Discovery metadata - Dublin Core, MODS, MARC, ...

Michael Day and Pete Johnston UKOLN, University of Bath m.day@ukoln.ac.uk

Cataloguing Online Resources: an Introduction to Metadata for Librarians, Manchester, 26 April 2006



http://www.ukoln.ac.uk/



Session overview

- Resource discovery
- Dublin Core
- The MARC formats
- MODS



http://www.ukoln.ac.uk/



Resource discovery (1)

- A basic function of metadata
- Part of information retrieval
- Cutter's principles from "Rules for a Dictionary Catalog" (1876), slightly paraphrased:
 - To enable a person to find a [book] of which either the author, title or subject is known
 - To show what a [library] has by a given author, on a given subject, or in a given kind of literature
 - To assist in the choice of a [book] as to its edition (bibliographically) or to its character (literary or topical)



http://www.ukoln.ac.uk/



Resource discovery (2)

- A particular challenge in Web environment
 - Resource providers have moved into a shared network space
 - Recognition that users wish:
 - "to refer to intellectual and cultural materials flexibly and transparently without concern for institutional or national boundaries" (Dempsey, 2000)
- This is the problem that Dublin Core is designed to address (cross-domain discovery)



http://www.ukoln.ac.uk/



Resource discovery (3)

- We will now look in more detail at three standards *primarily* developed to support resource discovery
 - Dublin Core
 - The MARC formats
 - MODS



http://www.ukoln.ac.uk/



Dublin Core basics

- Perhaps the most well-known metadata initiative (there are many implementations)
- Named after a workshop held in Dublin, Ohio a suburb of Columbus
- Mainly focused on cross-domain resource discovery
- A suite of standards (and other activities) organised as part of the Dublin Core Metadata Initiative (DCMI)



http://www.ukoln.ac.uk/



DCMI mission

- Providing simple standards to facilitate the finding, sharing and management of information, by:
 - Developing and maintaining international standards for describing resources
 - Supporting a worldwide community of users and developers
 - Promoting widespread use of Dublin Core solutions



http://www.ukoln.ac.uk/



DCMI brief history (1)

- Mid-1990s

- Challenge of discovery on the Web
- Search engines providing many hits, but little precision (pre Google)
- Recognition that the traditional library approach to cataloguing could not scale to Web resources

– 1995 - first workshop

- Hosted by OCLC at Dublin, Ohio
- Primarily focused on Web resource discovery (document-like objects)
- Resulted in interdisciplinary consensus on 13 metadata elements



http://www.ukoln.ac.uk/



DCMI brief history (2)

- 1996 2nd and 3rd workshops:
 - DC-2 (University of Warwick)
 - Recognised that DC elements would need to combine or co-exist with other types of metadata (modularity)
 - Warwick Framework devised to deal with this
 - DC-3 (Dublin, Ohio)
 - Workshop convened to deal with images (expanding beyond document-like objects)
 - Explicit focus now on cross-domain resource discovery
 - First identification of the 15 core elements

http://www.ukoln.ac.uk/



DCMI standardisation

- Dublin Core Metadata Element Set
 - Version 1.0: IETF RFC 2413 (1998)
 - Version 1.1: CEN Workshop Agreement CWA 13874 (2000), NISO Z39.85-2001, ISO 15836:2003
 - DCMI Recommendation (2004)
- DCMI Metadata Terms
 - DCMI Recommendation (latest version, 2005)
 - Specifies all metadata terms maintained by DCMI: elements, element refinements, encoding schemes, vocabulary terms
- DCMI Abstract Model
 - DCMI Recommendation (2005)



http://www.ukoln.ac.uk/



Dublin Core elements (1)

- Interdisciplinary consensus on simple element set for resource discovery
 - 15 elements
 - All optional
 - All repeatable
- Not intended for complex resource description
 - Initial idea of "simple document-like object"
 - Simplicity of semantics, ease of use
- Provides basic "semantic interoperability"
 - Across domains, across language communities
 - Does not provide detailed cataloguing rules
- A set of 15 broad "buckets"...



http://www.ukoln.ac.uk/



Dublin Core elements (2)

- Title
- Subject
- Description
- Creator
- Publisher
- Contributor
- Date

- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights



http://www.ukoln.ac.uk/



Dublin Core elements (3)

- Not a replacement for richer descriptive standards
- Can provide 15 "windows" into richer resource descriptions
 - disclose rich description in simple form
 - semantic cross-walks, mappings to existing data
 - export rather than create
- If metadata is language ...
 - ... then DC is a "pidgin" language for use by "tourists on the Internet commons" (Thomas Baker)



http://www.ukoln.ac.uk/





Dublin Core elements (4)

