Metadata - general introduction

Michael Day

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/

Event timetable

- 09:30 Registration
- **10:00 Metadata - general introduction**
- **10:15 Discovery metadata**
- 11:00 *Break*
- 11:15 Learning Object metadata
- **12:00 Other types of metadata**
- 13:00 *Lunch*
- 14:00 Metadata in practice - JORUM & LOM
- 15:00 Feedback and final discussion
- 15:30 *Close*

Session overview

- Metadata - general overview
  - Definitions
  - Some basic questions
  - Metadata standards

Defining metadata (1)

- Some definitions:
  - Literally, "data about data"
    - Defines the basic concept, but is (perhaps) not very meaningful
    - Refers to everything and nothing (Wendy Duff, 2004)
  - "Machine-understandable information about Web resources or other things" - Tim Berners-Lee, W3C (1997)

Defining metadata (2)

- "Structured data about resources that can be used to help support a wide range of operations - Michael Day, 2001
- "Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage" information objects - NISO, 2004
  - Hints at the many roles metadata can support

Defining metadata (3)

- Metadata is now typically defined by *function*
  - "Data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics" (Dempsey & Heery, 1998)
  - Popular categorisation:
    - » Descriptive metadata
    - » Structural metadata
    - » Administrative metadata

What functions can be supported?

- Resource disclosure & discovery
- The retrieval and use of resources
- Resource management, including preservation
- Verification of authenticity
- Intellectual property rights management
- Commerce
- Content-rating
- Authentication and authorisation
- Personalisation and localisation of services
- …

To what can metadata be applied?

- "Web resources or other things," e.g.:
- Web sites, Web pages, digital images, databases, books, museum objects, archival records, collections, services, geographical locations, organisations, events, concepts, … even metadata itself

Where can metadata be found?

- Within a resource, e.g.:
  - Title page and table of contents (books), META tags in document headers (Web pages), ID3 metadata (MP3), "file properties" (office documents), EXIF data (images)
- Directly linked to the resource, e.g.:
  - Link rel="meta" elements (Web pages)
- Independently managed in a separate database; can be linked by identifiers
  - This is the most common approach

How important is metadata?

- … "is recognised as a critically important, and yet increasingly problematic and complex concept with relevance for information objects as they move through time and space" -- Gilliland-Swetland (2004)

