# Metadata Standards and Applications

## 8. Metadata Interoperability and Quality Issues

# Goals of Session

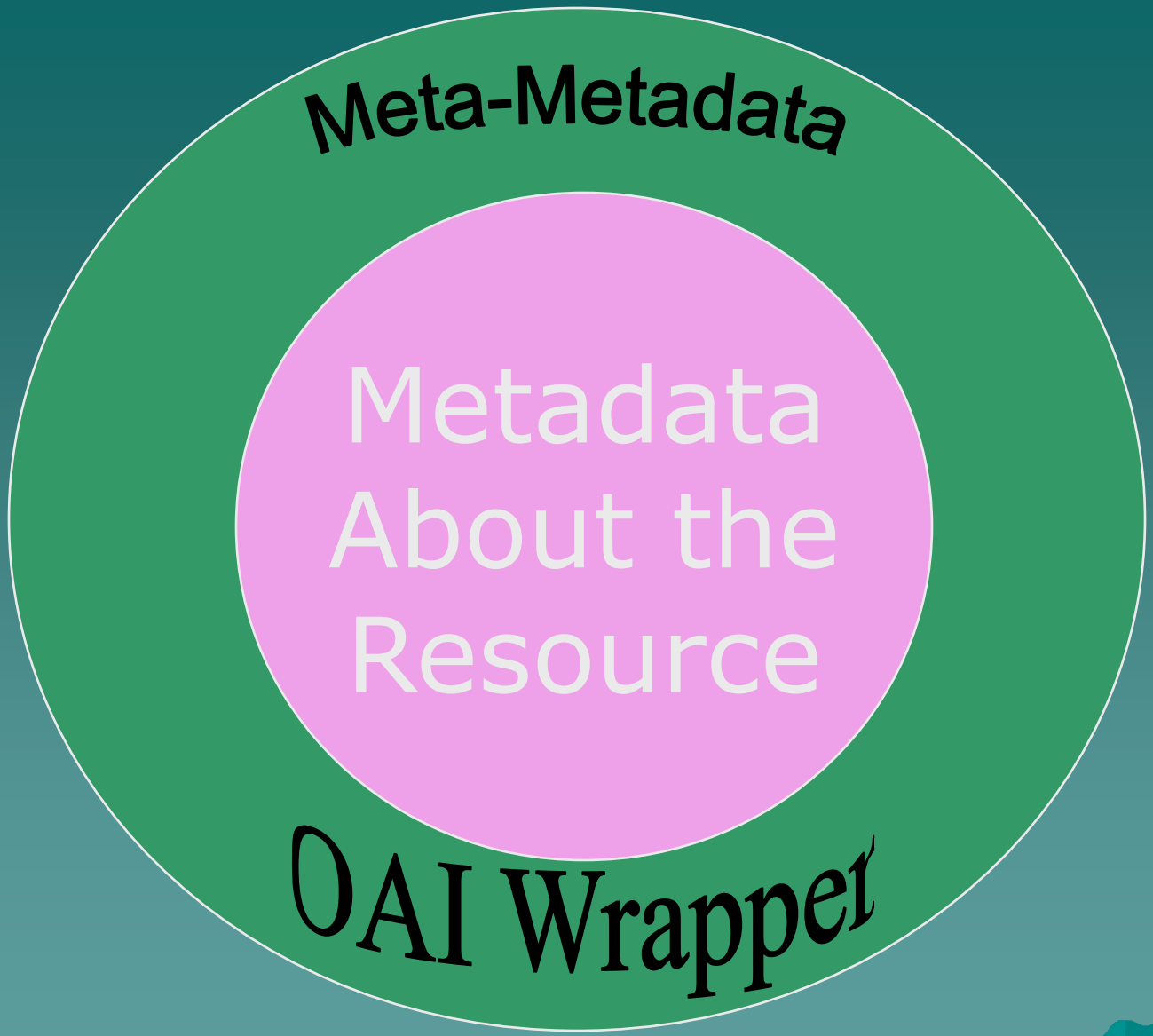
- Understand interoperability protocols (OpenURL for reference, OAI-PMH for metadata sharing)
- Understand crosswalking and mapping as it relates to interoperability
- Investigate issues concerning metadata quality

# What's the Point About Interoperability?

- For users, it's about resource discovery (user tasks)
  - What's out there?
  - Is it what I need for my task?
  - Can I use it?
- For resource creators, it's about distribution and marketing
  - How can I increase the number of people who find my resources easily?
  - How can I justify the funding required to make these resources available?

