Digital libraries and metadata interoperability

Lecture 1 at Masaryk University

Nils Pharo

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Who am I?

- Professor in knowledge organization and information retrieveal at Oslo and Akershus University College
- Teaching courses in Knowledge organization and IR (bachelor level) and Digital knowledge organization (master level)
- Research interests: interactive information retrieval, information searching behaviour, metadata interoperability

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Content

- Introduction to digital libraries
  - Technological development
  - History
  - Typology
- Indexing digital libraries
  - Owner
  - Collection
  - User
- Metadata interoperability problems
  - Metadata types
  - Case: library metadata interoperability
  - Solutions

From document surrogates to full text 1

- early 1960s: first complete texts (full text) made computer searchable
- 1960s: first full text information retrieval systems developed
- 1970s: common to use free text search in bibliographic databases
- 1980s: lots of experience in efficient full text search algorithms
- 1980s: comparisons of full text versus controlled vocabularies
- 1990s: the Web arrives
- 2000s: digital libraries and repositories of digital documents

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

What is a digital document?

- Digital originals
  - a web page?
  - a web site?
  - *.wpd, *.docx, *.pdf, *.ps, *.xls, *.*?
  - images, movies, songs etc?
  - ebooks?

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Digital library history

- Some «classic» digital libraries
- Project Gutenberg (1971) - the oldest digital library
- Perseus Project (1987) - classic texts
- Project Runeberg (1992) - Nordic equivalent to Project Gutenberg
- American Memory (1994) - Library of Congress Digital Archive
- The World Digital Library (2007) - Library of Congress
- IFLA's list of resources (from 2008)
- Wikipedia's list of digital library projects (dated 2013)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Typology of digital libraries

- discipline oriented
- format/genre oriented
- institutional
- task oriented

(Bearman, 2007)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Discipline oriented digital libraries

- designed for specific group of users
- barriers for non-experts
- internal documentation of systems

- Example: arXiv.org
- Example 2: Countway library of medicine

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Institutional DLs

- academic institutions
- external providers
- digital repositories

- Example1: ODA
- Example2: Repozitar.cz

Audience/Task-oriented DLs

- distance learning
- children

- Example: Open Library, from The Open University
- Example: International Children's Digital Library

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Digital library architecture

- collection
- metadata
- interface
- services
- authentication
- …

Challenges

- Acquisition
- Indexing
- Information retrieval
- User-centered design/personalization
- Economics

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Acquisition

- digitization
- harvesting
- collaboration/federation
- purchase
  - license agreements
  - consortium agreements

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Indexing

- What?
  - metadata
  - full text
  - semantic markup
- Who?
  - owners
  - Authors (Collection)
  - users

Information retrieval

- boolean perfect match
- full text best match
- combination

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

User-centered design/personalization

- user-centred interfaces
- authentication
- user profiles
- interaction
- recommendation
- privacy
- etc

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Economics

to secure economic sustainability it is crucial to integrate the DL in the mother organization!

- costs:
  - collection development (acquisition)
  - software & hardware
  - tech staff
  - technological updates

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Discussion

– What roles should the librarian take?

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Who should index?

1. owners
2. collection
3. users

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

1. Owner based indexing

- subject experts
- indexing experts

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Old tradition

– constitute important basis of librarians' profession
– based in the "second order of order“ (D. Weinberger)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Owners

- know their users
- know their collection
- should develop an indexing policy based on this knowledge

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Manual indexing

- consistent(?)
- language independent
- provides query support
- use subject languages

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Objectives of subject languages

- collocation of documents
- facilitate navigation

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Collocation objective

Charles Cutter's objective from 1876:

To show what the library has

- by a given author
- on a given subject
- in a given kind of literature

Collocation measures

recall

number of relevant documents found/number of relevant documents in collection

precision

number of relevant documents found/number of retrieved documents

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Navigation objective

"To navigate a bibliographic database (that is, to find works related to a given work by generalization, association, and aggregation; to find attributes related by equivalence, association, and hierarchy)." (Elaine Svenonius, 2000)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Subject language typology

- controlled vocabularies
- alphabetic subject-languages
- classification languages

