

Metadata Harvested through OAI-PMH

OCLC Western CONTENTdm Users Group Meeting

July 23, 2007

Reed College

Portland, Oregon

Amy Jackson, amyjacks@uiuc.edu

Myung-Ja Han, mhan3@uiuc.edu

Kurt Groetsch

Timothy W. Cole, t-cole3@uiuc.edu

University of Illinois at Urbana Champaign



IMLS Digital Collections and Content

- Project began December 2002 as an IMLS National Leadership Grant
 - Tim Cole, Principal Investigator
 - Amy Jackson, Project Coordinator

- Collaboration with UIUC Library and Graduate School of Library and Information Science

- <http://imlsdcc.grainger.uiuc.edu/>

IMLS Digital Collections and Content

- Project Objectives:
 - Implement a collection registry of digital collections created or developed with funding from IMLS NLG program
 - Use OAI-PMH to implement an item-level metadata repository for items contained in NLG collections
 - Carry out associated research related to:
 - Utility and usability of Registry & Repository
 - Current metadata practices of IMLS NLG grantees
 - Implications for interoperability (Framework of Guidance for Building Good Digital Collections)

Collection registry

- Collection registry
 - 180 NLG projects
 - 15 LSTA projects

Images 80%

Text 68%

Physical Object 29%

Sound 20%

Interactive Resource 10%

Collection Registry

- Top GEM subjects in Collection Registry
 - Social Studies 80%
 - United States history
 - State history
 - Arts 46%
 - Visual arts
 - Photography
 - Science 17%

Item-level repository

- Item-level Repository
 - Harvesting 71 of 195 Collections (36%)
 - 37 Repositories (some multiple institutions)
 - 10 ContentDM repositories
 - 310,448 records

- Item Records (self identified types)
 - 86% images
 - 14% text

Item-level repository

Top Item-level subjects

United States

People

Songs with piano

Trees

Tennessee Valley Authority

Archaeology Southern States

Works Progress Administration

Cities & towns

Women

Archaeology

Buildings

Photographers

Mountains

Men

Archaeological site

Insect

Bodies of water



Item-level repository

Number of harvested collections using each DC field

Field	Number of Collections	% of Collections
Title	35	100%
Identifier	35	100%
Subject	33	94%
Type	32	91%
Creator	32	91%
Description	31	89%
Date	30	86%
Publisher	30	86%
Format	28	80%
Rights	27	77%
Language	26	74%
Relation	23	66%
Contributor	21	60%
Source	20	57%
Coverage	18	51%

OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Low barrier protocol to enable the interoperability of digital libraries.
 - Data providers
 - Service providers (aggregators)
- Requires that metadata be exposed in Dublin Core
- Other formats are optional

5 of the 37 repositories export in schemas other than simple or qualified Dublin Core. Other schemas include MARC21, MODS, OLAC, and ETDMS.
- ContentDM supports export in simple and qualified Dublin Core.

- Barriers to sharing metadata through OAI-PMH
 - Technical Infrastructure
 - Metadata
 - Institution/Project

- ContentDM
 - Compliant with OAI-PMH
 - Metadata is mapped to DC

IMLS Digital Collections and Content

- Aggregated environment
 - Local ContentDM server
 - ContentDM multi-site server
 - Other service providers (IMLS DCC, OAIster)

Harvested Metadata

- How has use of Dublin Core changed over time?
Records harvested from January 1, 2001 and December 31, 2006.
 - Quantitative analysis
 - What measurable changes can we see in the metadata?
 - Qualitative analysis
 - How has use of fields changed over time?



Quantitative analysis

- Quantitative analysis
 - Repetition of elements
 - Length of fields
 - Use of core fields (Shreeves et al. (2005))