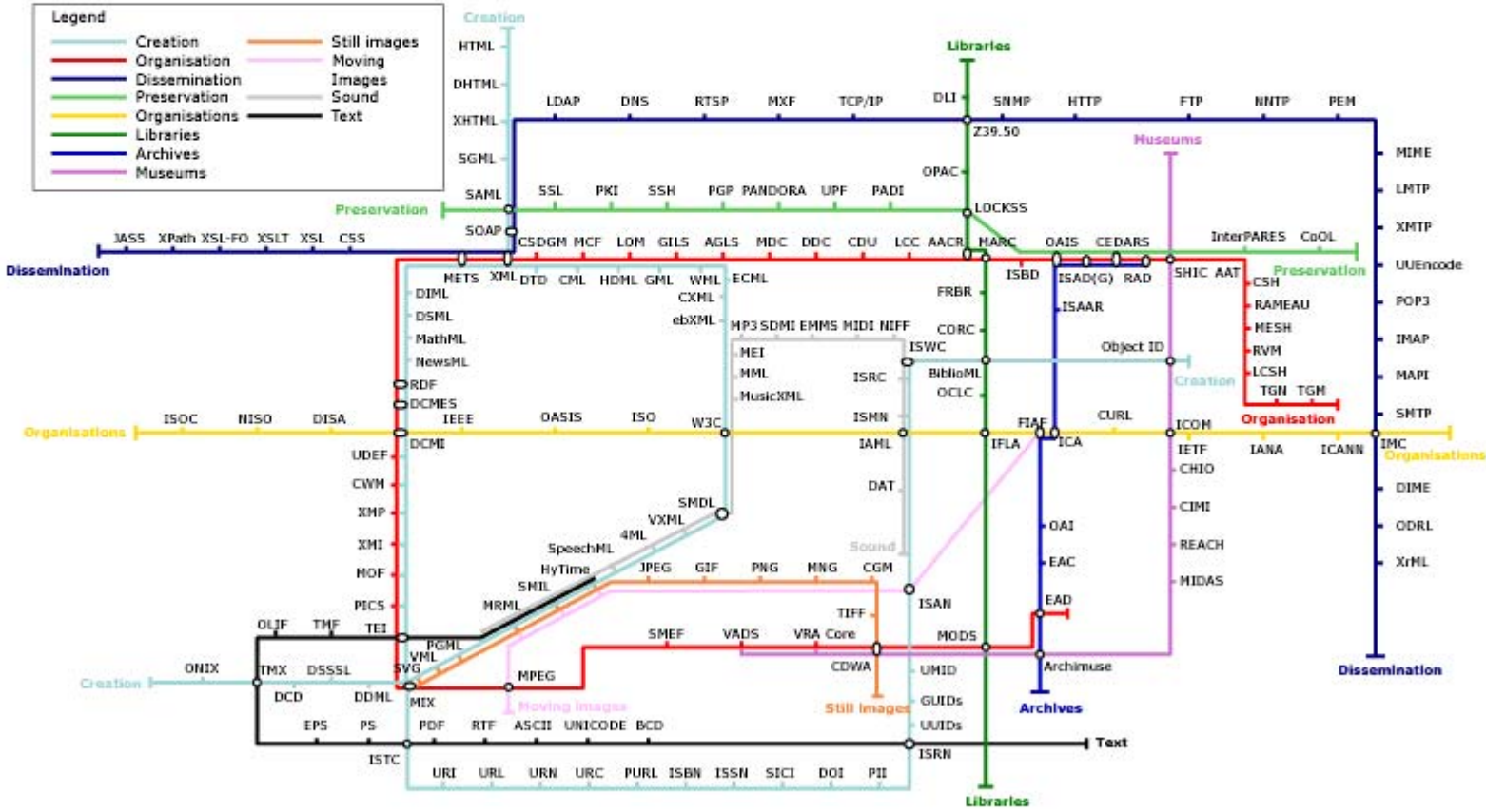
Metadata standards (1)

- But there are a large (and growing) number of metadata initiatives, formats, schemas, etc.
  - See James Turner's MetaMap for one attempt to visualise the metadata information space:
  - **http://mapageweb.umontreal.ca/turner/meta/english/**

MetaMap

[© 2004 James M. Turner, Véronique Moal, & Julie Desnoyers]

Position the mouse over an acronym to see what it stands for in a popup window.
Click on the acronym to see its definition and a link to its official site.



Metadata standards (2)

- Typically defined by "resource management communities"
  - Different traditions, perspectives, functional requirements
- Typically comprise:
  - A "conceptual model" (sometimes not explicit)
  - A set of named components ("terms", "elements" etc) and documentation on their meaning and use
  - A specification of how to represent a metadata instance in a digital format (binding)

Some examples (1)

- Bibliographic:
  - MARC (Machine-Readable Cataloguing) formats, e.g. MARC21
    - Exchange format since 1960s
    - Content often based on family of related standards, e.g. the ISBD series, AACR2
  - MODS (Metadata Object Description Schema)
    - A subset of MARC
  - ONIX
    - Used by publishers and the book trade

Some examples (2)

- Archives and records:
  - ISAD(G) (General International Standard Archival Description)
  - EAD (Encoded Archival Description)
  - EAC (Encoded Archival Context)
  - Recordkeeping metadata (e.g., ERMS (The National Archives), RKMS)
- Museum objects (and collections):
  - SPECTRUM

Some examples (3)

- Digital images:
  - VRA Core, NISO Technical Metadata for Digital Still Images
- Government information:
  - AGLS, e-GMS
- Learning objects:
  - IEEE LOM, UK LOM Core, IMS specifications
- Multimedia:
  - MPEG-7, MPEG-21 (for rights information)

Summing up

- Metadata is ubiquitous
- Metadata enables people and software applications to do things (functions)
  - Not only about "discovery"
  - Different functions require different metadata
- There are many different standards
- Challenges remain in working across standards, or in using standards in combination

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

Session overview

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

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)

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)

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

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)

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

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

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

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)

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”…

Dublin Core elements (2)

- Title
- Subject
- Description
- Creator
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

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)

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!

