

# *A Repository of Metadata Crosswalks*



Jean Godby, Devon Smith, Eric Childress, Jeffrey A. Young

OCLC Online Computer Library Center

Office of Research

DLF-2004 Spring Forum

April 21, 2004

# *Outline of this talk*

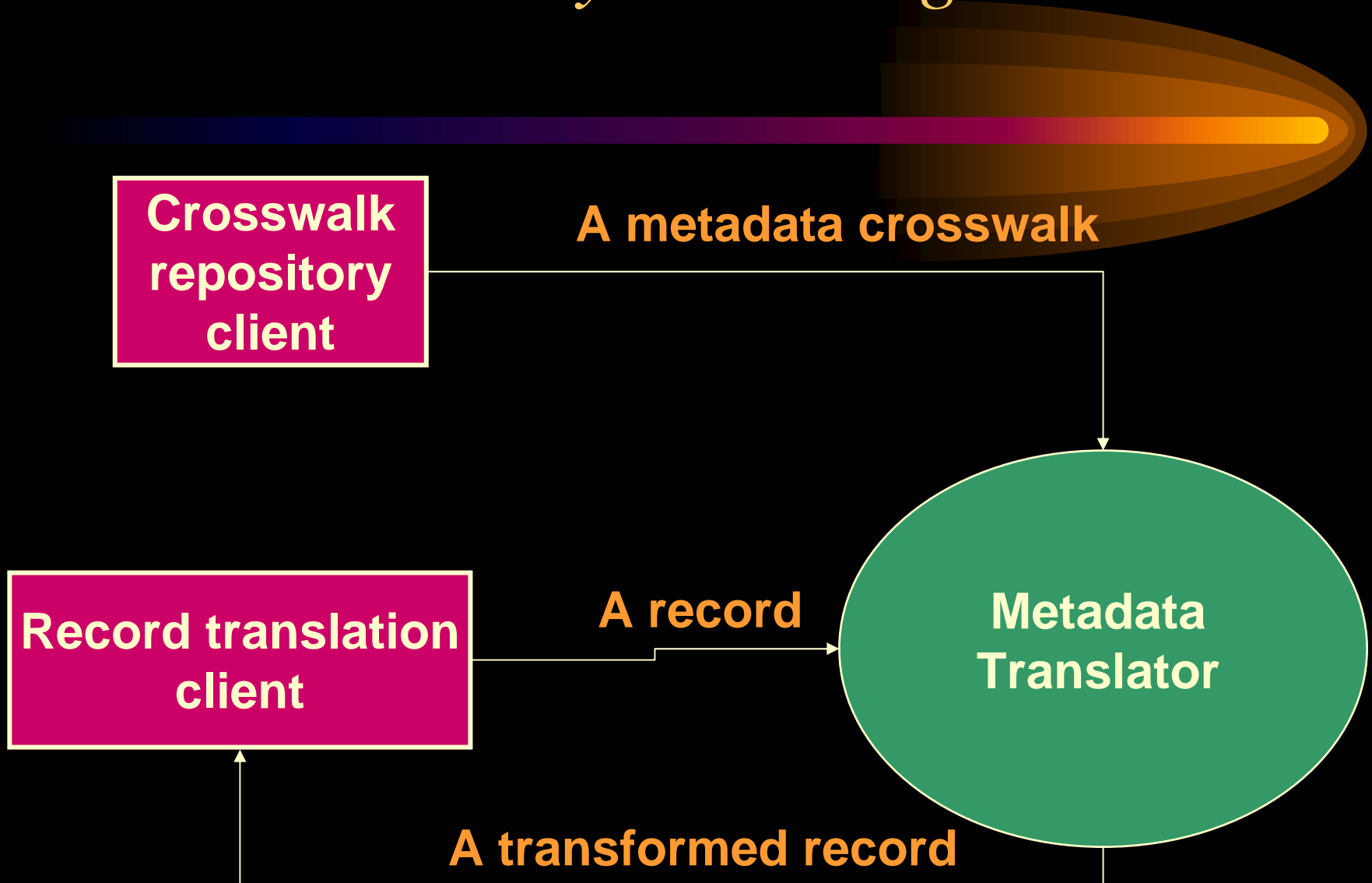


- Crosswalks and metadata translation
- Solution 1: A collection of XSLT scripts
- Solution 2: Pathfinders for crosswalks
- Solution 3: A crosswalk repository
- Open issues
- Project status