Semantics

- category semantics
- referential semantics
- relational semantics

Category semantics

subject languages can be:

- synthetic or
- enumerative

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Referential semantics

What a term in a subject language refers to:

- the set of all documents indexed with the term
- procedures for dealing with homonyms, polysemes and the unclarity of language
- semantic disambiguation

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Methods for clarifying language

- domain specification
- qualifiers
- notes
- hierarchy

Relational semantics

- hierarchical
- eqivalence (synonymy)
- relatedness

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Vocabulary selection

- terminology
- domain definition
- warrant

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Warrant

- literary warrant
- use warrant
- structural warrant

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

2. Collection based indexing

- authors' own words:
  - explicit keywords
  - Automatic extraction of meaning

Automatic indexing

- let the content represent itself - automatic indexing
- neutral
- fast
- cost-efficient

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

The basics of automatic indexing

The indexing task consists of

- first to assign to each stored item terms, or concepts, capable of representing document content, and
- second to assign to each term a weight, or value, reflecting its presumed importance for purposes of content identification (Salton & McGill, 1983)

Good index terms fulfill to purposes

1. The term must be related to the documents content so as to make it retrievable when it is wanted, but also
2. A good index term should distinguish between the documents it is assigned to from the remainder in order to prevent indiscriminate retrieval of all documents.

Inverse Document Frequency (IDF) Weight

term importance is:

- proportional to the occurrence frequency of each term k in each document i (FREQik)
- inversely proportional to the total number of documents to which each term is assigned

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Other components in automatic indexing weighting algorithms

- term placement
- term proximity
- element markup
- pagerank
- popularity

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

3. User based indexing

create folksonomies

- based on "folk" + "taxonomies"
- consists of tags (generating the verb "to tag")
- related to "free keywords"
- users make up their own index terms
- uncontrolled vocabulary
- Citeulike and Librarything

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

How are folksonomies used?

- users may tag their own and other users' collections
- used on a variety of digital collections (bookmarks, pictures, books, articles...)
- facilitates the indexing of new topics

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Some observation of tag content

describe different aspects (facets?) of the object they index

- aboutness (at variable levels of abstraction)
- emotional characteristics
- genre
- place
- space
- whereabouts
- +++

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Metadata interoperability

- Different types of metadata
- Case: library metadata
- Interoperability

Definition

Metadata is machine understandable information about web resources or other things (Berners-Lee, 1997)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Purposes of metadata

- facilitate information retrieval
- document management
- document encoding and analysis

OSLO AND AKERSHUS UNIVERSITY COLLEGE OF APPLIED SCIENCES

Metadata types

1. descriptive metadata
   - Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
2. structural metadata
   - Structural metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
3. administrative metadata
   - Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

(NISO, 2004)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Metadata, level of aggregation

- Collection level
- Series/volume level
- Document level
- Document part level

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Case: bibliographic metadata

- MARC
- Dublin Core

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Dublin Core

DCMES - Dublin Core Metadata Element Set

– initiated at a workshop in Dublin, Ohio in 1995
– consists of 15 core elements

(DC Metadata Element Set)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

DC characteristics

- very simple metadata schemes
- descriptive, structural and administrative elements
- no obligatory elements
- all elements are repetitive
- system independent
- several syntaxes
  - html
  - xml
  - rdf

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Qualified Dublin Core

- the objective is to refine the 15 core elements
- element represent same content with more specificity
- use *encoding schemes* to restrict interpretation
- in addition three new elements have been added to QDC
  - Audience
  - Provenance
  - RightsHolder

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Enchoding schemes

used to specify how the value taken by an element should be interpreted. There are two types of encoding schemes:

- vocabulary encoding schemes
  - DCMI type vocabulary
  - DDC
- syntax encoding schemes
  - ISO 3166

(Encoding schemes)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Example bibliographic record in MARC

*000 $a010041702
*008 apbv$cnob$hno
*009c $ak
*015 $anf0105587
*020 $a82-05-27748-6$bib.
*080c $a839.6
*082g $d839.823[S]
*082kj$a839.82
*082xn$a839.82
*086d $aS 4b
*100 $aHamsun, Knut$d1859-1952
*245 $aSult$cKnut Hamsun
*260 $a[Oslo]$bGyldendal$c2001
*300 $a147 s.
*440 $aGyldendals 10 store
*500 $a1. utg. København : Philipsen, 1890
*776 $w101353413