# OAI-PMH

- Open Archives Initiative-Protocol for Metadata Harvesting (http://www.openarchives.org/)
- Roots in the ePrint community, although applicability is much broader
- Mission: "The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content."
- Content in this context is actually "metadata about content"

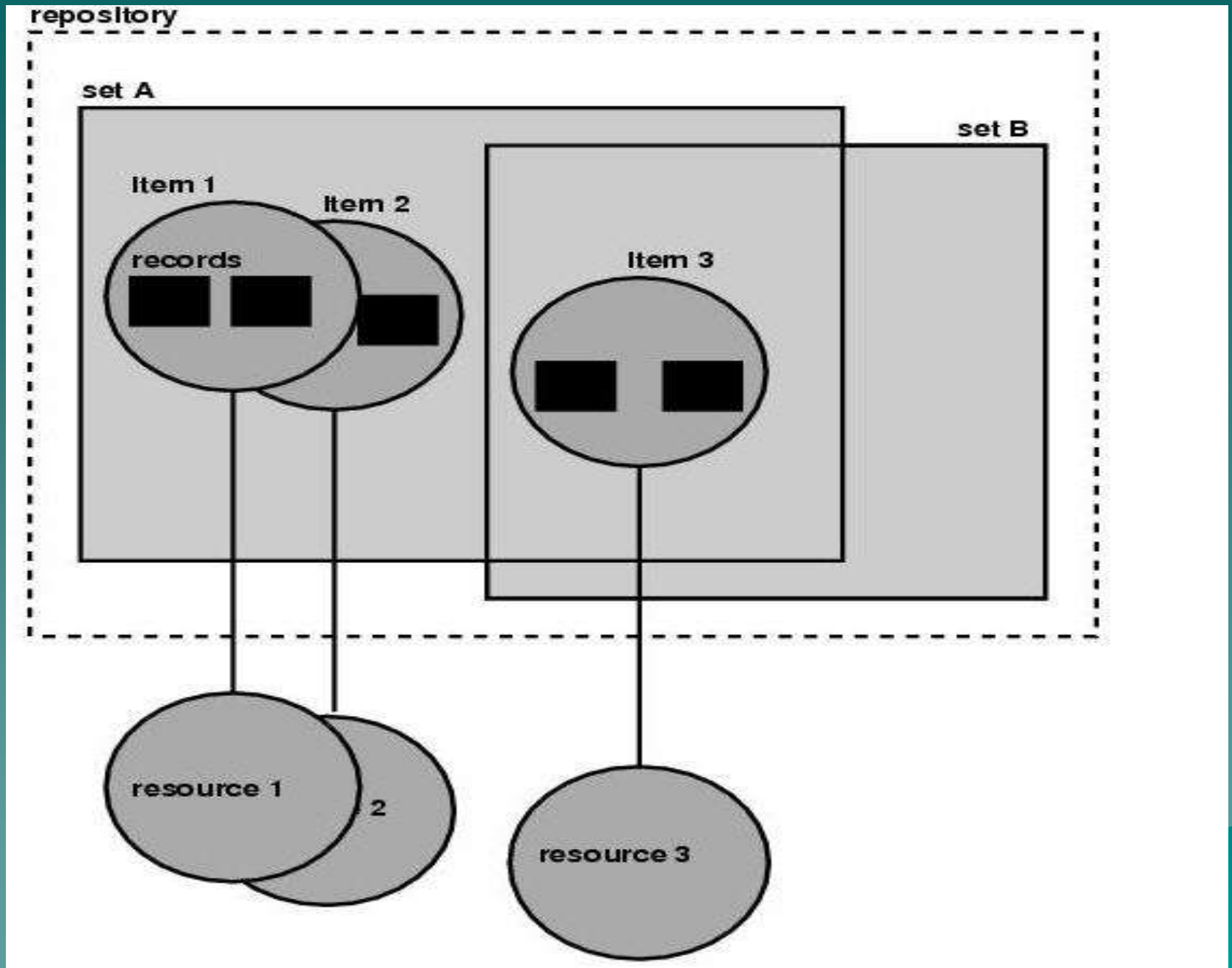

Meta-Metadata
Metadata About the Resource
OAI Wrapper

# OAI-PMH in a Nutshell

- Essentially provides a simple protocol for "harvest" and "exposure" of metadata records
- Specifies a simple "wrapper" around metadata records, providing metadata about the record itself
- OAI-PMH is about the metadata, not about the resources