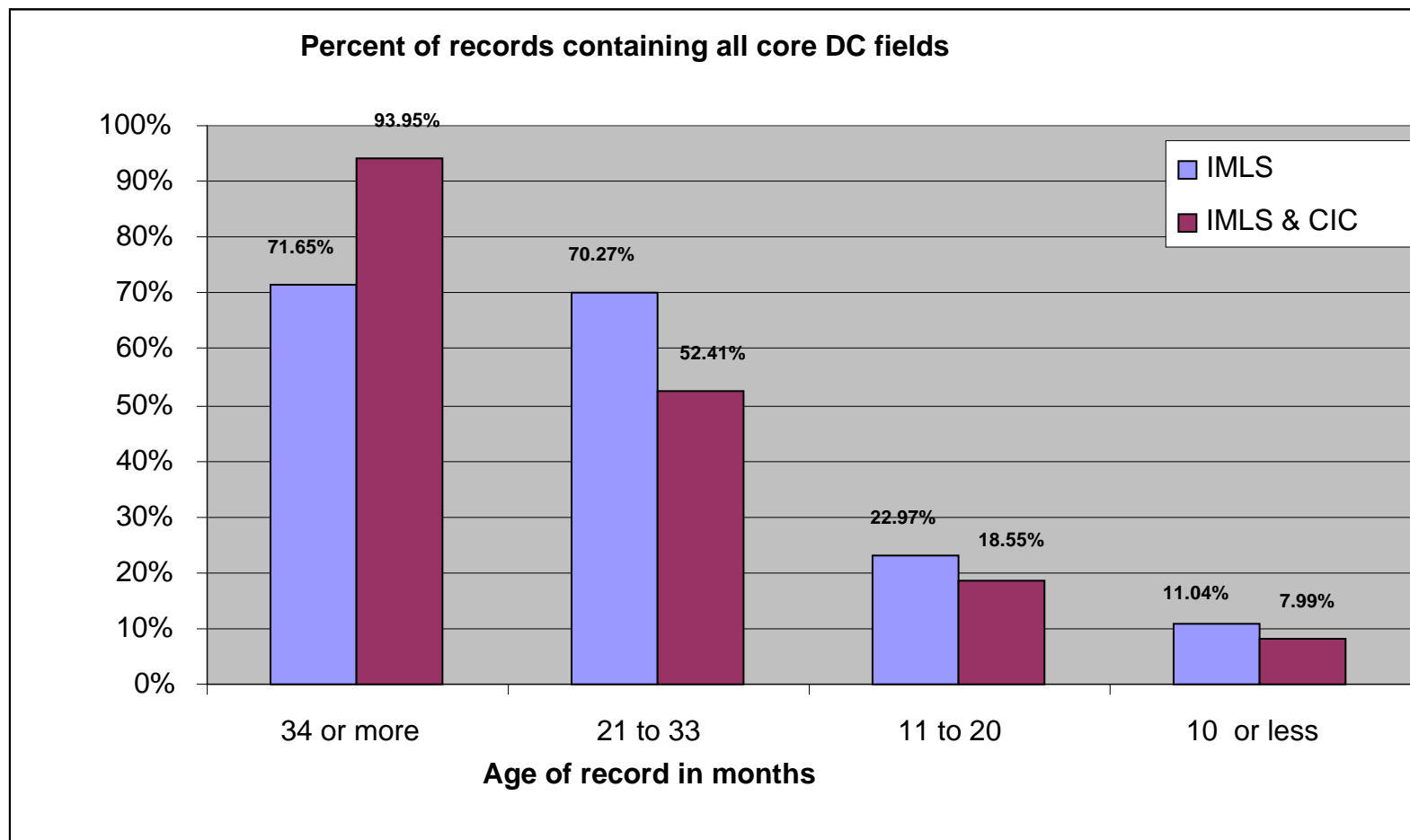
Title	Creator
Subject	Description
Date	Format
Identifier	Rights



Quantitative analysis

- Repetition of fields
 - Stable
- Length of fields
 - Stable
- Use of all 8 core fields
 - Declining

Quantitative Analysis



Quantitative Analysis

- Of these eight elements, the two elements most often missing are creator (used in 39% of records) and rights (52%).
- Identifier, title, and subject were each used in over 96% of all records.
- Format and description fields have shown the most significant decline in use since 2003.
- Decreased repetition and length of the description field, and an overall increase in use of the relation field.

Percent of Records Containing each DC field

DC field	ContentDM repositories	All repositories
Subject	93%	65%
Title	100%	96%
Contributor	5%	9%
Coverage	60%	22%
Dates	95%	65%
Description	92%	65%
Publisher	30%	77%
Relation	12%	63%
Type	98%	94%
Creator	36%	39%
Format	94%	61%
Source	88%	42%
Rights	42%	52%
Identifier	100%	100%
Language	80%	41%

Qualitative analysis

- Qualitative analysis
 - 225 records from 6 average repositories
 - Document changes in practice over time
 - 90 records from ContentDM repositories
 - 600 randomly selected records
- Only 1 observed change in practice over time
 - Early records:
<title>Frankie / Music by Neil Sedaka; words by Howard Greenfield </title>
 - Later records:
<title>Frankie</title>
<creator>Music by Neil Sedaka; words by Howard Greenfield</creator>

Qualitative analysis

- Unpacking MARC tags to Dublin Core elements
 - Merged publisher and date fields from MARC 260
 - Title field containing creator or contributor information from MARC 245
 - Confusion of type and format fields from MARC 300
 - Incorrect use of delimiter
 - <creator>Pika,</creator>
 - <creator>I.,</creator>
 - <creator>et al.,</creator>

Qualitative analysis

- Misuse of Dublin Core elements
 - Date and coverage fields
 - Source and relation fields
 - Format and description fields
 - Format

Qualitative analysis

- Misuse of Dublin Core Elements
 - Date and coverage fields

Item about the nineteenth century, published in 2007.

