Metadata harmonization

*for fun and profit*

Keynote, DC 2011

*Mikael Nilsson*
About me

➤ Who am I?
  • PhD thesis on metadata harmonization.
  • Worked on metadata interop for ten years
  • Former co-author/co-chair of several DCMI activities

➤ Why am I here?
  • To tell you about my experiences with metadata harmonization!
Humble beginnings...

Conzilla, in 1999
Looking for the right metadata format

- We had a tool, we needed interoperability
- Choose your community:
  - Web? Dublin Core or just <META> tags
  - Education?? IMS Metadata, later IEEE LOM
  - Libraries? MARC!
  - On your own? XML!
  - Video? MPEG-7!
<xml version='1.0' encoding='UTF-8' >
<mods ...>
  <subject authority="lcsh">
    <topic>Parachuting</topic>
  </subject>
  <extension>
    <!-- LOM fragment: -->
    <lom:description>
      <lom:string lom:language="en">
        Useful for learning some flight-related French terminology.
      </lom:string>
      <lom:string language="sv">
        Användbar för att lära sig lite flygrelaterad fransk terminologi.
      </lom:string>
    </lom:description>
    <!-- End LOM fragment. -->
  </extension>
</mods>
Vision: The metadata ecosystem
Metadata:

*Descriptive data about identifiable things*
Central definitions

 Metadata interoperability:

*the ability of two or more systems or components to exchange descriptive data about things, and to interpret the descriptive data that has been exchanged in a way that is consistent with the interpretation of the creator of the data*
This learning object explains parachuting.

Useful for learning some flight-related French terminology.

Edutella = P2P + RDF

- Reuse of learning content
- Information flexibility
- Ideology
- Agents & Smart systems
- Political/organizational issues
- Community processes

Information flexibility
Lessons from Edutella

- Common metadata semantics extremely powerful foundation
- But: Broad low-level system interoperability unlikely
- Strong need to focus on the standards
Core metadata concepts

- Metadata syntaxes
- Metadata vocabularies
- Abstract Models
- Application profiles
Concepts: Metadata syntaxes

- Concrete data formats used to exchange metadata between applications
- Used to implement metadata software and protocols
- Practical tool, but not essence of harmonization issues
- Examples: RDF/XML, DC-HTML, LOM XML
DC RDF

```xml
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dcterms="http://purl.org/dc/terms/"
         xmlns:ex="http://example.org/taxonomy/"
         xmlns:dcam="http://purl.org/dc/rdf/">

  <rdf:Description rdf:about="http://example.org/123">
    <dcterms:subject>
      <rdf:Description rdf:about="http://example.org/subject32">
        <dcam:memberOf
          rdf:resource="http://example.org/taxonomy/ExampleSubjects"/>
        <rdf:value xml:lang="en">Biology</rdf:value>
        <rdf:value xml:lang="sv">Biologi</rdf:value>
        <rdf:value
          rdf:datatype="http://example.org/taxonomy/SubjectEncoding">
          EA32</rdf:value>
      </rdf:Description>
    </dcterms:subject>
  </rdf:Description>
</rdf:RDF>
```
Concepts: Vocabularies

- Sets of descriptive terms for use in metadata descriptions
- Used by metadata “designers” to create metadata records
- Defined according to various models; conflicting models lead to harmonization issues
- Examples: DCMES, MARCREL, LCSH
<table>
<thead>
<tr>
<th><strong>URI</strong></th>
<th><a href="http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#InteractivityType-active">http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#InteractivityType-active</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Qualified Name:</strong></td>
<td>lomvoc:InteractivityType-active</td>
</tr>
<tr>
<td><strong>Defined By:</strong></td>
<td>LOM Vocabularies <a href="http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#">http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#</a></td>
</tr>
<tr>
<td><strong>Type:</strong></td>
<td>lom:InteractivityType</td>
</tr>
<tr>
<td><strong>Member Of:</strong></td>
<td>lomvoc:InteractivityType</td>
</tr>
<tr>
<td><strong>Label:</strong></td>
<td>active</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td>....</td>
</tr>
<tr>
<td><strong>Comment:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>See Also:</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>URI</strong></th>
<th><a href="http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#InteractivityType-expositive">http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#InteractivityType-expositive</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Qualified Name:</strong></td>
<td>lomvoc:InteractivityType-expositive</td>
</tr>
<tr>
<td><strong>Defined By:</strong></td>
<td>LOM Vocabularies <a href="http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#">http://ltsc.ieee.org/rdf/lomv1p0/vocabulary#</a></td>
</tr>
<tr>
<td><strong>Type:</strong></td>
<td>lom:InteractivityType</td>
</tr>
<tr>
<td><strong>Member Of:</strong></td>
<td>lomvoc:InteractivityType</td>
</tr>
<tr>
<td><strong>Label:</strong></td>
<td>expositive</td>
</tr>
<tr>
<td><strong>Definition:</strong></td>
<td>....</td>
</tr>
<tr>
<td><strong>Comment:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>See Also:</strong></td>
<td></td>
</tr>
</tbody>
</table>
Concepts: Abstract models

- Models used to define the meaning and usage of metadata terms
- Used by term designers and syntax creators to create interoperable specifications
- Incompatible models are major barrier to harmonization
- Examples: RDF triple model, IEEE LOM hierarchical model, DCMI abstract model
Example abstract model: DCAM
Concepts: Application profiles

- Defines the structure of metadata records in a particular context (domain, application, etc.)
  - use vocabularies on the basis of an abstract model to define a concrete syntax
- Used by application designers and domain experts to codify domain needs
- Useful for harmonization within the context of a single abstract model
- Examples: ePrints AP, OAI-DC, etc
Convert ALL the standards!!!
22