# The OAI World

- Divided into two categories:
  - Data providers: "A data provider maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata."
  - Service providers: "A service provider issues OAI-PMH requests to data providers and uses the metadata as a basis for building value-added services."


repository
set A
set B
Item 1
Item 2
records
Item 3
resource 1
2
resource 3

# Other important definitions

- *Archive*: Not the same as 'archive' used in libraries, more like "repository"
- *Protocol*: a set of rules defining communication between systems. FTP (File Transfer Protocol) and HTTP (Hypertext Transport Protocol) are other examples of Internet protocols
- *Harvesting*: the gathering together of metadata from a number of distributed repositories into a combined data store

# Inside OAI Repositories

- *repository* - A repository is a network accessible server that can process requests. A repository is managed by a data provider to expose metadata to harvesters
- *resource* - A resource is the object or "stuff" that metadata is "about," whether physical or digital, stored in the repository or a constituent of another database
- *item* - An item is a constituent of a repository from which metadata about a resource can be disseminated
- *record* - A record is metadata in a specific metadata format

# OAI Goals

- Low barrier to participation
  - Server software available in many programming languages, intended to be easy to install
  - Server-less implementation available now via "Static repository" (essentially a web page that looks like an OAI response and can be harvested as such)
- Limited set of commands
- Predictable responses and flows of data

# Other OAI Info

- Responses are encoded in XML syntax
- OAI-PMH supports any metadata format encoded in XML—Simple Dublin Core is the minimal format specified
- Data Providers may define a logical set hierarchy to support levels of granularity for harvesting by Service Providers
- Date stamps flag the last change of the metadata set, and thus provide further support for granularity of harvesting
- OAI-PMH supports flow control

# OAI Requests

- Identify-->Returns general information about the particular OAI server
- ListMetadataFormats-->returns formats available
- ListSets-->returns list of sets available
- ListIdentifiers-->returns identifiers only
- ListRecords-->returns record ids in a set
- GetRecord-->returns particular record
- Try it out at the UIUC OIA Registry: (http://gita.grainger.uiuc.edu/registry/searchform.asp)

# Dates Used in OAI-PMH

- Datestamps are used as values in requests to support selective harvesting by date (generally latest update date of the metadata record)
- Datestamps are also used in record headers in responses
- Datestamps are particular to a repository
- Repeat: OAI dates are about the metadata, not the resources

# OAI-PMH Optional Containers

- Repository level
  - Rights
  - Branding
- Record level
  - About
    - Provenance
    - Rights

# About Container Example

```
<about>
<provenance xmlns="http://www.openarchives.org/OAI/2.0/provenance"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/provenance
                      http://www.openarchives.org/OAI/2.0/provenance.xsd">

<originDescription harvestDate="2002-02-02T14:10:02Z" altered="true">
  <baseURL>http://the.oa.org</baseURL>
  <identifier>oai:r2.org:klik001</identifier>
  <datestamp>2002-01-01</datestamp>
  <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
  <originDescription harvestDate="2002-01-01T11:10:01Z" altered="false">
    <baseURL>http://some.oa.org</baseURL>
    <identifier>oai:r2.org:klik001</identifier>
    <datestamp>2001-01-01</datestamp>
    <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
  </originDescription>
</originDescription>

</provenance>
</about>
```

# OAI Rights Expressions

- Rights expressions are valid at three levels:
  - Repository
  - Set
  - Record
- Rights expressed at the Repository and Set levels are not a substitute for expressions at the Record Level

# OAI Best Practices (DLF & NSDL)

- Guidelines for data providers and service providers
  - http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/Main_Page
- Best Practices for Shareable Metadata
  - http://webservices.itcs.umich.edu/mediawiki/oaibp/?PublicTOC

# OAI In Practice

- The UIUC OAI-PMH Data Provider Registry
  - http://gita.grainger.uiuc.edu/registry/searchform.asp
- Includes most known data providers
- Link on home page to Service Providers
- Provides multiple reports, sample records, browses, search, etc.
- Ex.: Show report from left hand menu: "Distinct Metadata Schemas"
  - http://gita.grainger.uiuc.edu/registry/ListSchemas.asp
  - Choose a schema, look for providers and sample records

