

ICDAT 2005

Metadata aggregation for digital libraries

Muriel Foulonneau (mfoulonn@uiuc.edu)

Grainger Engineering Library
University of Illinois at Urbana-Champaign
USA

June 2005





Outlines

- Role and practices of actors in OAI-PMH
 - Data provider role and practices
 - Service provider role and practices
 - Digital Library Federation / National Science Digital Library best practices for OAI and shareable metadata

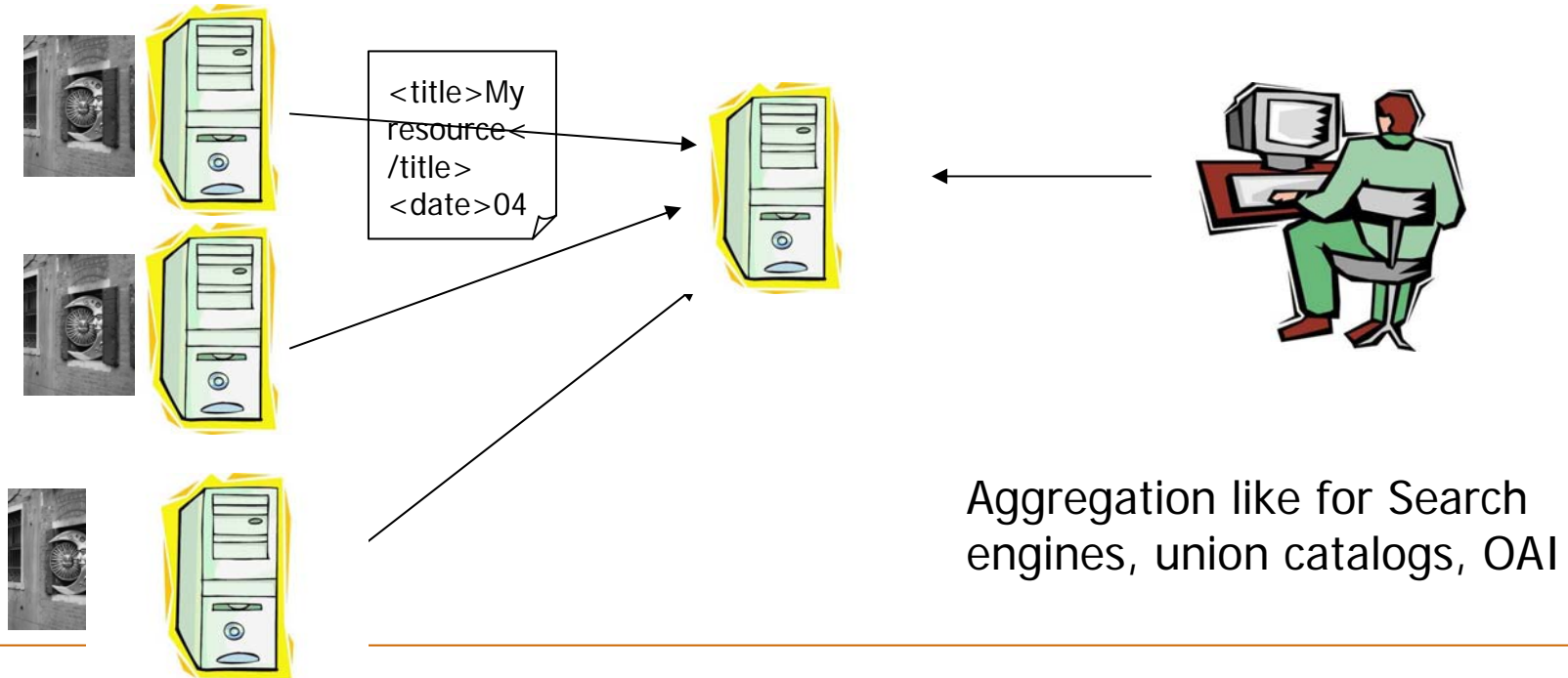
- Current evolutions
 - Metadata aggregation and resource discovery
 - Collection description
 - A distributed collection of resources

- Orientations for the future



Metadata harvesting

- Divides the worlds between data providers, service providers and intermediary aggregators
- Asynchronous model





Data providers





The role of a Data Providers

- Repository
 - Independent / Semi-independent / Integrated to the local system
- Update repository
- Make shareable metadata
 - Richest possible
 - The most adapted to SP usage





