added to a Connexion constant data record.
2.2. Item-level metadata are created in a MARC21 format template which uses the constant data.
2.3. Item-level headings are controlled against authority files.
2.4. Metadata records are added to WorldCat, a global retrieval system (the format must be MARC21).
2.5. Metadata for a local retrieval system are created by casting the MARC21 records to DC format and exporting them.

3. Alternatively

3.1. Item-level metadata are created locally ... (as in 1.1.)
3.2. Metadata are exported in DC format ... (as in 1.3.–1.4.)
3.3. Document URLs are added to a temporary list of hyperlinks.
3.4. Metadata are harvested by Connexion by following the links.
3.5. Metadata are cast to MARC21 format.
3.6. Collection-level metadata are added using constant data, and headings are controlled (as in 2.1. and 2.3.)
3.7. Metadata are added to WorldCat (as in 2.4.)
3.8. Metadata are cast to Resource Description Framework (RDF) format and exported to a local system.

Metadata conversion

The following example from the ASPECT service shows the results of metadata conversion using the toolkit. The item is one of a set of web pages describing each candidate in the first elections to the new Scottish Parliament in 1999.

040	CX@GDLASP $c CX@
049	CX@A
245 00	Angus Mackay.
260	[Glasgow]: $b [Centre for Digital Library Research], $c 2002.
500	HTML Title: Aspect: Angus MacKay, Scottish Labour Party candidate, Edinburgh South, 1999.
522	Constituency: Edinburgh South.
540	GDL: Glasgow Digital Library.
651 0	Edinburgh (Scotland)
856 40	$u http://gdl.cdlr.strath.ac.uk/aspect/lab/labeds.htm $q text/html

Figure 1: This is an extract from a MARC21 display record generated from MS Access and imported into Connexion.

The result of using Connexion to cast to DC format in HTML:

<meta name="DC.Title" content="Angus Mackay">
<meta name="DC.Coverage.spatial" content="Constituency: Edinburgh South">
<meta name="DC.Publisher" content="[Centre for Digital Library Research],">
<meta name="DC.Publisher.place" content="[Glasgow]:">
<meta name="DC.Date.issued" scheme="MARC21-Date" content="1999">
<meta name="DC.Description.note" content="HTML Title: Aspect: Angus MacKay, Scottish Labour Party candidate, Edinburgh South, 1999">
<meta name="DC.Identifier" scheme="URI" content="http://gdl.cdlr.strath.ac.uk/aspect/lab/labeds.htm">

And the result of casting to RDF:

<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.0/"
xmlns:dcq="http://purl.org/dc/qualifiers/1.0/">
<rdf:Description about="http://gdl.cdlr.strath.ac.uk/aspect/lab/labeds.htm">
<dc:title>Angus Mackay</dc:title>
<dc:coverage>Constituency: Edinburgh South</dc:coverage>
<dc:publisher>[Centre for Digital Library Research],</dc:publisher>
<dc:publisher>[Glasgow]:</dc:publisher>
<dc:date>1999</dc:date>
<dc:description>HTML Title: Aspect: Angus MacKay, Scottish Labour Party candidate, Edinburgh South, 1999</dc:description>
<dc:identifier>http://gdl.cdlr.strath.ac.uk/aspect/lab/labeds.htm</dc:identifier>

Workflow issues

remaining issues

Use of the CDLR toolkit to support workflows involves a number of issues that remain to be resolved.

— The Connexion tool for casting MARC21 to DC ignores local MARC21 tags and drops their content.

— Casting MARC21 records to DC results in data loss because MARC21 has a richer structure.

— MARC21 records cast from DC usually require additional content to meet international cataloguing standards.

— Headings in DC records can only be controlled in Connexion by casting to MARC21, and then recasting to DC.

— The use of Connexion constant data records to add collection level metadata requires mediation, and can only be carried out on one record at a time.

Conclusion

OCLC Connexion offers a useful set of tools for authority control and conversion of metadata between MARC21 and Dublin Core formats.

These tools in combination with tools developed locally for database maintenance can support many different workflows and methods for creating and maintaining metadata for digital collections.

The toolkit can be employed to optimise the use of local expertise in a global environment and reduce the need for retroconversion of legacy metadata in the future.

Acknowledgement

This paper was prepared with the invaluable assistance of Alan Dawson, who created many of the CDLR metadata tools, the »Dawson toolkit«.

References and links

[1] CDLR: http://cdlr.strath.ac.uk/ (checked 24 Feb 2003)

[2] HILT project: http://hilt.cdlr.strath.ac.uk/ (checked 24 Feb 2003)

[3] HaIRST project: http://hairst.cdlr.strath.ac.uk/

[4] CC-interop project: http://ccinterop.cdlr.strath.ac.uk/ (checked 24 Feb 2003)

[5] SPEIR project: http::://speir.cdlr.strath.ac.uk/ (checked 24 Feb 2003)

[6] BUBL service: http://bubl.ac.uk/ (checked 24 Feb 2003)

[7] CAIRNS service: http://cairns.lib.strath.ac.uk/ (checked 24 Feb 2003)

[8] SCONE service: http://scone.strath.ac.uk/service/index.cfm (checked 24 Feb 2003)

[9] RCO service: http://scone.strath.ac.uk/rco/index.cfm (checked 24 Feb 2003)

[10] Creating a permanent digital archive of local materials: a SLIC funded project / project manager, Don Martin; researcher, Stephen Winch: http://www.slainte.org.uk/slicpubs/cpdalm.pdf (checked 24 Feb 2003)

[11] e-GIF: http://www.govtalk.gov.uk/interoperability/egif.asp (checked 24 Feb 2003)

[12] Glasgow Digital Library service: http://gdl.cdlr.strath.ac.uk (checked 24 Feb 2003)

[13] ASPECT service: http://gdl.cdlr.strath.ac.uk/aspect/ (checked 24 Feb 2003)

[14] Springburn Virtual Museum: http://gdl.cdlr.strath.ac.uk/springburn/ (checked 24 Feb 2003)

[15] Red Clydeside: http://gdl.cdlr.strath.ac.uk/redclyde/ (checked 24 Feb 2003)

[16] Victorian times: http://vt.cdlr.strath.ac.uk/ (checked 24 Feb 2003)

[17] Voyage of the Scotia: http://gdl.cdlr.strath.ac.uk/springburn/index.html (checked 24 Feb 2003)

[18] Connexion: http://www.oclc.org/connexion (checked 24 Feb 2003)

[19] Glasgow Digital Library draft metadata standards policy: http://gdl.cdlr.strath.ac.uk/documents/gdlmetadatapolicy.htm (checked 24 Feb 2003)

[20] NACO: http://www.loc.gov/catdir/pcc/naco.html (checked 24 Feb 2003)

[21] SACO: http://www.loc.gov/catdir/pcc/saco.html (checked 24 Feb 2003)

Der Verfasser

Gordon Dunsire ist stellvertretender Direktor des Centre for Digital Library Research, Strathclyde University, Glasgow, G4 0NS
g.dunsire@strath.ac.uk