# What's an OpenURL?

- The OpenURL provides a standardized format for transporting bibliographic metadata about objects between information services
- Provides a basis for building services via the notion of an *extended service-link*, which moves beyond the classic notion of a *reference link* (a link from metadata to the full-content described by the metadata)

"The OpenURL standard enables a user who has retrieved an article citation, for example, to obtain immediate access to the "most appropriate" copy of that object through the implementation of extended linking services. The selection of the best copy is based on user and organizational preferences regarding the location of the copy, its cost, and agreements with information suppliers, and similar considerations. This selection occurs without the knowledge of the user; it is made possible by the transport of metadata with the OpenURL link from the source citation to a "resolver" (the link server), which stores the preference information and the links to the appropriate material."

*--OpenURL Overview, SFX website*

# OpenURL Characteristics

- Protocol operates between an information resource and a service component
- Service component is called a "link server" or "link resolver"
- Link server defines the user context
- Takes source citation and determines whether a user has access

# Distinguishing Users

- Uses information stored in a cookie (the CookiePusher mechanism)
- Uses information contained in a digital certificate, such as the one proposed by the DLF digital certificates prototype project
- Identifies a user's IP address
- Obtains user attributes via the Shibboleth framework

# Examples of Extended Service Links

- From a record in an abstracting and indexing database (A&I) to the full-text described by the record
- From a record describing a book in a library catalogue to a description of the same book in an Internet book shop
- From a reference in a journal article to a record matching that reference in an A&I database
- From a citation in a journal article to a record in a library catalogue that shows the library holdings of the cited journal

# OpenURL Examples & Demo

- http://sfxserver.uni.edu/sfxmenu?issn=1234-5678&date=1998&volume=12&issue=2&spage=134
- An OpenURL demo:
  - http://www.ukoln.ac.uk/distributed-systems/openurl/

# Defining and Ensuring Metadata Quality

- What constitutes quality?
- Techniques for evaluating and enforcing consistency and predictability
- Automated metadata creation: advantages and disadvantages
- Metadata maintenance strategies

# Beginning to Define Quality

- Experience of the library community--BIBCO & NACO
  - Agreed upon standards for library quality
  - Training and documentation in support of practitioners
  - Review and enforcement of standards by means of institutional "buddy system"

# How Does Quality Happen?

- Lessons from the library community
  - Quality is quantifiable and measurable
  - To be effective, enforcement of standards of quality must take place at the community level
- Furthermore:
  - Data problems are not unique to particular communities
  - general strategies can improve interoperability

# Quality Measurement: Criteria

- Completeness
- Accuracy
- Provenance
- Conformance to expectations
- Logical consistency and coherence
- Timeliness (Currency and Lag)
- Accessibility

# Completeness

- "Metadata should describe the target objects as completely as economically feasible"
- "Element set should be applied to the target object population as completely as possible"

# Accuracy

- Information provided in values should be correct and factual
- Editing applied to:
  - Eliminate typos
  - Ensure conforming name expressions
  - Ensure standard abbreviations, usages in general

# Provenance

- Who prepared the metadata? What do we know about the preparer?
- What methods were used to create the metadata? Is it human created or created by machine?
- What transformations have been applied since creation?
- Where has it been before?

# Conformance to Expectations

- Contains elements a community would expect to find
- Controlled vocabularies are well-chosen and explicitly exposed to downstream users
- Metadata is reflective of community thinking about necessary compromises

# Logical Consistency/Coherence

- Standard mechanisms like application profiles and common crosswalks are used
- Similar structures and appearance are enabled for search results
- There is very limited reliance on defaulted values

# Timeliness

- Currency
  - Target object changes but metadata does not
- Lag
  - Target object disseminated before some or all metadata is available
- "Metadata aging" is affected by cultural differences between librarians and technologists
  - Librarians: once and it's done
  - Technologists: metadata as an iterative process

# Accessibility

- Barriers to accessibility may be economic, technical or organizational
  - Metadata as "premium" or proprietary information
  - Unreadable for technical reasons (file formats, etc.)
  - Metadata may not be properly linked to relevant object(s)

# Evaluating Metadata (1)

- Random sampling (XMLSpy)
  - Advantages
    - Includes some formatting and color coding
  - Disadvantages
    - Assumes consistency/predictability
    - Difficult to determine extent of problems found
    - Tedious, at best

