Collection mapping: models and standards for international cooperation

Juha Hakala
Director, IT
The National Library of Finland
Just for comparison…

• There is a good understanding on how to share the burden of bibliographic description
  – IFLA Universal Bibliographic Control and International MARC Core activity (UBCIM)
  – AACR2 & RDA development
  – MARC21 & other MARC formats
  – UNICODE character set

• With appropriate tools (good Z39.50 client and MARC converter) copy cataloguing works well, although there is still room for improvement
It's different with collections…

- There is little political support and coordination on behalf of IFLA
- There are no cataloguing rules
- Common format and exchange syntax have been missing
  - But this problem has been fixed recently!
- And there is no common understanding of where collection description fits in into the larger picture
Setting the scene: portal metadata

- Metasearch portals need two kinds of metadata
  - Information retrieval service descriptions
    - E.g. Z39.50 or SRU server of Helka, the Helsinki University Libraries OPAC
  - Collection descriptions
    - E.g. Nordenskiöld map collection in Helsinki University Library

- No portal or Web OPAC will work without service descriptions; the need for collection descriptions becomes urgent only when there are a lot of remote databases to search

- Generally, portal metadata helps the human and software agents to locate & search resources in the deep Web (databases)
The NISO Metasearch Initiative

• Response to content provider/service provider concerns and libraries’ need to organize the production of “metadata for metasearch” better
• Bring together
  – Content providers
  – System vendors
  – Library service providers
  – Standards developers
• "To identify, develop, and frame the standards and other common understandings that are needed to enable an efficient and robust information environment"
The NISO MI: aims

• Aims to enable
  – metasearch service providers to offer more effective and responsive services
  – content providers to deliver enhanced content and protect their intellectual property
  – libraries to deliver services that distinguish their offerings from other free web services

• A set of draft standards and other documents published in November 2005
  – [http://www.niso.org/committees/MS_initiative.html](http://www.niso.org/committees/MS_initiative.html)
The NISO MI Task Groups

- **Task Group 1: Access Management**
  - Gather requirements for access/authentication
  - Describe existing processes
  - Develop use cases

- **Task Group 2: Collection Description**
  - Establish metasearch services' requirements for description of
    - Collections
    - Services which provide access to Collections ("Informational Services")
  - Select/develop metadata schemas
  - Recommend syntax for representation & data exchange

- **Task Group 3: Search & Retrieval**
  - Describe existing practice
  - Metadata to describe result sets
  - Metadata to describe article-level citations
Term definitions

• Item
  – A physical or digital entity

• Collection
  – An aggregation of one or more items

• Service
  – The provision of, or system of supplying, one or more functions of interest to an end-user or software application.
  – Physical or digital
  – Digital services may be "structured" or "unstructured"

• Informational services
  – Services that provide access to, or metadata about, items and/or collections
    – JISC Information Environment Architecture: Glossary
Functional requirements

• Allow an agent to
  – **Discover** collections of potential interest
  – **Identify** a collection
  – **Select** one or more collections from amongst a number of discovered collections
  – **Identify** the informational services that provide access to the collection

  – **Select** a service with which to interact
  – **Interact** with service
    • Subject to "knowledge" of interface semantics
Relations between collections and services

• Relationships exist
  – Between collections and services
  – Between collections

• In NISO MI conceptual model
  – A collection *is-made-available-by* zero or more services
  – A service *makes-available* exactly one collection
  – A collection *is-part-of* zero or more (super-) collections (parent)
  – A collection *has-part* zero or more (sub-) collections (child)
Collection & service descriptions

- NISO MI TG2 has specified metadata for collections & services
  - Data model (as outlined in the previous slide; in practice there are elements in both formats to facilitate the linking)
  - Metadata semantics
  - Syntax(es) for representation and data exchange
  - Guidelines for use (under development)
- N.B. TG2 has **not** (and will not):
  - build a service; or
  - specify the architecture within which a service might operate
  - specify the protocols for the exchange of collection/service metadata (though OAI-PMH is a good candidate for this)
- Has built on existing work
  - RSLP & DC Collection Description Working Group
  - ZeeRex
RSLP

– support for academic research
– improve discovery of library/archive collections
– also collaborative collection management
– recognition of collection level description as important mechanism for disclosure/discovery
– Collection description project, 1999-2000
  • DC-based metadata schema developed by Andy Powell from UKOLN
  • Concerns over maintenance & persistence
Dublin Core Collection Description Working Group

• Active 2001 (really 2003!) -
• Provide forum for sharing information about collection level description activity
• Develop a DC Application Profile for collection-level description
  – Specification of how DC (and other) properties are used for describing collections
• Develop supporting materials for use of AP
• Informed by experience of RSLP CD implementers and other initiatives
  – RSLP projects, TEL, others
DC CD Application Profile

• A "core" set of collection description properties
  – For simple collection-level descriptions
  – Suitable for a broad range of collections
  – Primarily to support discovery of collections
• Examine collection attributes (only) of RSLP CD Schema as starting point
• DC CD AP building on Michael Heaney’s E-R model
  – introduces Service as entity-type
  – describes Collection-Location, Collection-Service, Collection-Agent relationships
  – but excludes Location, Service, Agent description
NISO Z39.91-200X Collection Description Specification

- Compatible with DC CD AP, but extends it a bit
  - No changes or deletions
  - The plan is to align the development of DC CD AP and NISO standard, in order to prevent confusion

- Has been published as a draft for comments; comment period will last until November 4, 2006
  - Final draft including term definitions and comments up to that point will be released in Spring 2006

- The main author of the text is Pete Johnston from UKOLN, who also chairs the DC CD WG
NISO Z39.91-200X Contents

• Topics covered
  – Identification of collection
  – Content of items in collection
  – Completeness of collection (using Conspectus)
  – Form of items in collection
  – Process by which items gathered into collection
  – Ownership of collection
  – Rights of access to/use of collection
  – Location of collection
  – Services that provide access to collection
  – Relationships between collections
Some issues

• Existence of an element does not mean that it is easy to use it (wisely)
  – There is no standard collection identifier; using local identifiers is not a good idea
    • Helsinki University library is developing International Standard Collection Identifier NWI for ISO TC 46, building upon ISIL (ISO 15511)

• Existence of an exchange format does not mean that it makes sense to exchange the records (internationally)
  – Multilinguality, common subject headings & classifications, shared guidelines / cataloguing rules, systematic description of completeness
Some further issues

- Portal and ILMS vendors are likely to implement service descriptions first
  - The need is clear, and technical problems are smaller: two-step searching (first the collections, then the databases themselves) is difficult to implement well

- Current business model is in favour of the vendors providing service descriptions
  - Libraries should not accept this; vendor approach is not scaleable and the quality and timeliness of the metadata has not been sufficient

- What is the libraries’ role in portal metadata production – is it an important part of what we do?
Conclusion

- Libraries no longer use just ILMS; there are also portals, digital asset management systems and ERMs
- These new applications require new kinds of metadata; first examples of this are IR service descriptions and collection descriptions (with preservation metadata to follow)
- Providing this new metadata is as important as production of bibliographic data, since otherwise the new systems will not work
  - This is a challenge for library management
  - Shift from bibliographic description; automatic processing of (digital) text (20,000 pages per month per operator in my library, including scanning, OCR and automatic metadata creation) can complement traditional workflows
- What kind of impact does this have on our bibliographic products, such as the national bibliography?
  - Should service & collection descriptions be part of them?