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

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)

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

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

DC embedded in X/HTML

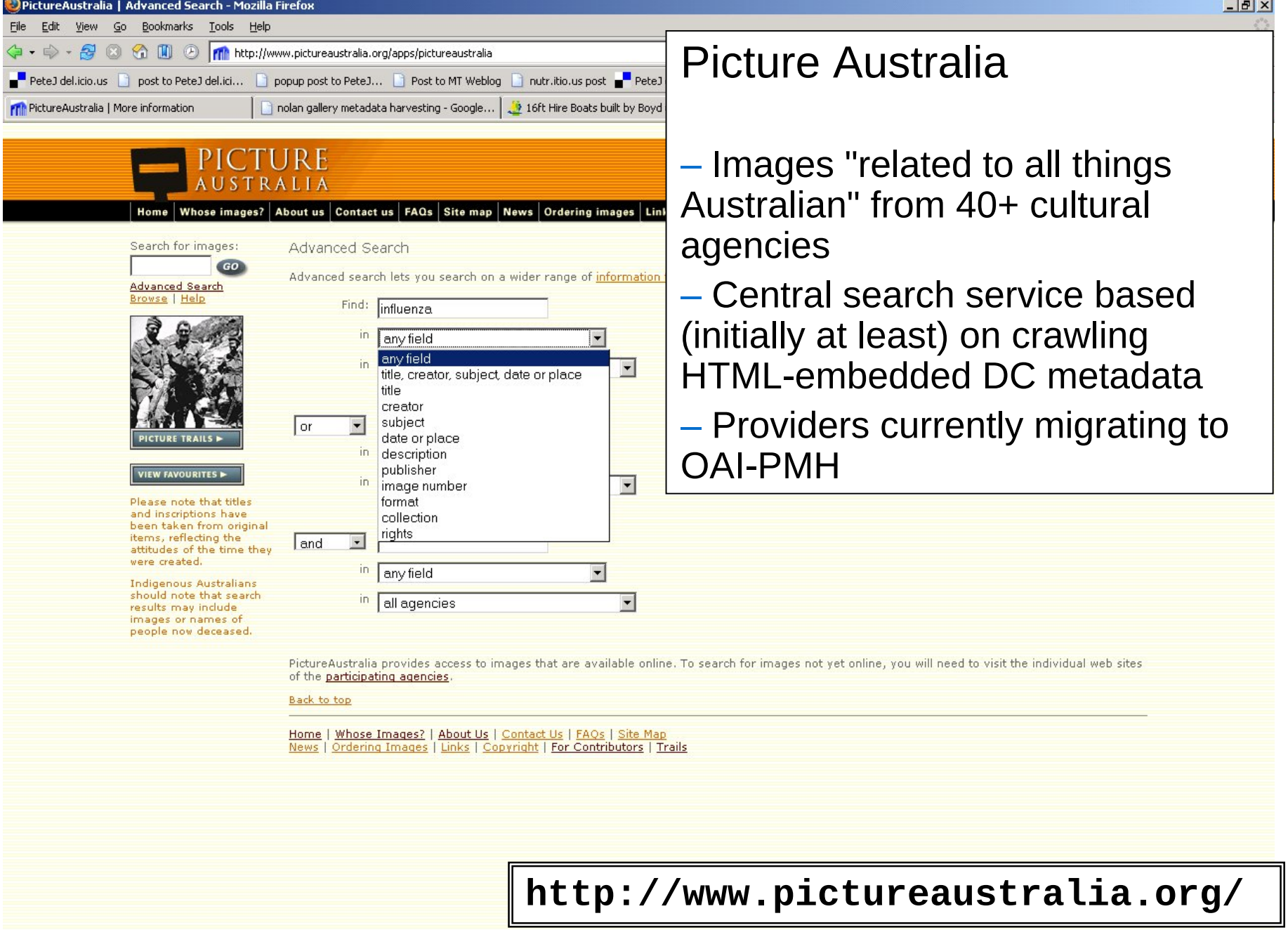
- Search crawlers can extract metadata from individual pages
- However, little or no use by the major search engines
  - Robot spamming problems
  - Lack of metadata (or quality-control)
  - Availability of better indexing tools, e.g. Google's PageRank algorithm
- But, useful in controlled environments