# *Our research project goals*

- **A robust design for metadata translation**
  - A clean separation of:
    - document data model
    - schema translations
    - machinery
  - Support for current practice and foreseeable innovation
- **A metadata translation system/toolkit**
  - An ‘unplugged’ service for metadata translation
  - A place for human input (intellectual mappings) in an automated system

# *Our system design*



# *The problem*

- Reusing metadata often requires translation
- Available translation options typically:
  - Support single use cases
  - Produce low-fidelity translations
  - Are cumbersome to maintain

100 ‡a Shakespeare,  
William, ‡d 1564-1616.ℙ  
245 ‡a Hamlet.ℙ  
260 ‡a New York :  
‡b Penguin Books,  
‡c c2003.ℙ

<dc:creator>Shakespeare,  
William,1564-1616</>  
<dc:title>Hamlet</>  
<dc:publisher>Penguin Books</>  
<dc:date>2003</>

## *What is a crosswalk?*

**“Crosswalks are used to ‘translate’ between different metadata element sets. The elements (or fields) in one metadata set are correlated with the elements of another metadata set that have the same or similar meanings. This is also sometimes called ‘semantic mapping.’”**

Source: Canadian Heritage Information Network  
(<http://www.chin.gc.ca/English/>)

# *An example: Dublin Core to Encoded Archival Description (EAD)*

Dublin Core to EAD1 - Microsoft Internet Explorer provided by OCLC, Inc.

**Dublin Core to EAD\***

Dublin Core	EAD <eadheader>	EAD <archdesc>
CONTENT		
Coverage		<geogname> (spatial) <unitdate> (temporal)
Description	<notestmt><note>	<abstract>
Type	***	<archdesc> with LEVEL attribute
Relation		
Source		
Subject	<notestmt><subject>	<controlaccess><subject>
Title	<titleproper>	<unittitle>
INTELLECTUAL PROPERTY		
Creator	<author>	<origination><persname> <origination><corpname> <origination><famname>
Contributor	<author>	<origination><persname>

# *Why use XSLT for crosswalks?*

- It capitalizes on current trends that model structured text in XML.
- It is a reasonable solution for lightweight processes and simple semantic mappings.

## **So, an XSLT repository would:**

- Reduce duplication of effort.
- Promote the use of standards.



# *A test client*

Which  
crosswalks  
are  
equivalent  
?

If they're not  
equivalent,  
how do they  
differ?

Which  
crosswalks  
have XML  
schemas that  
match my  
data?

<http://smithsonian.org/devon/marcRe>

( URI of Record )

( Overrides the Upload )

OR

Browse... Record (XML)

--//DOCUMENT XSLT MarcXML to DC 2002//EN  
--//DOCUMENT XSLT MarcXML to DC 2002//EN  
--//OCLC//DOCUMENT XSLT Dummy Transform 20021021//EN  
--//OCLC//DOCUMENT XSLT Marc21slim to Dublin Core Record Set 20020919//EN  
--//OCLC//DOCUMENT XSLT Marc21slim to RDF Dublin Core 20021122//EN

# A crosswalk pathfinder

The screenshot shows the top of the DLESE website. It has a blue header with the DLESE logo on the left. To the right of the logo are several orange buttons: 'Educational Resources', 'For Educators', 'News & Opportunities', 'People & Groups', 'For Developers', and 'About DLESE'. Below these buttons is a search bar with the text 'Search over: Educational resources' and a 'Search' button. On the left side, there is a vertical navigation menu with links: 'Metadata home', 'ADN framework', 'Annotation framework', 'Archived documents', 'Collection building', 'Collection framework', 'Contribute to DLESE', 'Controlled vocabularies', and 'Crosswalks'.