- **IEEE LOM**
- **NorLOM**
- **UK LOM Core**
- **IMS**
- **MARC21**
- **MARC-XML**
- **METS**
- **ISO MLR**
- **RDA**
- **Dublin Core**
- **Semantic Web**

### Timeline:

- **2001:** IMS Metadata RDF binding
- **2002-2006:** LOM RDF binding
- **2006:** DCMI/LTSC Task force (chair)
- **2005:** Abstract Model
- **2007:** RDF expression
- **2007:** Architecture Forum (co-moderator)
- **2008:** Application Profile Models
- **2008:** Interoperability Levels
- **2006:** DCMI submission on interoperability
- **2008:** "Requirements for ISO MLR interoperability"
- **2007-:** DCMI/RDA Task Group
- **2008-:** Application Profile Models
- **2008-:** Interoperability Levels

---

- **RDF**
- **Dublin Core**
- **DC APs**
- **Semantic Web**
- **ISO MLR**
- **IMS**
- **2001-:** IMS Metadata RDF binding
- **2005-:** Abstract Model
- **2006:** DCMI/RDA Task Group
- **2007-:** Abstract Model
- **2005:** Abstract Model
- **2007:** RDF expression
- **2007:** Architecture Forum (co-moderator)
- **2008:** Application Profile Models
- **2008:** Interoperability Levels
Convert all the standards?
Central definitions

Metadata harmonization:

the ability of two or more systems or components to exchange ”combined metadata” conforming to two or more metadata specifications, and to interpret the metadata that has been exchanged in a way that is consistent with the intentions of the creators of the metadata.
Harmonization in software

Fedora – digital repository

- Storage of multiple metadata standards
- Exposes two standards for search:
  - Dublin Core (15 properties)
  - Object relations (RDF)

SCAM – RDF-based repository

- Import of multiple metadata standards
- Exposes full metadata for search
Semantic metadata interoperability

- When two systems can exchange machine-processable semantics alongside the metadata and interpret this semantics correctly.
Recipe for harmonization

1. **Adopt a core model** with support for machine-processable semantics

2. **Construct mappings** of other standards that preserve semantics

[Diagram showing a core model at the top with semantics-preserving mappings to Standard A, Standard B, and Standard C]
Deconstructing the harmonization recipe

Core model

Vocabulary 1
Vocabulary 2
Vocabulary 3
Vocabulary 4
Vocabulary 5

Semantic embeddings

Standard A
Standard B
Standard C
Promising harmonization technology

- **Linked data** (Link **ALL** the data!)
- Mapping technology (GRDDL, RDB2RDF)
- RDFS conversions for existing standards
  - schema.org
  - RDA
  - etc
Conclusions & future directions

- Harmonization is not the same as interoperability
- Application profiles are useful for harmonization within specifications, but not between them
- Semantic metadata interoperability basis for harmonization
- Focus on abstract models
- Increased focus on harmonization in standardization activities
- Modularization of standards an important tool
- Harmonization as basis for Linked Data efforts
Various metadata standards (LOM, ISO MLR, DCAM, etc.)

Metadata formats (XML, N3, XHTML, etc.)

Profile Models (DSP etc.)

Vocabulary Models (SKOS etc.)

Application profiles

Element vocabularies

Value vocabularies

Ontologies

LOM AP

EdNA

UK LOM Core

Example APs

Example element vocabs

Example value vocabs

Core model

Specifications

Domain definitions

Ontology Models (OWL etc.)

DC

RDA

LOM

RDA Vocabs

LOM Vocabs

LOM

Vocabs

UK LOM Core

LCSH
Syntactic combinations (MODS & LOM)

```xml
<?xml version='1.0' encoding='UTF-8' ?>
<mods ...>
  <subject authority="lcsh">
    <topic>Parachuting</topic>
  </subject>
  <extension>
    <!-- LOM fragment: -->
    <lom:description>
      <lom:string lom:language="en">
        Useful for learning some flight-related French terminology.
      </lom:string>
      <lom:string language="sv">
        Användbar för att lära sig lite flygrelaterad fransk terminologi.
      </lom:string>
    </lom:description>
    <!-- End LOM fragment. -->
  </extension>
</mods>
```
Why syntactic combinations fail

<table>
<thead>
<tr>
<th>Base format</th>
<th>Extended with</th>
<th>Processable by LOM</th>
<th>Processable by MODS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOM XML</td>
<td>MODS</td>
<td>Only LOM part</td>
<td>No</td>
</tr>
<tr>
<td>MODS</td>
<td>LOM XML</td>
<td>No</td>
<td>Only MODS part</td>
</tr>
</tbody>
</table>

**Diagram:**

- **Application A (MODS)**
  - express/encode
  - transport MODS (XML)
- **Application B (MODS)**
  - interpret
  - No interpretation!
  - NO agreement
  - express
  - encode
  - insert LOM fragment
- **Application B (LOM)**
  - express
  - encode
  - insert LOM abstract syntax

**Description:**

- **LOM XML binding**
  - Only LOM part

**MODS (XML)**

**LOM XML**

**Base format**

**Extended with**

**Processable by LOM**

**Processable by MODS**
Successful combinations - RDF

This learning object explains parachuting.

Useful for learning some flight-related French terminology.
Semantics: the interpretation of data

- Makes metadata out of data

Three kinds

- **Informal**: human semantics (language) – “A is a kind of B”
- **Machine-processable**: semantics encoded as data – `<A> rdfs:subClassOf <B>`.
- **Formal**: logical semantics
  \[ \langle x, y \rangle \in \text{IEXT}(\text{l(rdfs:subClassOf)}) \Rightarrow \]
  \[ \text{x and y} \in \text{IC and ICEXT(x)} \subseteq \text{ICEXT(y)} \]