Review of existing implementations

- ½ home-grown
- ½ turn key systems and modules
 - Most popular ones
 - OAICat, VTOAI, DLESE, CWIS, ContentDM

http://nsdl.comm.nsdl.org/meeting/session_docs/2004/2620_National_Science_Digital_Library_Conference.doc

And

<http://oai-best.comm.nsdl.org/>





Review of optional OAI features

- Optional features of packages
 - OAI sets
 - Multiple metadata formats
 - Resumption tokens and record counts
 - Persistent deleted records
 - Granularity of datestamp



Parameters that NSDL data providers use

.OAI sets	<50% between 1 and 175
.Multiple metadata formats	>50% 15 different schemas, 6 non standard ones
.Resumption token and records count	85%
.Persistent deleted records	20% persistent, 20% transient
.Granularity of datestamp	25% second based





Service providers



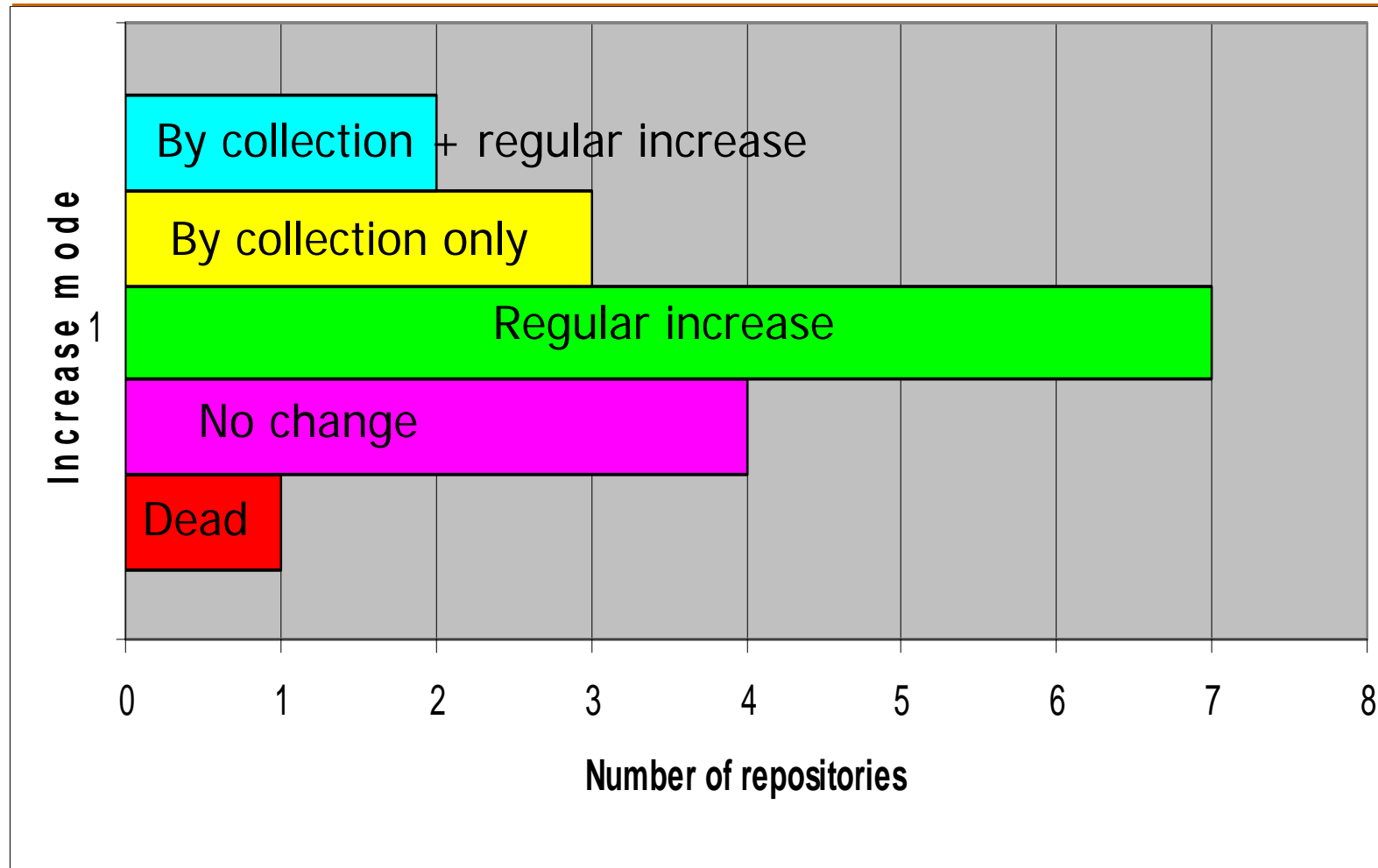


Service Provider role

- Support for building data providers
 - (coordinated / uncoordinated repositories)
- Harvesting
- Metadata normalization and enhancement
 - Semantics
 - Syntax
 - Traceability, provenance ...
- Data publication
 - ➔ Building application on top of a centralized collection of metadata records and a distributed collection of resources
 - ➔ Building of an aggregated collection through development and selection of information sources