## Crosswalks

Describes the concept of crosswalks within DLESE.

### Crosswalks defined

A crosswalk is a semantic and/or technical mapping (sometimes both) of one another metadata framework.

The screenshot shows a table titled 'AGLS' which maps AGLS terms to DC-Elements and DC Definitions. The table has four columns: 'AGLS', 'Definition AGLS', 'DC-Element', and 'DC Definition'.

AGLS	Definition AGLS	DC-Element	DC Definition
Title	The name given to the record	DC.Title	A name given to the resource
Title Alternative	An alternative name by which the record is known	DC.Title Alternative	Any form of the title used as a substitute or alternative to the formal title of the resource.
Title Words	The words used to name the record; is the title.	DC.Title Alternative	Any form of the title used as a substitute or alternative to the formal title of the resource.
Scheme Type	The type of naming convention used to title records		
Scheme Name	The name of the specific internal/external standard, method or convention used to title the record.		
Agent	A corporate entity or organisational element which is responsible for some action on or usage of a record. An individual who performs some action on a record, or who uses a record in some way.	DC.Creator	An entity primarily responsible for making the content of the resource
	The name of an individual who performs		An entity primarily responsible for

The screenshot shows a PDF document titled 'Draft Standard for Learning Object Metadata'. The document is dated 15 July 2002 and is sponsored by the Learning Technology Standards Committee of the IEEE. It is copyrighted by the Institute of Electrical and Electronics Engineers, Inc. in 2002. The document is a draft of a proposed IEEE-SA Standard 1484.12.1. It includes a disclaimer: 'USE AT YOUR OWN RISK.' and a note that the document is a copy of the draft approved by the IEEE Review Committee on June 12, 2002. The document is 1 of 44 pages.

resource type

metadata records computer files.  
Automatically change extensible  
the following shows Dublin

to World</dc.title>  
</educational> GOES TO

12\_1\_v1\_Final\_Draft.pdf Annex B.

[s/index.htm](#)

## *Some problems*

- **In the XSLT collection**
  - Information needed for executing the scripts is missing.
  - Undocumented XSLT scripts aren't crosswalks, as we have defined them.
  - Syntax and semantics have been dissociated.
  - The collection can't be mined.
- **In the pathfinder page**
  - The documents vary in scope and granularity.
  - Executable code is often difficult to locate.
  - Pathfinders aren't designed for browsing and searching.

# *The crosswalk repository (3.0)*

## **Components of our solution:**

- Model the crosswalk as a complex object using the Library of Congress Metadata Encoding and Transmission (METS) standard.
- Assemble the records into a searchable repository built on Open Archives Initiative (OAI) standards.

## *A crosswalk as a METS record*

- Describe the crosswalk object in the METS header.
- Assemble and identify six objects in the METS structural map:
  - The source metadata schema
  - The target metadata schema
  - The crosswalk
  - Human-readable and executable versions of each
- Associate metadata for each file in the METS Descriptive Metadata Section.

# *OCLC's OAI repository*

**“The Open Archives Initiative develops and promotes interoperability standards to facilitate the efficient dissemination of content. Primary among these is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).”** [Source:oclc.org/research](http://oclc.org/research)

- Collects publicly accessible XML-encoded metadata on a range of research projects into a searchable collection, including **thesauri** and **Electronic Theses and Dissertations**.
- Includes XML tools required to manage the data, such as XSLT scripts.
- Serves as a testbed for research on information registries, URL maintenance, and searching.