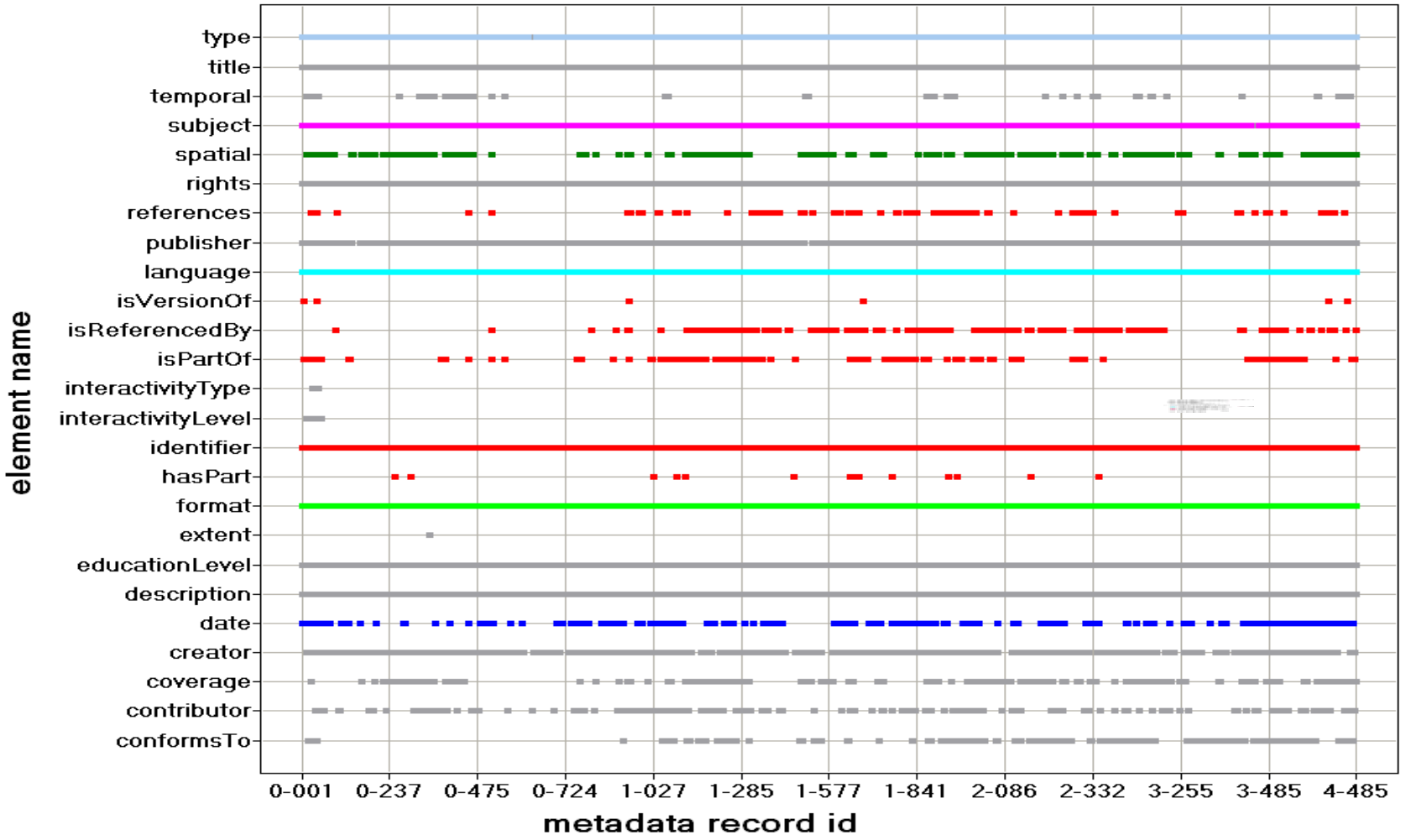
# Evaluating Metadata (2)

- Spreadsheets (Microsoft Excel)
  - Advantages
    - Better sorting and control by reviewer
  - Disadvantages
    - Unwieldy for large files
    - Requires sustained focus from reviewer
    - Requires translation into tab-delimited file

