

Uniform Access to Interlinked (Digital Library) Sources

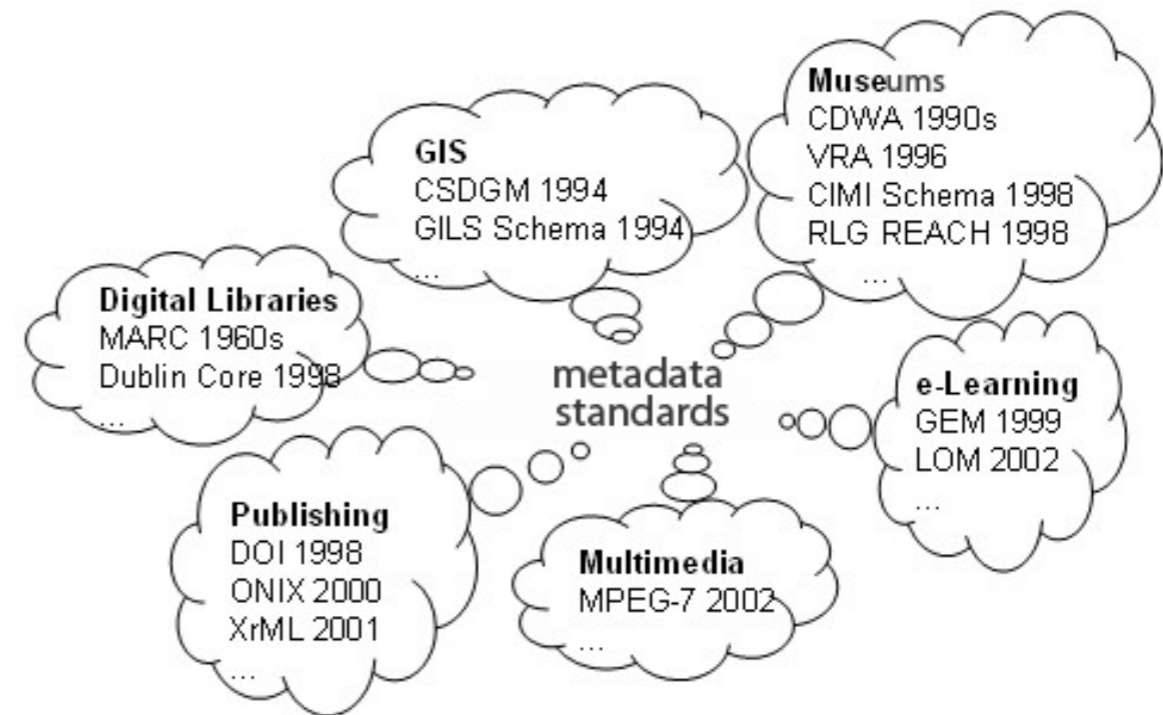
Bernhard Haslhofer
University of Vienna

Motivation

- Institutions use various incompatible metadata schemes*

- some are standardised

- many are proprietary



- How to establish uniform access to institutional metadata repositories?

* the terms schema / ontology / vocabulary are used interchangeably

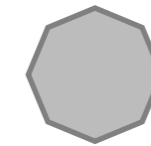
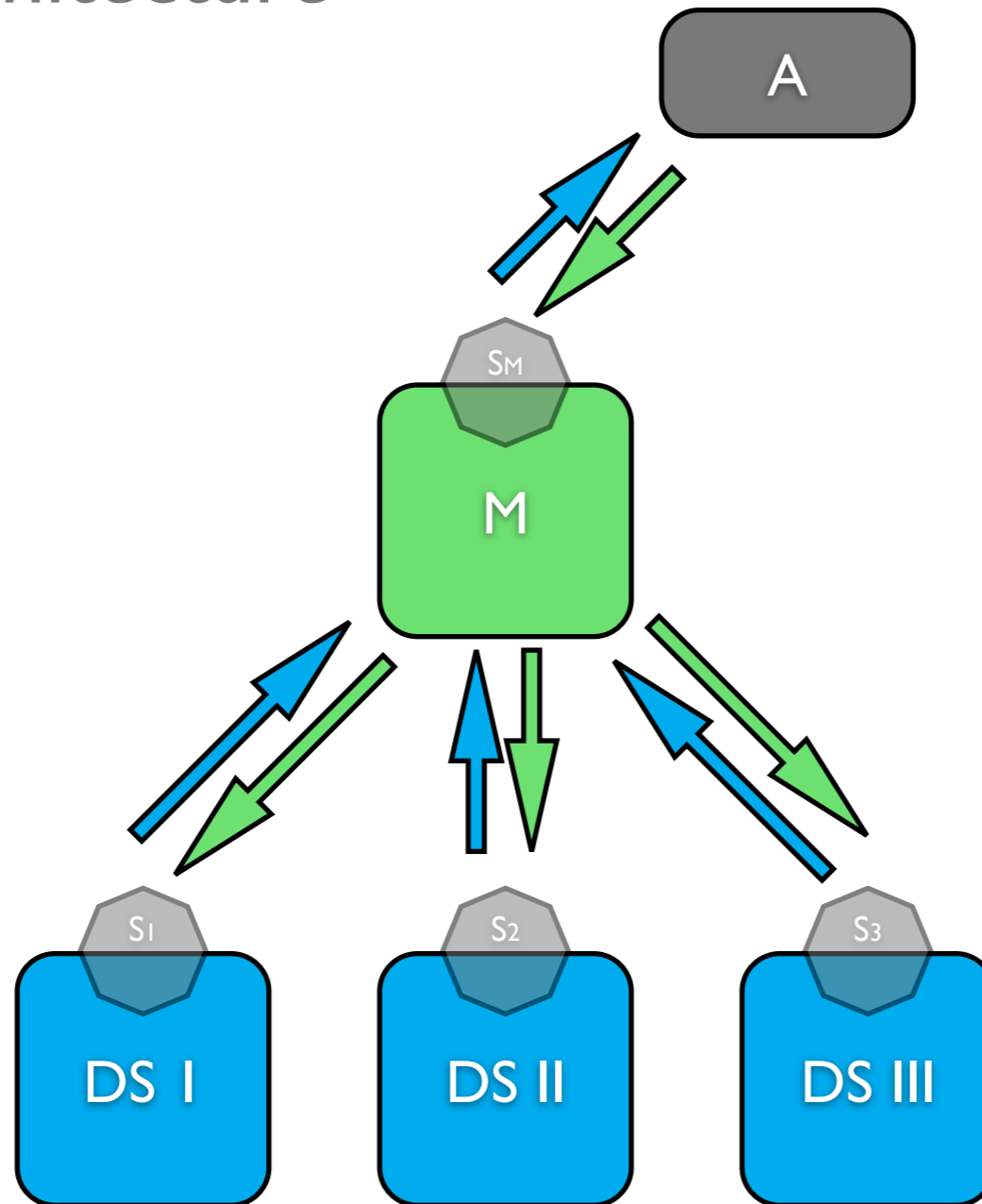
Motivation