# *A crosswalk METS record in the OAI repository*

The screenshot shows a web browser window with the OAI XML/XSD/XSL Registry interface. At the top, a navigation bar contains four steps: 1. Select a repository, 2. Filter sets and formats, 3. Browse list records, and 4. View record (highlighted in red). Below this, the page title is 'XML/XSD/XSL Registry'. A red button labeled 'ERRoLs' is visible. The main content area shows the 'Metadata Formats (Item):' section with a dropdown menu set to 'mets'. The 'Identifier:' field contains 'oai:xmlregistry.oclc.org:sts/lom2dc'. The 'Sets (Record):' dropdown menu is open, showing 'XSLStylesheets', 'Projects', and 'Projects:MetadataSwitch'. Below this, a table displays the record details for the 'Agent: INDIVIDUAL CREATOR'.

1 Select a repository 2 Filter sets and formats 3 Browse list records 4 View record

## XML/XSD/XSL Registry


ERRoLs


Metadata Formats (Item): mets

Identifier: oai:xmlregistry.oclc.org:sts/lom2dc

Sets (Record): XSLStylesheets  
Projects  
Projects:MetadataSwitch

Agent: INDIVIDUAL CREATOR	
name	Carol Jean Godby
note	This record documents a crosswalk between the 12-01-2001 version of IEEE LOM and unqualified Dublin Core, Version 1.1. It includes the following elements: 1) An XSL stylesheet; 2) XML schemas for the source and target documents; 3) human-readable descriptions of the source and target metadata schemes. The XML schema for LOM in this package is non-standard.





**Dublin Core Metadata Initiative®**

ABOUT THE INITIATIVE

DCMI NEWS

DOCUMENTS

TOOLS AND SOFTWARE

GROUPS

PROJECTS

RESOURCES

AskDCMI

Home > Documents > Dces >

Enter keyword

target

Application

dc:subject	DC, Dublin Core, u
dc:description	An XML encoding o

<http://dublincore.org/schemas/xmls/sin>

Reference

**Title:**

**Dublin Core Metadata Element Set, Version 1.1: Reference Description**

**Date Issued:** 2003-06-02

**Identifier:** <http://dublincore.org/documents/2003/06/02/dces/>

**Supersedes:** <http://dublincore.org/documents/2003/02/04/dces/>

**Latest version:** <http://dublincore.org/documents/dces/>

**Translations:** <http://dublincore.org/resources/translations/>

**Status of document:** This is a DCMI Recommendation.

**Description of document:** This document is the reference description, version 1.1 of the Dublin Core Metadata Element Set.

dc:subject

dc:title

dc:date

DC, Dublin Core

Dublin Core Metada

2003-06-02

**Introduction**

The Dublin Core metadata element set is a standard for cross-domain information resource description. Here an information resource is defined to be "anything that has identity". This is the definition used in Internet RFC 2396, "Uniform Resource

```

<?xml version="1.0" encoding="UTF-8" ?>
- <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:dc="http://purl.org/dc/elements/1.1/" elementFormDefault="qualified"
  attributeFormDefault="unqualified">
- <xs:annotation>
  <xs:documentation xml:lang="en">Simple DC container XML Schema Created 2003-04-02
    Created by Tim Cole (t-cole3@uiuc.edu) Tom Habing (thabing@uiuc.edu) Jane Hunter
    (jane@dstc.edu.au) Pete Johnston (p.johnston@ukoln.ac.uk), Carl Lagoze
    (lagoze@cs.cornell.edu) This schema declares a container element for a Simple DC
    application. The declaration of the simpledc element uses the dc:elementContainer
    complexType. Note that this schema does not define a target namespace. The expectation
    is that the simpledc element is assigned to a namespace by an application schema which
    includes this schema.</xs:documentation>
  </xs:annotation>
  <xs:import namespace="http://purl.org/dc/elements/1.1/" schemaLocation="dc.xsd" />
  <xs:element name="simpledc" type="dc:elementContainer" />
</xs:schema>