Harvesting in practice





Normalization

Attribute	present	after	% of records in the repository
Type			99%
Format			66%
Language			60%
Collection			100%
Resource URL			72%

Also Where : 31% and When : (80%)



SP as a metadata publisher

- Identifying data sources
- Sending user to expected locations
- Representing the resources
 - Homogeneous categories
 - Text
 - Images
 - Video
 - Etc ...
 - Homogeneous codes





Representing the resources

Address <http://cic harvest.granger.uiuc.edu/spatial/index.asp> Go Google

CIC Resources Online Project : metadata portal

the most popular places US Middle East Africa Europe Asia Central America South America

Which African State are you interested in?
Notice : The area covered by the resources have been automatically interpreted; this may lead to discrepancies

The CIC librarian is searching for you...

Country : Sudan 166 results

page 1 / 9 Go to page 1 2 3 4 5 6 7 8 9 Next page >>

title	American Library Association Archives: Executive Board and Executive Director: Executive Director: Subject File
type	The materials described in this descriptive record are archives (official records of an organization).
Collection	American Library association finding-aid - University of Illinois at Urbana-Champaign
	More info View
title	Fine Arts and Applied Arts: Urban and Regional Planning Department: Scott Keyes Papers
type	The materials described in this descriptive record are manuscripts (personal papers of an individual).
Collection	University of Illinois at Urbana-Champaign archives finding-aid - University of Illinois at Urbana-Champaign
	More info View
title	Sugar Plantation in Gunaid, Blue Nile Province
Collection	Africa Focus - Sights and Sounds of a continent - University of Wisconsin-Madison
	More info View
title	Peanut Cultivation in the Jazira Area of the Nile Valley
Collection	Africa Focus - Sights and Sounds of a continent - University of Wisconsin-Madison
	More info View
title	Young Girl in Field of Long Staple Cotton in Blue Nile Province
Collection	Africa Focus - Sights and Sounds of a continent - University of Wisconsin-Madison
	More info View
title	Harvesting from a Date Palm
Collection	Africa Focus - Sights and Sounds of a continent - University of Wisconsin-Madison
	More info View
title	Aerial View of Desert Development Project in the Gezira
creator	Lewis, Herbert S. (Photography)

Website maintained by the University of Illinois at Urbana-Champaign
 In partnership with CIC institutions
 For information and comments, please contact [Muriel Foulonneau](#)

Done Internet

2005



Best practices for OAI and shareable metadata





Best practices - OAI and shareable metadata

- Launched by the Digital Library Federation and the National Science Digital Library in 2004
- Associates service providers
 - NSDL, Metascholar, American South, American West, CIC metadata portal, Aquifer, IMLS registry ...
- And data providers
 - Indiana University, Library of Congress, University of Tennessee, Princeton ...

<http://oai-best.comm.nsdlib.org>





Practices and communication DP/SP

- Which are the problems?
 - Turn-key systems compliance
 - Local implementations
 - Metadata issues

- What is the impact of implementing OAI features
 - Granularity of datestamp, Sets, Multiple metadata formats ... etc

- What is the impact of metadata shareability issues
 - What if I encode erroneously a date?





Metadata aggregation and resource discovery





SP role in improving resource discovery

- Collection registries
 - IMLS, Aquifer ...