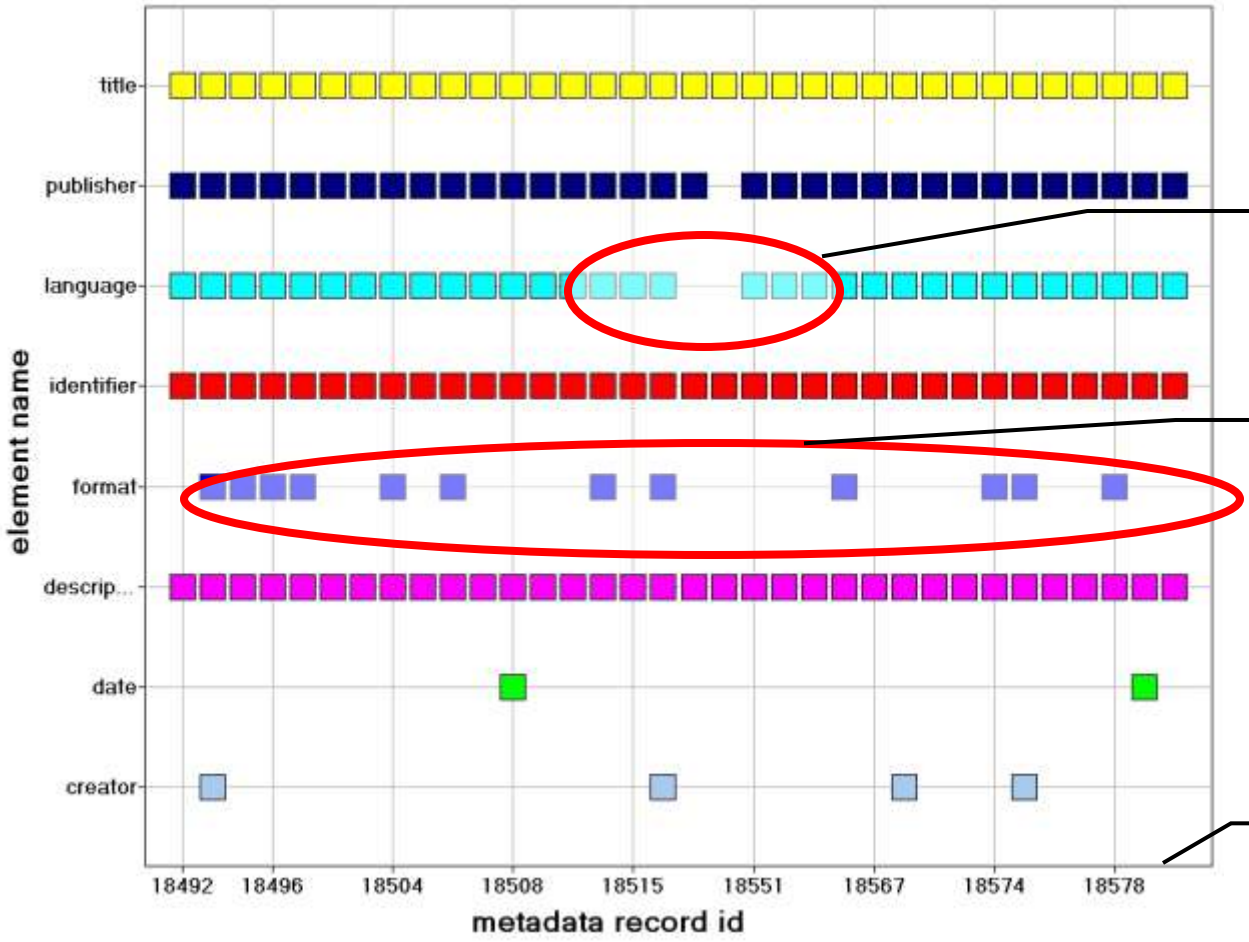
# Evaluating Metadata (3)

- Visual Graphical Analysis (Spotfire)
  - Advantages
    - View of several data dimensions simultaneously
    - Reviewer controls data display
    - Tends to pull reviewer focus to anomalies
    - Handles fairly large files at one time, while allowing subset views
    - Display manipulation possible without programmers
  - Disadvantages
    - High cost of software
    - Requires translation into tab-delimited file