- Small vocabulary, simple grammar/structure
 - Resource has Title "An Introduction to Dublin Core and the DCMI"
 - Resource has Subject "Metadata"
- Not as subtle/powerful as separate languages but can be useful!



http://www.ukoln.ac.uk/



Extending Dublin Core

- Element refinements:
 - Narrow the meaning of a DC element
 - e.g. "date modified" v "date"
- Encoding schemes:
 - Provide additional information about a value
 - e.g. can indicate that a subject value is a Library of Congress Subject Heading
- The "Dumb-Down" principle
 - Provides rules for transforming "qualified" description into "simple" description
- the "One-to-One" rule
 - A DC description describes exactly one resource



http://www.ukoln.ac.uk/





Dublin Core Application Profiles

- In practice, metadata implementers
 - Combine elements from different sources (e.g. DC plus elements from other schemas, "local" elements)
 - Refine definitions of elements
 - **Constrain** use of elements
- Application profiles
 - If simple DC is a "pidgin", an application profile is a "regional idiom or creole"! (Thomas Baker)
 - Element set plus policies, guidelines
 - Some DCMI Working groups developing application profiles for specific domains (government, education)



http://www.ukoln.ac.uk/



DC Application Profiles: examples

- "Simple Dublin Core"
 - Use of the 15 properties of the DCMES
 - All optional and repeatable
 - Values represented by value strings
 - No vocabulary or syntax encoding schemes
- UK eGovernment Metadata Standard (eGMS)
 - Use of selected properties from DCMI vocabularies, additional properties
 - Guidelines on use of properties
 - Some properties mandated/recommended
 - Some vocabulary encoding schemes mandated/recommended
 - Guidance on content of value strings



http://www.ukoln.ac.uk/



Some applications of Dublin Core

- Embedded in Web pages
- Integrated resource discovery services
 - For example
 - Subject Gateways Resource Discovery Network
 - OAI Service Providers OAIster
 - Image services Picture Australia



http://www.ukoln.ac.uk/



DC embedded in X/HTML

- Search crawlers can extract metadata from individual pages
- However, little or no use by the major search engines
 - Robot spamming problems
 - Lack of metadata (or quality-control)
 - Availability of better indexing tools, e.g. Google's PageRank algorithm
- But, useful in controlled environments



http://www.ukoln.ac.uk/



Dicturo Australia
d
– Images "related to all things
Australian" from 40+ cultural
agencies
ageneics
– Central search service based
(initially at least) on crawling
HTMI -embedded DC metadata
Providers currently migrating to
ine. To search for images not yet online, you will need to visit the individual web sites
ap Trails

http://www.pictureaustralia.org/

🔀 Find: user

PictureAustralia More informal	tion - Mozilla Firefox				
<u>File E</u> dit <u>V</u> iew <u>G</u> o <u>B</u> ookmarks	Tools Help				
🔄 • 🗣 • 🛃 😣 😚 🔳 🕑	http://www.pictureaustralia.org/app	ps/pictureaustralia?a	action=PADisplay&mode=display&rs=resultset-134	418no=2	🗾 🔀 etadata harvestin
m PictureAustralia	📔 nolan gallery metadata har	vesting - Google	🛛 🎂 16ft Hire Boats built by Boyd (25 were built)	PictureAustralia More information	(
	PICTURE				
Á lega	USTRALIA				
Home Who	se images? About us Contact us F	AQs Site map	News Ordering images Links		
Search for im	More information				
Advanced Sea	Your search for influ results Displaying image: 2 arch	uenza in title fou	und 28 images.		
Browse Help	< <u>View first page</u> -	< <u>View previous p</u>	page Brief view View next page > View	ast page >>	
		Add to fav	vourites		
	Title	[Army person	nnel being inoculated against influenza at ar	n Army depot near Melbourne]	
PICTORE TRAIL	Subject	Walter and El	liza Hall Institute of Medical Research.		
VIEW FAVOURIT	res ► Subject	Immunization	n Australia		
	Description	Shows Capt. S	Steven Williams of the Walter and Eliza Hall	Institute, on loan to the army to vaccinate per	sonnel.
	Image number	H99.201/112	23		
	Format	photograph :	: gelatin silver ; 12.5 × 16.5 cm. approx.		
	Managed by	Item held by	State Library of Victoria		
	Collection or series	IspartOf Argu	us newspaper collection of war photographs.	World War II.	
	Date or place	[ca. 1944]			
	Rights	Reproduction	rights: State Library of Victoria.		
	Go to	Original imag	<u>ae</u>		
	Page: <u>1</u> 2 <u>3</u> 4 <u>5</u> 6 <u>7</u> 8 << <u>View first page</u> < Search for images: GO In these results <u>Advanced Search</u> <u>Browse</u> <u>Help</u> <u>Back to top</u>	8 9 10 11 12 13 1	14 15 16 17 18 19 20 next	ast page >>	
	Home Whose Image	es? About Us	Contact Us FAQs Site Map		
	News Ordering Ima	ges Links Cor	pyright For Contributors Trails		