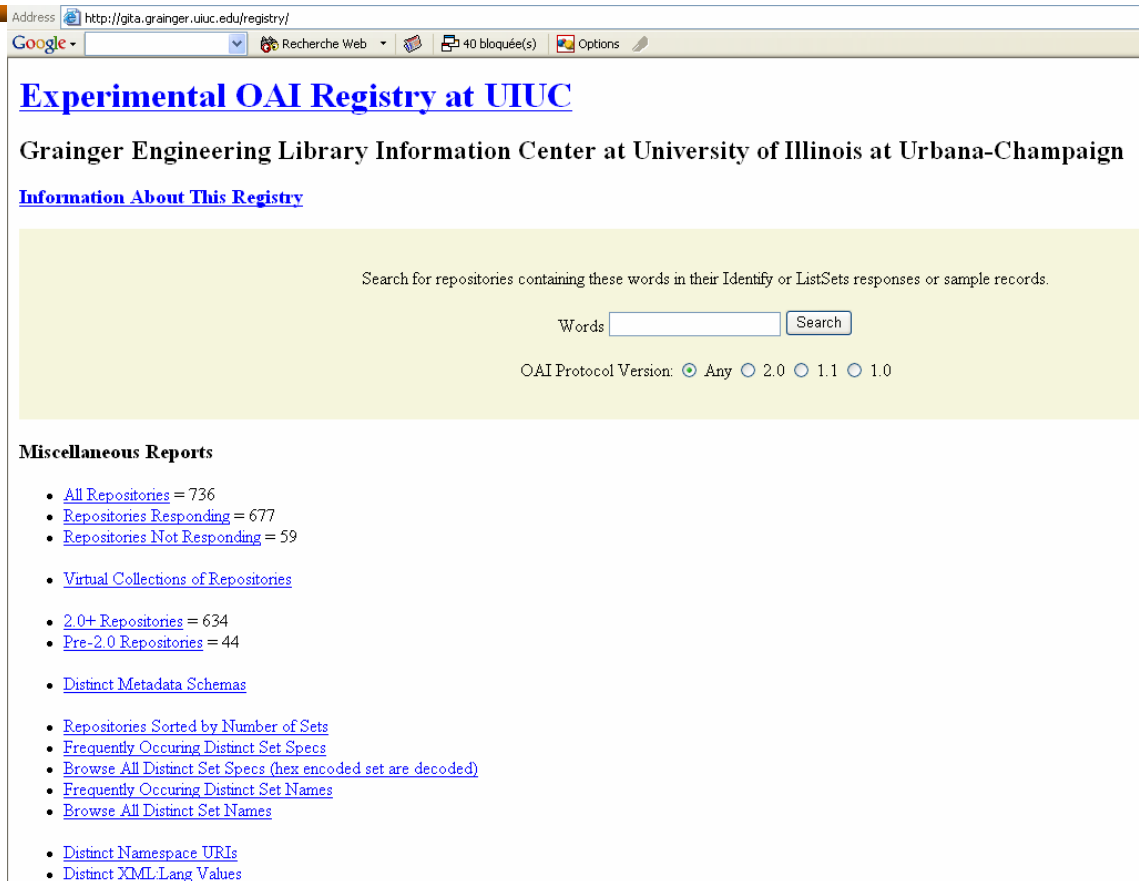
- Service registries
 - OKHAM, IESR

- Better discoverability of resources through traditional search engines
 - Yahoo! And Google agreements with OAIster
 - ePrints / IR collections in Google Scholar





Improving discoverability of OAI services



Address <http://gita.grainger.uiuc.edu/registry/>

Google Recherche Web 40 bloquée(s) Options

Experimental OAI Registry at UIUC

Grainger Engineering Library Information Center at University of Illinois at Urbana-Champaign

[Information About This Registry](#)

Search for repositories containing these words in their Identify or ListSets responses or sample records.

Words

OAI Protocol Version: Any 2.0 1.1 1.0

Miscellaneous Reports

- [All Repositories](#) = 736
- [Repositories Responding](#) = 677
- [Repositories Not Responding](#) = 59
- [Virtual Collections of Repositories](#)
- [2.0+ Repositories](#) = 634
- [Pre-2.0 Repositories](#) = 44
- [Distinct Metadata Schemas](#)
- [Repositories Sorted by Number of Sets](#)
- [Frequently Occuring Distinct Set Specs](#)
- [Browse All Distinct Set Specs \(hex encoded set are decoded\)](#)
- [Frequently Occuring Distinct Set Names](#)
- [Browse All Distinct Set Names](#)
- [Distinct Namespace URIs](#)
- [Distinct XML:Lang Values](#)

<http://gita.grainger.uiuc.edu/registry/>



Improving discoverability of collections



The screenshot shows a web browser displaying the IMLS Digital Collections Registry. The page title is "IMLS Digital Collections Registry" with the subtitle "A gateway to digital collections funded through the IMLS National Leadership Grant Program". The browser address bar shows the URL: <http://imlsdcc.grainger.uiuc.edu/collections/GEMTopPlusSubs.asp>. The page features a navigation menu with links for Home, About, and Contact. The main content area is titled "Browse Collections by Subject" and lists various subject categories with their respective collection counts and sub-topics. On the left side, there is a "Browse Collections by:" section with links for Subject, Object, Place, and Title, and a "Search Collections" box with a search input field and a "Go" button. Below the search box is an "Advanced Search" link.