# Element Names vs. Record Ids (Scatter Plot)



# Missing Elements (Scatter Plot)




**2 records without language element**

**format element present inconsistently**

**Easy to rescale axis on the fly and scroll through records**

# Table View

| metadata record id | element namespace | element name | element value |
|---|---|---|---|
| oai:scout.wisc.edu:ScoutNSDL-373 | http://purl.org/dc/elements/1.1/ | date | - |
| oai:scout.wisc.edu:ScoutNSDL-374 | http://purl.org/dc/elements/1.1/ | date | - |
| oai:scout.wisc.edu:ScoutNSDL-377 | http://purl.org/dc/elements/1.1/ | date | - |
| oai:scout.wisc.edu:ScoutNSDL-906 | http://purl.org/dc/elements/1.1/ | date | 1992 |
| oai:scout.wisc.edu:ScoutNSDL-308 | http://purl.org/dc/elements/1.1/ | date | 1993 |
| oai:scout.wisc.edu:ScoutNSDL-1013 | http://purl.org/dc/elements/1.1/ | date | 1994 - 2000 |
| oai:scout.wisc.edu:ScoutNSDL-391 | http://purl.org/dc/elements/1.1/ | date | 1995 |
| oai:scout.wisc.edu:ScoutNSDL-1533 | http://purl.org/dc/elements/1.1/ | date | 1997 |
| oai:scout.wisc.edu:ScoutNSDL-870 | http://purl.org/dc/elements/1.1/ | date | 1997 - 1999 |
| oai:scout.wisc.edu:ScoutNSDL-251 | http://purl.org/dc/elements/1.1/ | date | 1999 |
| oai:scout.wisc.edu:ScoutNSDL-1263 | http://purl.org/dc/elements/1.1/ | date | 1999, 2002 |
| oai:scout.wisc.edu:ScoutNSDL-694 | http://purl.org/dc/elements/1.1/ | date | 2000 |
| oai:scout.wisc.edu:ScoutNSDL-293 | http://purl.org/dc/elements/1.1/ | date | [-] |
| oai:scout.wisc.edu:ScoutNSDL-253 | http://purl.org/dc/elements/1.1/ | date | [-] |
| oai:scout.wisc.edu:ScoutNSDL-1168 | http://purl.org/dc/elements/1.1/ | date | [1993] |
| oai:scout.wisc.edu:ScoutNSDL-1103 | http://purl.org/dc/elements/1.1/ | date | [1996] |
| oai:scout.wisc.edu:ScoutNSDL-1055 | http://purl.org/dc/elements/1.1/ | date | [1997 - 2001] |
| oai:scout.wisc.edu:ScoutNSDL-256 | http://purl.org/dc/elements/1.1/ | date | [2002] |
| oai:scout.wisc.edu:ScoutNSDL-575 | http://purl.org/dc/elements/1.1/ | date | [c-] |
| oai:scout.wisc.edu:ScoutNSDL-710 | http://purl.org/dc/elements/1.1/ | date | [c-] |
| oai:scout.wisc.edu:ScoutNSDL-204 | http://purl.org/dc/elements/1.1/ | date | [c1940] |
| oai:scout.wisc.edu:ScoutNSDL-135 | http://purl.org/dc/elements/1.1/ | date | [c1995 - 1998] |
| oai:scout.wisc.edu:ScoutNSDL-375 | http://purl.org/dc/elements/1.1/ | date | [c1997] |
| oai:scout.wisc.edu:ScoutNSDL-212 | http://purl.org/dc/elements/1.1/ | date | [c2002] |
| oai:scout.wisc.edu:ScoutNSDL-542 | http://purl.org/dc/elements/1.1/ | date | c- |
| oai:scout.wisc.edu:ScoutNSDL-549 | http://purl.org/dc/elements/1.1/ | date | c1991 |
| oai:scout.wisc.edu:ScoutNSDL-531 | http://purl.org/dc/elements/1.1/ | date | c1993 |
| oai:scout.wisc.edu:ScoutNSDL-153 | http://purl.org/dc/elements/1.1/ | date | c1993 - 2001 |
| oai:scout.wisc.edu:ScoutNSDL-224 | http://purl.org/dc/elements/1.1/ | date | c2001 |
| oai:scout.wisc.edu:ScoutNSDL-246 | http://purl.org/dc/elements/1.1/ | date | c2001, 2002 |
| oai:scout.wisc.edu:ScoutNSDL-82 | http://purl.org/dc/elements/1.1/ | date | c2002 |

**Only DC Date elements are selected for display**

**Sorted by element value**

**Non-empty, "no information" values that may confuse end users**

**The only W3CDTF syntax present is four digits.**

# Improving Metadata Quality …

- Documentation
  - Basic standards, best practice guidelines, examples
  - Exposure and maintenance of local and community vocabularies
  - Application Profiles
  - Training materials, tools, methodologies

# … Over Time

- Culture change
  - Support for documentation and exchange of knowledge and experience
  - Routine contribution to the "general good"
  - More focused research on practical metadata use and quality considerations
  - Better project-based and community-wide documentation

# Crosswalking

"Crosswalks support conversion projects and semantic interoperability to enable searching across heterogeneous distributed databases. Inherently, there are limitations to crosswalks; there is rarely a one-to-one correspondence between the fields or data elements in different information systems."