```



[http://ltsc.ieee.org/wq12/files/LOM\\_1484\\_12](http://ltsc.ieee.org/wq12/files/LOM_1484_12)

crosswalk

Application

dc:date

dc:creat

dc:cont

dc:subj

dc:desc

<http://err>

OAI Respo

IEEE 1484.12.1-2002  
15 July 2002

## Draft Standard for Learning Object Metadata

Sponsored by the  
Learning Technology Standards Committee  
of the IEEE

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
- <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns="http://www.openarchives.org/OAI/2.0/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:lom="http://www.imsglobal.org/xsd/imsmd_v1p2" exclude-result-prefixes="lom"
  version="1.0">
  <xsl:output doctype-public="//OCLC//DTD Dublin Core Record 20020906//EN" doctype-
    system="http://opaque.dev.oclc.org:8080/service/schema?id=-%2F%2FOCLC%2F%
    2FDTD+Dublin+Core+Record+20020906%2F%2FEN" indent="yes" method="xml" />
  - <xsl:template match="/lom:lom">
    - <dublincore>
      <xsl:apply-templates />
    </dublincore>
    </xsl:template>
  - <xsl:template match="lom:general/lom:identifier">
    - <identifier>
      <xsl:value-of select="." />
    </identifier>
    </xsl:template>
  - <xsl:template match="lom:general/lom:language">
    - <language>
      <xsl:value-of select="." />
    </language>
    </xsl:template>
  - <xsl:template match="lom:general/lom:title/lom:langstring">
    - <title>
      <xsl:value-of select="." />
    </title>
    </xsl:template>
  - <xsl:template match="lom:general/lom:description/lom:langstring">
    - <description>
      <xsl:value-of select="." />
    </description>
```

# *What the METS encoding solves*



- The semantic and syntactic information required for interpreting and executing a crosswalk is collected into a single object.
- The repository is searchable by humans and automated processes.
- Services can be built on top of it.
- It encourages the development and standardization of crosswalks.

**These outcomes are possible because every component in the system is a standard.**

# *Some possible services*

- Translations

- Queries

Which encodings have been done for a given metadata schema or namespace?

- Interactions with data

Given the XML schema and namespace referenced in my data, does this repository have any XSLT scripts that process it?

- Documentation

The METS crosswalk object can be associated with a given set of records to document which standards/versions/scripts were used to convert it.

# *Open issue 1*

*Are crosswalks a potential standard? Or just a local solution for the management of heterogeneous data?*



“The crosswalk is a preliminary one aimed at transforming relatively simple METS documents...”

**Source: Yee and Beaubien, 2003**

“Crosswalks that extend interoperability are essential so that the digital library collections can be accessible through a variety of portals and search interfaces. As more organizations share what they have learned...the development of crosswalks will be better understood and more easily accomplished.” **Source: Lightle and Ridgway 2003**

## *Open issue 2:*

### *Is XSLT the best tool for metadata translation?*

- XSLT is cumbersome when there is a need for high-fidelity translations.
- More precise associations between syntax and semantics may be necessary.
- The supporting documentation required for verifiable translation is daunting.

Metadata Schema X (versions \* encodings) \*

Metadata Schema Y (versions \* encodings)

# *Project status*

- The OCLC OAI repository is accessible at:  
<http://errol.oclc.org/xmlregistry.oclc.org.html>
- Advanced searching using the SRW (Search/Retrieve Web service) protocol is currently being implemented.
- The registry is being populated with crosswalk records.
- We welcome your comments and participation!

## *For further information*

- Metadata Schema Transformation Services  
[http://oclc.org/research/projects/mswitch/1\\_schematrans.htm](http://oclc.org/research/projects/mswitch/1_schematrans.htm)
- The Open Archives Initiative Project  
<http://oclc.org/research/projects/oai/default.htm>
- “Two Paths to Interoperable Metadata”  
<http://oclc.org/research/publications/archive/2003/godby-dc2003.pdf>

# References

- Raymond Yee and Rick Beaubien. 2003. “A Preliminary Crosswalk from METS to IMS Content Packaging.” *Library Hi Tech*.
- Kimberly Lightle and Judith S. Ridgway. 2003. “Generation of XML Records across Multiple Metadata Standards.” *DLIB*  
<http://dlib.org/dlib/september03/lightle/09lightle.html>