🔯 Find: user

-

5 • 🖒 • 🎜 🔊 🤗 🔟 /	D film http://www.sly.vic.gov.au/argus/0/0/2/doc/ap002100_shtml	▼ C etelete herved
M PictureAustralia	📔 nolan gallery metadata harvesting - Google 🛛 🌅	1 😻 view-source: - Source of: http://www.slv.vic.gov.au/argus/0/0/2/doc/an002190.shtml - Mozilla Fi 💶 🕮 🗶
		<u>Eile Edit V</u> iew
State Library of Victoria		
	Catalogues & Databases Collections Services	<title>Army personnel being inoculated against influenza at an Army depot near Melbourne [picture] by </title>
		<meta content="text/html; charset=utf-8" http-equiv="Content-Type"/>
Pictures Catalogue > Army personnel being inoculated against influenza at a near Melbourne [picture] Date(s) of creation: [ca. 1944] photograph : gelatin silver ; 12.5 x 16.5 cm. approx. Reproduction rights: State Library of Victoria. Accession Number: H99.201/1123 Umage Number: ap002190	<pre><meta content="Shows Capt. Steven Williams of the Walter
and Eliza Hall Institute, on loan to the army to vaccinate personnel. " name="description"/> a <meta content="Ships,Gelatin silver photographs,Walter and
Eliza Hall Institute of Medical Research,Walter and Eliza Hall Institute of
Medical Research,state,library,image,victoria,melbourne" name="keywords"/> k rel=" schema.AGLS" href="http://www.naa.gov.au/recordkeeping/gov online/agls/1.2"></pre>	
	Image Number: an002190	<pre>hret="http://www.haa.gov.au/recordReeping/gov_online/agis/1.2"> (meta NAME="DC.Title"LANG="en"CONTENT="[Army personnel being inoculated against influenza at an Army depot near Melbourne] "> (meta NAME="DC.Subject" SCHEME="LCNA" CONTENT="Walter and Eliza Hall Institute of Medical Research."> (meta NAME="DC.Subject" SCHEME="LCSH" CONTENT="Walter and Eliza Hall Institute of Medical Research."> (meta NAME="DC.Subject" SCHEME="LCSH" CONTENT="Immunization Australia"> (meta NAME="DC.Date.Created" SCHEME="ISO8601" CONTENT="5 September 2005"> (meta NAME="DC.Type" CONTENT="Gelatin silver photographs."> (meta NAME="DC.Identifier.URL" (CONTENT="DC.Identifier.URL" (CONTENT="DC.Identifier.URL" (CONTENT="Immunization silver photograph."> (meta NAME="DC.Identifier.URL" (CONTENT="DC.Identifier.URL" (met</pre>
	<u>Feedback on this image</u> State Library of Victoria is a contributor to <u>PictureAust</u>	<pre>state NAME = DC.Rights" CONTENT="Reproduction rights: State Library of Victoria."></pre>
		<style media="screen" type="text/css"> @import "/headerpages/picsdb.css"; </style>
		Victoria

Dublin Core and the OAI-PMH (1)

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
 - Fairly simple mechanism for sharing metadata records between applications
 - Has origins in "e-prints" community
 - Built on HTTP, XML
 - Allows a harvester to ask a repository for all or some of its metadata records (in a specified metadata format)
 - e.g., "Give me all your records updated since yyyymm-dd"



http://www.ukoln.ac.uk/



Dublin Core and the OAI-PMH (2)

- "OAI-DC" (Simple DC) is mandatory format
 - But no limitation on format that can be transferred (as long as can be described by XML Schema)



http://www.ukoln.ac.uk/



The MARC formats (1)

- Machine-Readable Cataloguing (MARC)
 - Will be familiar to most librarians
 - Integral to bibliographic cataloguing practice in many countries since the 1960s
 - Not a single standard, but a family of formats, e.g. MARC 21, UNIMARC, UKMARC
 - Facilitates the exchange of bibliographic data (shared cataloguing)
 - Determines search functionality in library catalogues (OPACs)



http://www.ukoln.ac.uk/



The MARC formats (2)