-- Mary Woodley, *"Crosswalks: The Path to Universal Access?"*

"Metadata schema transformations are more complex than purely structural transforms because they require a set of equivalences identified by human experts—Dublin Core title can be mapped to MARC 245, Dublin Core author can be mapped to MARC 100 and so on—but this important knowledge is recorded in a multitude of ways that are not standardized and not always machine-processable, including Web pages, databases, spreadsheets, PDF documents, and the source code of many computer languages."

-- Jean Godby, *Two Paths to Interoperable Metadata*

# Crosswalks

- In general: Semantic mapping of elements between source and target metadata standards
- The process of metadata conversion specification includes transformations required to convert a metadata record content to another format, including:
  - Element to element mapping
  - Hierarchy and object resolution
  - Metadata content conversions
  - Stylesheets can be created to transform metadata based on crosswalks

## II. MARC to Dublin Core Crosswalk (Unqualified).

Conventions: "$" is used to represent the control character subfield delimiter.

| DC Element | MARC Fields | Implementation notes |
|---|---|---|
| Title | 245 | |
| Creator | 100, 110, 111, 700, 710, 711 | See Appendix 1 below; Contributor element not used. |
| | 720 | |
| Subject | 600, 610, 611, 630, 650, 653 | |
| Description | 500-599, except 506, 530, 540, 546 | |
| Contributor | | See Appendix 1 below; Contributor element not used. |
| Publisher | 260$a$b | |
| Date | 260$c | |
| Type | Leader06, Leader07 | See Appendix 2 for Leader-Type rules |
| | 655 | |
| Format | 856$q | |
| Identifier | 856$u | |
| Source | 786$o$t | |
| Language | 008/35-37 | |

## III. MARC to Dublin Core Crosswalk (Qualified).

| DC Element | DC Qualifier(s) | MARC Fields | Implementation notes |
|---|---|---|---|
| Title | | 245 | |
| Title | Alternative | 130, 210, 240, 242, 246, 730, 740 | |
| Creator | | 100, 110, 111, 700, 710, 711 | See Appendix 1 below. |
| | | 720 | |
| Subject | LCSH | 600, 610, 611, 630, 650 | Second indicator=0 |
| Subject | MeSH | 600, 610, 611, 630, 650 | Second indicator=2 |
| Subject | LCC | 050 | |
| Subject | DDC | 082 | |
| Subject | UDC | 080 | |
| Description | | 500-599, except 505, 506, 520, 530, 540, 546 | |
| Description | TableofContents | 505 | |
| Description | Abstract | 520 | First indicator=3 |
| Contributor | | | See Appendix 1 below; Contributor element not used. |
| Publisher | | 260$a$b | |

# Available Crosswalks

- Library of Congress
  - http://www.loc.gov/marc/marcdocz.html
- MIT
  - http://libraries.mit.edu/guides/subjects/metadata/mappings.html
- Getty
  - http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html

# Problems With Converted Records

- Differences in granularity (complex vs. simple scheme)
  - Some data might be lost
  - Differences in semantics can occur
  - Differences in use of content standards make sharing sometimes problematic
  - Properties may vary (e.g. repeatability)
- Converting everything may not always be the best solution

# Example:Mapping MODS:title to DC:title

- Includes attribute for type of title
  - Abbreviated
  - Translated
  - Alternative
  - Uniform
- Other attributes:
  - ID,authority,displayLabel,xLink
- Subelements: title, partName, partNumber, nonSort

# Mapping MODS:title to DC:title

- DC has one element refinement:

Alternative

  - DC title has no substructure; MODS allows for subelements for partNumber, partName
- Best practice statement in DC-Lib says to include initial article
  - MODS parses into<nonSort>
- MODS can link to a title in an authority file if desired

# Exercise

- Evaluate a small set of human and machine-created metadata.