Correct: `<date>2007</date>`

`<coverage>1800-1899</coverage>`

Incorrect: `<date>1800-1899</date>`

Qualitative analysis

- Misuse of Dublin Core Elements
 - Source and Relation fields

Source: "The resource from which the described resource is derived."

Relation: "A related resource."

(From: *Dublin Core Metadata Element Set, Version 1.1*, <http://dublincore.org/documents/dces/>)

Example: Series title belongs in relation not source.

Qualitative analysis

- Misuse of Dublin Core elements
 - Format and Description fields

Example:

Notes Material: Whale Bone (MARC 500)

Incorrect: <description>Material: Whale Bone</description>
 <description>9 in. x 6 in.</description>

Correct: <format>Material: Whale Bone; 9 in. x 6 in.</format>

Qualitative analysis

- Misuse of Dublin Core fields
 - Format “The file format, physical medium, or dimensions of the resource.” (From: *Dublin Core Metadata Element Set, Version 1.1*, <http://dublincore.org/documents/dces/>)

Incorrect:

`<format>Available via the World Wide Web</format>`

`<format>web browser</format>`

`<format>Any machine capable of running graphical Web browsers, 640x480 minimum monitor resolution</format>`

Correct:

`<format>image/jpeg</format>`

Qualitative analysis

- Confusion in Descriptive metadata and Administrative metadata (too much information exposed)
 - Administrative metadata is not helpful for discovery in the aggregated environment.
 - software used for digitization, master file format, storage equipment
 - Exposed metadata as one “view” of all associated metadata.

Qualitative analysis

- Lost information (not enough information exposed)
 - Additional metadata fields helpful for discovery could be mapped to Dublin Core and exposed.
 - ContentDM allows local fields to be mapped or not mapped to Dublin Core fields for exporting through OAI-PMH; be sure that helpful information isn't being hidden from the service provider.

Current collection: American Library Association Archives Digital Collections

[change](#)

Field properties

View, add, edit and delete fields. Enable full text searching and controlled vocabulary. Once you have added, changed, or deleted fields, index

	Field name	DC map	Data type	Large	Search	Hide	Vocab		add field
1	Title	Title	Text	No	Yes	No	Yes	move to ▼	edit delete
2	Translation Title	Title	Text	No	No	No	No	move to ▼	edit delete
3	Alternative Title	Title	Text	No	No	No	No	move to ▼	edit delete
4	Type	Type	Text	No	Yes	No	Yes	move to ▼	edit delete
5	Digitized Material	Description	Text	No	Yes	No	Yes	move to ▼	edit delete
6	Year-Coverage	Coverage-Temporal	Text	No	Yes	No	No	move to ▼	edit delete
7	PeriodName-Coverage	Coverage-Temporal	Text	No	No	No	Yes	move to ▼	edit delete
8	Geographic-Coverage	Coverage-Spatial	Text	No	Yes	No	Yes	move to ▼	edit delete
9	Names-Subject	Subject	Text	Yes	Yes	No	Yes	move to ▼	edit delete
10	Subject	Subject	Text	Yes	Yes	No	Yes	move to ▼	edit delete
11	Description	Description	Text	Yes	Yes	No	No	move to ▼	edit delete
12	Creator Personal	Creator	Text	No	No	No	Yes	move to ▼	edit delete

Conclusions

- Conclusions
 - Native metadata records are rich in meaning in their own environment, but lose richness in the aggregated environment due to mapping errors and misunderstanding and misuse of Dublin Core fields.
 - Mapping is often based on semantic meanings of metadata fields rather than value strings.
 - Correct mapping could improve metadata quality significantly.



Recommendations

- Recommendations
 - Publish local metadata practices
 - Publish crosswalking information
 - Expose native metadata in addition to Dublin Core
 - Ensure metadata creators receive appropriate training



IMLS Digital Collections and Content

Reference: Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., & Cole, T.W. "Is 'quality' metadata 'shareable' metadata? The implications of local metadata practice on federated collections." In H.A. Thompson (Ed.) *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*, April 7-10 2005, Minneapolis, MN . Chicago, IL: Association of College and Research Libraries. p.223-237.

Amy Jackson

Project Coordinator, IMLS Digital Collections and Content

University of Illinois at Urbana Champaign

amyjacks@uiuc.edu