Picture Australia

– Images "related to all things Australian" from 40+ cultural agencies

– Central search service based (initially at least) on crawling HTML-embedded DC metadata

– Providers currently migrating to OAI-PMH



http://www.pictureaustralia.org/

PICTURE AUSTRALIA

More information

Your search for **influenza** in title found 28 images.
Displaying image: 2

Add to favourites

Title	[Army personnel being inoculated against **influenza** at an Army depot near Melbourne]
Subject	Walter and Eliza Hall Institute of Medical Research.
Subject	Immunization -- Australia
Description	Shows Capt. Steven Williams of the Walter and Eliza Hall Institute, on loan to the army to vaccinate personnel.
Image number	H99.201/1123
Format	photograph : gelatin silver ; 12.5 x 16.5 cm. approx.
Managed by	Item held by State Library of Victoria
Collection or series	IspartOf Argus newspaper collection of war photographs. World War II.
Date or place	[ca. 1944]
Rights	Reproduction rights: State Library of Victoria.
Go to	Original image

Search for images: GO
in these results
Advanced Search
Browse | Help

Army personnel being inoculated against influenza at an Army depot near Melbourne [picture] by - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.slv.vic.gov.au/argus/0/0/2/doc/an002190.shtml

PictureAustralia | nolan gallery metadata harvesting - Google...

State Library of Victoria

Catalogues & Databases | Collections | Services

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
Image Number: an002190

Feedback on this image

State Library of Victoria is a contributor to PictureAustr

Victoria The Place To Be

view-source: - Source of: http://www.slv.vic.gov.au/argus/0/0/2/doc/an002190.shtml - Mozilla Fi...

File Edit View

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN">
<html xmlns:SLV>
<head>
<title>Army personnel being inoculated against influenza at an Army depot
near Melbourne [picture]   by  </title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta NAME="description" CONTENT="Shows Capt. Steven Williams of the Walter
and Eliza Hall Institute, on loan to the army to vaccinate personnel. ">
<meta NAME ="keywords" 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">
<link rel=" schema.AGLS"
href="http://www.naa.gov.au/recordkeeping/gov_online/agls/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="Immunization -- Australia">
<meta NAME="DC.Description" CONTENT="Shows Capt. Steven Williams of the
Walter and Eliza Hall Institute, on loan to the army to vaccinate
personnel. ">
<meta NAME="DC.Date.Created" SCHEME="ISO8601" CONTENT="5 September 2005">
<meta NAME="DC.Type" CONTENT="Gelatin silver photographs. ">
<meta NAME="DC.Format" CONTENT="photograph : gelatin silver ; 12.5 x 16.5
cm. approx. ">
<meta NAME="DC.Identifier.URL"
CONTENT="http://www.slv.vic.gov.au/argus/0/0/2/doc/an002190.shtml">
<meta NAME="DC.Identifier.URL.thumbnail"
CONTENT="http://www.slv.vic.gov.au/argus/0/0/2/tn/an002190.jpg">
<meta NAME="DC.Identifier" CONTENT="H99.201/1123">
<meta NAME="DC.Source" CONTENT="Item held by State Library of Victoria">
<meta NAME="DC.Language" SCHEME="RFC 1766" CONTENT="en">
<meta NAME="DC.Relation" CONTENT="IspartOf Argus newspaper collection of
war photographs. World War II.">
<meta NAME="DC.Coverage.Temporal" SCHEME="ISO8601" CONTENT="[ca. 1944] ">
<meta NAME="DC.Rights" CONTENT="Reproduction rights: State Library of
Victoria.">
<style type="text/css" media="screen">
    @import "/headerpages/picsdb.css";
</style>
```

Find: user | Find Next | Find Previous | Highlight | Match case

Applet v51 started

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 yyyy-mm-dd"

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)

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)

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)

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

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)

MARC 21 Structure (2)

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


```
[...]
100 1  Hardy, Thomas $d 1840-1928
245 14 The return of the native $c Thomas Hardy
260 0  London $b Macmillan $c 1927
300    ix, 482 p $b map $c 19 cm.
[...]
```

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

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.

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

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

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

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/

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/

UKOLN

MLA MUSEUMS LIBRARIES ARCHIVES

JISC

UNIVERSITY OF BATH

DCC

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/