Browse Collections by Subject
(Collections may be listed under more than one subject)

Arts (65 Collections)
[Architecture](#) (8), [Computers in art](#) (3), [Dance](#) (2), [Drama/dramatics](#) (4), [Film](#) (3), [History of art](#) (6), [Informal education in art](#) (2), [Music](#) (9), [Photography](#) (28), [Popular culture](#) (8), [Technology in art](#) (3), [Theater arts](#) (5), [Visual arts](#) (33)

Educational Technology (9 Collections)
[Educational media](#) (3), [Informal education in educational technology](#) (3), [Language laboratories](#) (1), [Multimedia education](#) (6), [Technology planning](#) (1)

Foreign Languages (3 Collections)
[History of foreign languages](#) (1), [Linguistics](#) (2)

Health (6 Collections)
[Aging](#) (2), [Chronic conditions](#) (1), [Consumer health](#) (2), [Death and dying](#) (3), [Disease](#) (1), [Environmental health](#) (3), [Family life](#) (1), [History of health](#) (2), [Human sexuality](#) (1), [Informal education in health](#) (1), [Mental/emotional health](#) (3), [Nutrition](#) (5), [Safety](#) (2), [Smoking](#) (2), [Substance abuse prevention](#) (3), [Technology in health](#) (2)

Language Arts (12 Collections)
[Alphabet](#) (2), [Grammar](#) (1), [Listening comprehension](#) (1), [Literature](#) (9), [Reading](#) (1), [Story telling](#) (2), [Writing \(composition\)](#) (1)

Mathematics (3 Collections)
[Algebra](#) (1), [Calculus](#) (1), [Geometry](#) (1), [Number theory](#) (1), [Patterns](#) (1), [Statistics](#) (1), [Trigonometry](#) (1)

Search Collections

Go
[Advanced Search](#)