- Problem:
 - structural and semantic conflicts among metadata schemes/elements
- Solution:
 - metadata mappings / crosswalks

MARC	DC
100 a Shakespeare William d 1564 - 1616	<dc:creator> Shakespeare, William, 1564 - 1616 </dc:creator>
245 a Hamlet	<dc:title> Hamlet </dc:title>
260 a New York b Penguin Books c 2003	<dc:publisher> Penguin Books </dc:publisher>
	<dc:date> 2003 </dc:date>

Approach

- Architecture



RDF/S,
OWL



SPARQL



RDF



Mediator



(Meta)Data
Source

Approach

- Workflow (high-level view)
 1. Application (A) formulates SPARQL query over a user selected mediation schema (SM)
 2. Mediator (M) reformulates query according to mappings / crosswalks into queries over the source schemes (Sx)
 3. Query plan calculation
 4. Mediator federates the queries to the data sources (DSx), collects the results and returns them to the user application (A)

Approach

- Query reformulation - **Step 1**
 - compile crosswalks into SPARQL query templates (assuming that DC is the mediation schema)

100 a Shakespeare William d 1564 - 1616

<dc:creator> Shakespeare, William, 1564 - 1616 </dc:creator>

```
CONSTRUCT { ?x dc:creator ?c }
WHERE {
  ?x marc:100a ?name
  ?x marc:100d ?date
  (?name ` , ` ?date) ext:concat ?c
}
```

Approach

- Query reformulation - **Step 2**
 - analyse **graph pattern** of input query

```
SELECT ?x
WHERE {
  ?x dc:creator ?name .
  FILTER {REGEX(?name,i,"Shakespeare" )}}
```

- choose and execute the matching templates

```
CONSTRUCT { ?x dc:creator ?c }
...
```

Related Issues

- Where are the SPARQL data sources?
- **Linking Open Data (LOD) Initiative** (<http://linkeddata.org/>)
 - exposing, sharing, and connecting data sources using URI, HTTP (and SPARQL)
 - Example Data Sources:
 - DBPedia (exposes Wikipedia data): 103 M triples
 - USCensus 2000 dataset: ~ 1 B triples
 - DBLP Bibliography: 15 M triples
- Up-to-date list of all available sources:
 - <http://esw.w3.org/topic/TaskForces/CommunityProjects/>

Related Issues

- Issue No. 1 when talking about SemWeb Technologies: “Triple Stores do not perform”
- Possibility #1: use a SPARQL-SQL translator
 - D2RQ Server: <http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/>
- Possibility #2: use a triple store
 - they are not yet perfect; but they are getting better
 - Large Triple Stores: <http://esw.w3.org/topic/LargeTripleStores>
 - AllegroGraph + Virtuoso (1 Bn), others 70 Mn - 650 Mn

Status & Future Work

- OAI2LOD Server v.0.1 - ready
 - exposes any OAI-PMH compliant repository as LOD
 - Demo: <http://mediaspaces.mminf.univie.ac.at:2020/>
- Online mapping platform - work in progress
 - choose data sources → retrieve schema info
 - create and deploy reusable mappings on the Web

Contact

- bernhard.haslhofer@univie.ac.at

Motivation

- Current implementation of crosswalks: **XSLT**
 1. fetch original metadata record from source repository
 2. apply XSLT transformation using an XSL style sheet
 3. deliver transformed metadata record in target format
- This is fine for metadata **aggregation** but not for **virtual integration**
 - aggregation = copy / duplicate (and transform) metadata to a central repository
 - virtual integration = leave metadata in their repositories