- Format first developed long before the term "metadata" was coined
- The MARC formats are standards for "the representation and communication of bibliographic and related information in machine-readable form" (MARC 21)
 - Machine-readable = data can be read, interpreted and manipulated by computers
 - Integrally linked with a range of standards that define field content, e.g. the International Standard Bibliographic Description (ISBD) series, cataloguing rules (e.g. AACR2, RDA)



http://www.ukoln.ac.uk/



MARC 21 basics

- Format resulted from the "harmonisation" of USMARC and CANMARC
- Maintained by Library of Congress and Library and Archives Canada
- Separate formats defined for:
 - Bibliographic data
 - Authority data
 - Holdings data
 - Classification data
 - Community Information data



http://www.ukoln.ac.uk/



MARC 21 Structure (1)

- Structure (based on ISO 2709)
 - Leader (24 characters)
 - Directory
 - Control fields (fixed length, mostly codes)
 - Variable fields
- Variable fields include:
 - Bibliographic description (0XX, 2XX, 3XX, 4XX), including notes (5XX) - typically follows ISBD
 - Main Entry (1XX), Other Added Entries (7XX, 8XX)
 - Subject Entries and Classification (0XX, 6XX)
 - Electronic Location and Access (856)



http://www.ukoln.ac.uk/



MARC 21 Structure (2)

 Variable Fields in a typical bibliographic record will look something like:





http://www.ukoln.ac.uk/



MARC 21 - main features

- Builds on 150 years of modern cataloguing practice
- Builds on external standards (e.g. ISBN) and controlled vocabularies (e.g., name authorities, subject headings)
- Used for many types of bibliographic item: books, serials, maps, music, electronic resources, ...
- The basis of shared cataloguing services (e.g. OCLC's WorldCat)
- Many million MARC records in library systems



http://www.ukoln.ac.uk/



MARC 21 in XML

- MARC is now quite an old standard:
 - Initially developed to automate the printing of catalogue cards in 1960s
 - Legacy of the card format remains, e.g. the concept of main entry, poor linking between related items
 - Other legacy issues related to structure and character sets, e.g. ISO 2709
- MARC 21 XML Schema
 - Library of Congress developing a framework for working with MARC 21 in an XML environment
 - More flexibility for internal linking, developing crosswalks with Dublin Core and other formats, etc.



http://www.ukoln.ac.uk/



MODS basics (1)

- Metadata Object Description Schema
 - Maintained by Library of Congress Network
 Development and MARC Standards Office
 - More extensive than Dublin Core, 19 top-level elements, 64 at lower levels
 - Grounded in MARC 21
 - Includes a subset of MARC 21 fields (logically restructured), inherits some MARC semantics
 - Expressed in the XML Schema language
 - Is extensible
 - Can integrate with standards like METS



http://www.ukoln.ac.uk/



MODS basics (2)

- Specifically designed for library operations
 - e.g., Digital library systems, digitisation projects
- A possible alternative to Dublin Core?
 - Integrates better with the existing MARC corpus
 - Worth investigating for library-type operations
 - Untested in cross-domain contexts, MARC legacy may not be so useful here



http://www.ukoln.ac.uk/





Summing up:

- Three standards:
 - Dublin Core
 - A "core" for discovery of wide range of resources, focus on cross-domain discovery
 - Limited functionality, unless extended, e.g. using Application Profiles
 - MARC 21
 - A proven role for bibliographic data; not particularly suitable (or designed) for other resource types
 - MODS
 - Promising new XML-based standard, less complex than MARC 21, will have applications in libraries



http://www.ukoln.ac.uk/



More information:

- DCMI: http://dublincore.org/
- MARC 21: http://www.loc.gov/marc/
- MARC 21 XML Schema: http://www.loc.gov/standards/marcxml/
- MODS: http://www.loc.gov/standards/mods/



http://www.ukoln.ac.uk/



Discovery metadata - Dublin Core, MODS, MARC, ...

Michael Day and Pete Johnston UKOLN, University of Bath m.day@ukoln.ac.uk

Cataloguing Online Resources: an Introduction to Metadata for Librarians, Manchester, 26 April 2006



http://www.ukoln.ac.uk/



Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union, and other sources. UKOLN also receives support from the University of Bath, where it is based.

http://www.ukoln.ac.uk/

The *Digital Curation Centre* is funded by the JISC and the UK Research Councils' e-Science Core Programme.

http://www.dcc.ac.uk/



http://www.ukoln.ac.uk/