<http://imlsdcc.grainger.uiuc.edu/collections/index.htm>

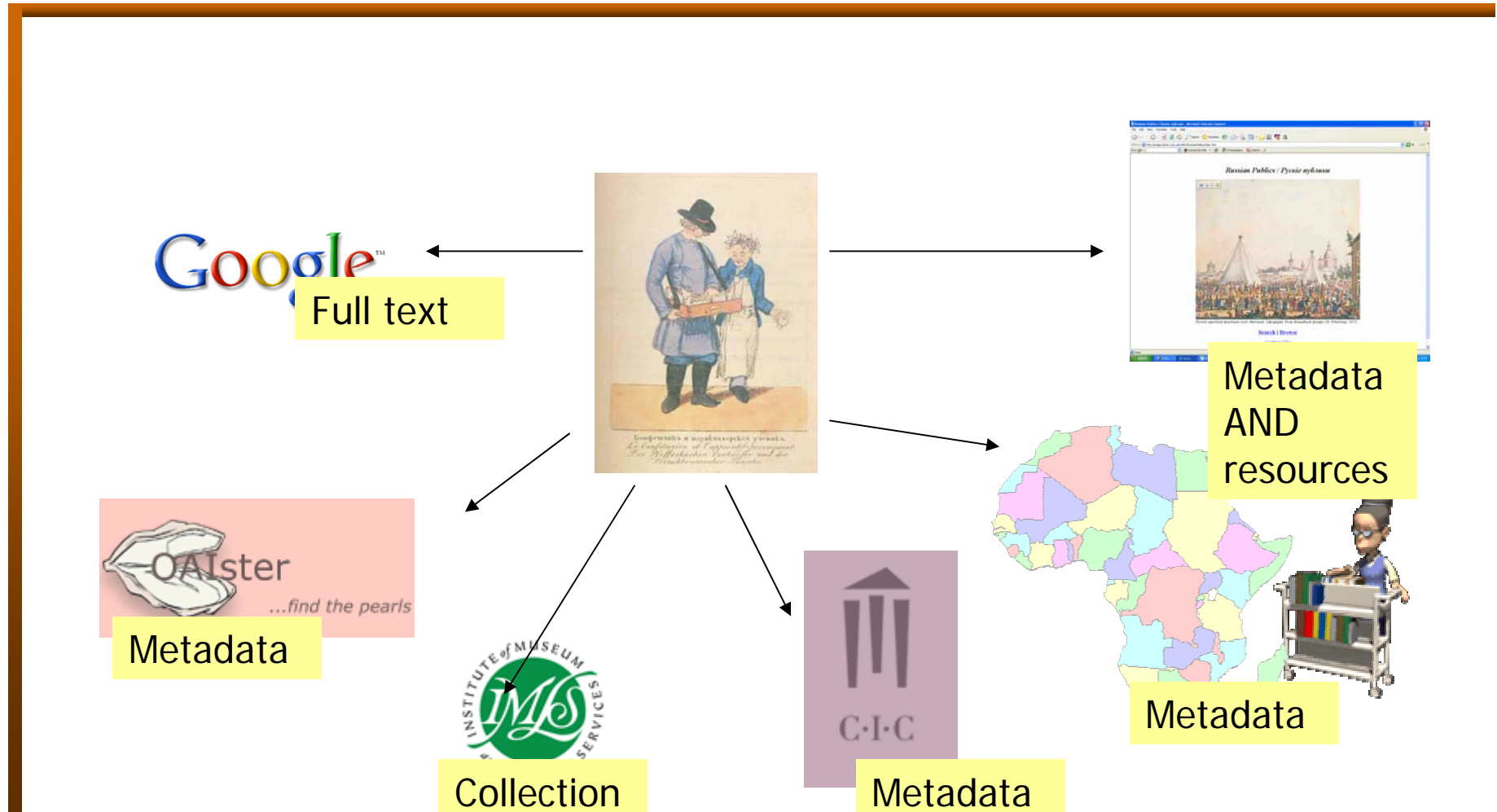


Performance and usability of the data

- Quality issue
 - Particularly completeness and consistency
 - Context?
 - On a horse problem
- Which metadata content for which usage?
 - Human friendly
 - Search engine friendly
 - Machine friendly
 - Codes and references
- Usage models
 - Find, Select, Identify, Obtain, Interpret, Co-locate



Multiple services use different features



Information puzzle

Resources



Metadata

Collections

**Interpretation /
relation services**





Collection descriptions





The information landscape

- Granularity of descriptions vs granularity of user expectations
- Information implied in local context and missing when aggregated

Record 2 of 8

Title	[Food for criticks, in] The Pennsylvania gazette. Containing the freshest Advices Foreign and Domestick. Numb. 190. From Monday, July 10. to Monday, July 17. 1732
Author/Creator	Richard Lewis, c.1700-1734
Publisher	Printed by Hall and Sellers
Type	text
URL	http://name.umd.umich.edu/AM1315
Collection	American Poetry Database





Filtering by collection

CIC-OAI Metadata Search Portal

Home
Search
View Institutions
Help

Your search was in **all fields** for **"russian"**. The search was limited to the **"University of Illinois Library DIMTI"** archive.

You found **83 records**.
 * [Revise your search](#) to retrieve fewer records.
 * View your results, starting with **records 1 to 25 of 83**.

Results by Institution

Indiana University:
Indiana University Digital Library Program
[90 records](#)

Michigan State University:
Michigan State University Digital and Multimedia Center
[52 records](#)

Ohio State University:
Ohio Link Electronic thesis and dissertations
[10 records](#)

Pennsylvania State University:
Pennstate libraries Image Collections
[3 records](#)
Pennstate thesis and dissertations
[2 records](#)

University of Chicago:
University of Chicago Library
[4 records](#)


University of Illinois at Urbana-Champaign:
Engineering Documents Center Collection
[1 record](#)
University of Illinois Finding Aids
[49 records](#)
University of Illinois Sheet Music
[24 records](#)
University of Illinois Library DIMTI
[83 records](#)

Sort by title


[Next 10 Records](#)

Jump to Records: [1](#) | [26](#) | [51](#) | [76](#)

Record 1 of 83

Title	
Author/Creator	(unknown)
Description	(unknown)
URL	http://images.library.uiuc.edu:8081/RussianPublics/image/97752201452003_fonar60-5.jpg
Collection	Early 19th Century Russian Readership & Culture

Record 2 of 83

Title	
Author/Creator	(unknown)
Description	(unknown)






Searching collections

Address <http://cic harvest.grainger.uiuc.edu/colls/collections.asp> Go Links >>






Google Recherche Web 40 bloquée(s) Options



CIC Metadata Portal

Access to digital library resources of
major midwestern universities
(CIC -- Committee on Institutional Cooperation)