BIBSYS-MARC

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

MARC-record transformed to Dublin Core

DC.Format="p"
DC.Type="v"
DC.Language="nob"
DC.Identifier="82-05-27748-6"
DC.Subject="839.823[S]"
DC.Subject="839.82"
DC.Subject="839.82"
DC.Creator="Hamsun, Knut"
DC.Creator="1859-1952"
DC.Title="Sult"
DC.Publisher="Gyldendal"
DC.Date="2001"
DC.Description="1. utg. København : Philipsen, 1890"
DC.Relation="101353413"

Mapping from BIBSYS-MARC to Dublin Core

Example 2

Movie: True Grit

MARCXML

Dublin Core

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Problem

Similar kind of content is described using different metadata standards and syntaxes

Why?

Create groups of four and discuss possible reasons why such situations occur?

Interoperability

"Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality."

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Metadata interoperability

- 3 levels of interoperability
  - schema level
  - record level
  - repository level

(Chan & Zeng, 2006; Zeng & Chan, 2006)

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

MARCsism: the uniform solution

Everybody should use the same system! :-)

top down approach (MARCsism): "everybody should use MARC", is in theory the optimal solution, but is no longer realistic

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Schema level interoperability

Efforts are focused on the elements of the schemas, being independent of any applications:

- derivation
- application profiles
- crosswalks
- switching-across
- metadata framework
- metadata registry

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Derivation

To create a new schema from an existing source schema, e.g.:

- USMARC and MARC21 har been the basis for a number of national MARC formats, including NORMARC
- MARCXML is also based on MARC21, example from Library of Congress
- Qualified Dublin Core is derived from simple DC, e.g. <abstract> is a refinement of <description>

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

The DC dumb down principle

- The fifteen elements should be usable and understandable with or without the qualifiers
- Like saying that nouns can stand on their own without adjectives
- If your search engine encounters an unfamiliar qualifier, look it up somewhere -- or just ignore it!
- To test whether a qualifier is "good", cover the qualifiers with your hand and ask:
  - Does the statement still make sense?
  - Is it correct?

OSLO AND AKERSHUS UNIVERSITY COLLEGE OF APPLIED SCIENCES

Application profiles

Solutions based on a (combination of) existing schemas optimized for a specific community. The developers will typically have a bottom-up approach

- an AP may also specify the specific value schemes, cardinality and syntaxes that are allowed used
- Dublin Core is often used to provide the core set of elements
- combination of DC and LOM for education purposes
- BIBLINK Core BIBLINK was an attempt at coupling ePublishers and national bibliographic agencies.
- new elements must be accompanied by a namespace declaration

OSLO AND AKERSHUS UNIVERSITY COLLEGE OF APPLIED SCIENCES

Crosswalks

specifies the mapping of elements, semantics and syntax between different schemas. Very common way of securing interoperability. Two forms of crosswalks, absolute and relative

- absolute crosswalk:
  - exact mapping between elements in the two schemas: MARC.260$c = DC.date.created
- relative crosswalk:
  - mapping between elements that do not share eqivalent meaning:MARC.240 (Uniform title) = DC.title

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Crosswalk challenges

- different degres of equivalence; one-to-one, one-to-many, many-to-one, one-to-none
- crosswalks work better from a complex schema to a simpler one
- crosswalk from a complex to a simple schema results in data loss. Several different MARC fields map to DC.subject

Examples

- TEI header to MARC
- Dublin Core to LOM
- Dublin Core to MARC
- MARC to Dublin Core

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Switching across

To map between multiple schemas one of the schemas are used as a switch

Example: Getty's crosswalk where CDWA (Categories for the Description of Works of Art) is used to map between 12 different schemas

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Metadata frameworks

Created to provide guidelines for developers of metadata schemas in specific environments. Frameworks can be developed based on existing schemas or prior to any schema has been developed

Example: OAIS reference model

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Metadata registries

Provide overview of relevant metadata schemas, their elements, syntax, semantics etc to facilitate adoption and reuse of existing schemas in favour of creating new (and redundant) schemas

Example: Dublin Core Metadata Registry

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Record level interoperability

Efforts are intended to integrate the metadata records through the mapping of the elements according to the semantic meanings of these elements.