Current CIC collections in the metadata portal

Indiana University	
<p><u>Wright American Fiction Project</u></p>	<p>2839 items</p> 
<p><u>IU Bio-Archive of biology data and software</u></p> <p>This is an archive of biology data and software, established in 1989 to promote public access to freely available information, primarily in the field of molecular biology. This archive is maintained by the Indiana University Biology department.</p>	<p>504 items</p> 
<p><u>U.S. Steel Gary Works Photograph Collection, 1906-1971</u></p>	<p>2227 items</p> 
<p><u>Victorian Women Writers Project</u></p>	<p>169 items</p> 
<p><u>Charles W. Cushman Photograph Collection</u></p>	<p>14425 items</p> 

Menu

[Home](#)

Search the metadata portal

- ▶ [CIC keyword search](#)
- ▶ [CIC geographic browse](#)
- ▶ [CIC collection browse](#)

Project details

- ▶ [Project presentations & reports](#)
- ▶ [CIC collection development](#)
- ▶ [CIC Metadata guidelines](#)
- ▶ [Schemas and crosswalks](#)
- ▶ [Metadata aggregation](#)
- ▶ [OAI Content Providers](#)
- ▶ [Institutions participation](#)
- ▶ [The DLXS software](#)
- ▶ [Resources](#)

Contacts

[Partner section](#)

The DC Collection standard

Collection Working Group

> InstanceTemplate

Title of Collection

Collection Description

Resource URI	http://resourceURI/
Property QName	Value String
dc:identifier	value string
dc:title	value string
dcterms:alternative	value string
dcterms:abstract	value string
dc:format	value string
dcterms:extent	value string
dc:language	value string (term)
dc:type	value string (term)
dc:type	value string (term)
dc:rights	value string
dcterms:accessRights	value string





Current work to describe collections

- IMLS collection registry
<http://imlsdcc.grainger.uiuc.edu/registry/default.asp>
- Digital collections of Canada <http://collections.ic.gc.ca/>
- TEL <http://www.theeuropeanlibrary.org>
- Minerva / Michael <http://www.minervaeurope.org/>
<http://www.michael-culture.org/>
- Catalogue des fonds numerises
http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm
- Enrich UK <http://www.enrichuk.net/>



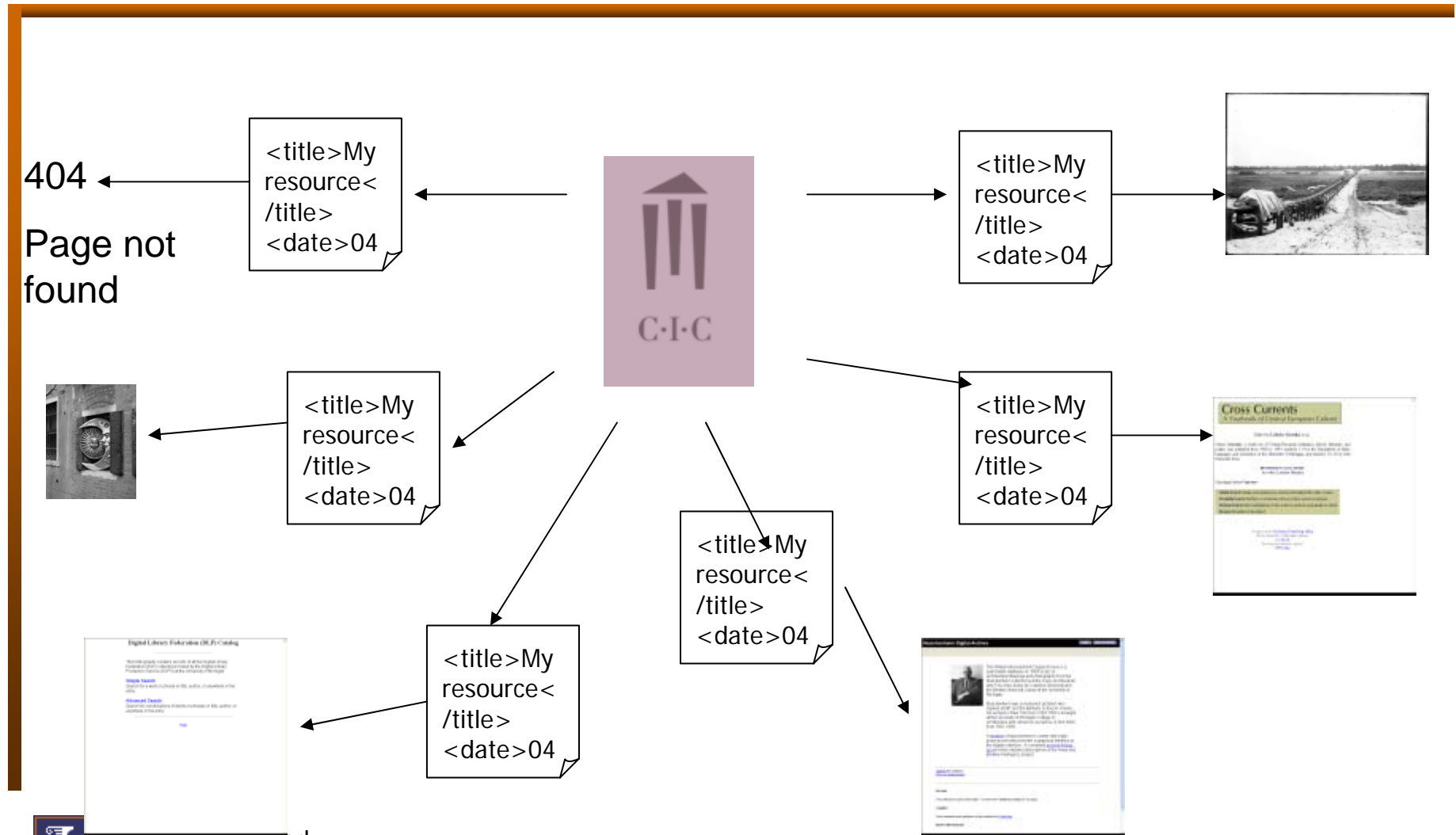


A distributed collection of resources





Behind the metadata



Adding the resources



- A Webpage structure contains information
 - Links density
 - Images density

Adding the full text of resources

- ePrints project
- NSDL, Digital Collections of Canada ...
 - A Webpage textual content
 - Metadata
 - Description, interpretation
 - Complementary resources





Metadata generation and topic classification

- MetaCombine project
 - application of semantic clustering to digital libraries
- American West
- NSDL – Riverside work for metadata generation
- IMLS: Manual clustering at collection level





Orientalions





Data and service providers collaboration

- DP implemented with more and more features
 - In 1 year CIC data providers changed a lot
 - Set descriptions
 - Thumbnails
 - Multiple metadata formats
 - More OAI DP
 - Better quality of metadata





SP struggling with heterogeneity of DP

- Community building
 - Best practices for DLF libraries : MODS as a major metadata formatetc size of institution, competences etc
- Training material
 - OAForum tutorial a first step <http://www.oaforum.org/>
 - A lot more to come : book, Aquifer, OAI best practices Website / handbook ...
- Metadata quality services [?]





Extensions

- Sharing resources and non descriptive metadata
 - MOD_OAI
 - Resource harvesting

- Coupling with other technologies
 - SRU/SRW (eg. OAI to Z39.50 and SRW packages)
 - Spidering resources





Experiences at UIUC – Grainger library

- **Metadata reprocessing**
 - Metadata quality
 - Reprocessing strategies
 - Assessing the relative importance of information in metadata records
 - Multiplicity of metadata formats : Harvesting in MODS, MARC ...

- **Collection description**
 - Automating OAI sets processing
 - Using collection descriptions to improve data discoverability
 - Building relevant collection descriptions

- **Resources**
 - Analyzing collections of distributed resources
 - Agreement for thumbnails generation





More experiences at UIUC – Grainger library

- **Usability testing** : students, teachers, professionals in libraries and museums
 - Specific expectations for aggregated collections
 - Under which conditions is this useful?
- **Discovery services**
 - The UIUC OAI providers registry
 - SRU access to OAI repositories
- **Tools** <http://sourceforge.net/projects/uilib-oai/>
 - OAI harvester and repositories
 - Thumbnails generation tool





Several references

- The CIC metadata portal <http://cicharvest.grainger.uiuc.edu/>
- IMLS registry of digital collections
<http://imlsdcc.grainger.uiuc.edu/collections/>
- UIUC registry of OAI repositories
<http://gita.grainger.uiuc.edu/registry/>
- The DLF / NSDL best practices for OAI and shareable metadata <http://oai-best.comm.nsd.org/>
- The Open Archives Initiative <http://www.openarchives.org/>