- conversion
- data reuse and integration

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Metadata conversion

convert the content of metadata records in one schema into another

– problem: data loss

– Example from Zeng & Chan

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Data reuse and integration

This approach is based on combining metadata from various sources in one common format. METS (Metadata Encoding and Transmission Standard) and RDF provide two different solutions for this.

- a METS record (may) contain its own metadata on an item in addition to pointers to metadata records that describe the same item
- RDF records use XML's namespace declaration to combine the values from different metadata schemas

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

RDF example

```
<?xml version="1.0"?>
 <rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.0/"
 xmlns:skos="http://www.w3.org/2004/02/skos/core#">
 <rdf:Description
rdf:about="http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=3792">
 <dc:title>The Intellectual Foundation of Information Organization</dc:title>
 <dc:creator>Elaine Svenonius</dc:creator> <dc:publisher>MIT Press</dc:publisher>
 <dc:date>1999</dc:date>
 <dc:language>en</dc:language>
 <skos:prefLabel>information organization;/skos:prefLabel>
 </rdf:Description>
 </rdf:RDF>
```

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Repository level interoperability

Secure interoperability of harvested or integrated records from varying sources

Different solutions:

– keep each providers original format
– aggregation of metadata from different sources
– convert/integrate into a standard format

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

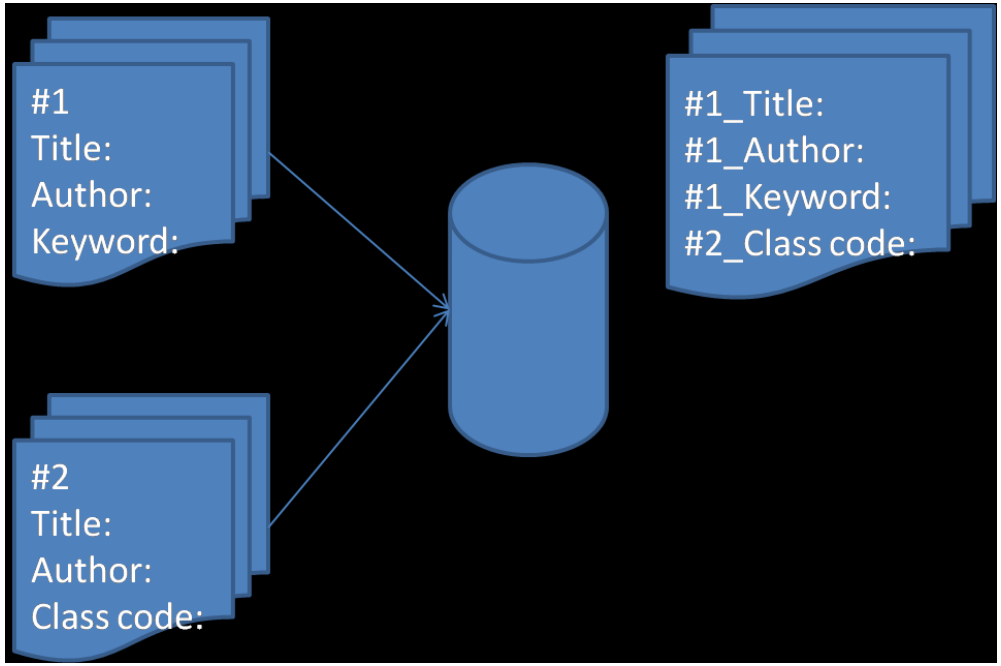
No conversion

- harvest from several sources
- store metadata in original schema
- add collection meta data
- possible to enter the individual source via the repository

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Content aggregation

The content of a metadata record for an item is the aggregated contents of all the item's metadata records



What's next?

- Wednesday: semantic web and linked data
- Thursday: ontology modelling, RDF and Topic Maps