A FRAMEWORK FOR ONTOLOGY-BASED LIBRARY DATA GENERATION, ACCESS AND EXPLOITATION

Daniel Vila Suero

Advisors:
Prof. Dr. Asunción Gómez-Pérez and Dr. Jorge Gracia del Río

PhD in Artificial Intelligence,
Defense, Madrid, 27th of July 2016
Library records describing **persons, concepts, works**, and other **entities**.

**MARC 21 international standard** describing **millions** of records.
(e.g., European Library +80 million records)
CATALOGUE RESOURCES ARE DISCONNECTED FROM THE WEB

Online catalogues

Web forms
HTML

WWW

Specialized users
Applications
Search engines
Semantic web and Linked Data technologies (Ontologies, RDF, HTTP URIs):

“Creating and delivering library data while providing a natural extension to the collaborative sharing models historically employed by libraries.”
OUTLINE

• Problem and state of the art
• Research methodology and hypotheses
• Contributions
  – Mapping
  – Ontology development
  – Library applications
• Conclusion and future directions
INTRODUCTION

RESEARCH PROBLEM

PRINCIPLED TRANSFORMATION OF LIBRARY CATALOGUES INTO ONTOLOGY-BASED DATA AND THEIR PUBLICATION AND CONSUMPTION ON THE WEB.
INTRODUCTION

THE PROBLEM: **GLOBAL-AS-VIEW** ONTOLOGY-BASED DATA INTEGRATION

**SOURCE SCHEMA**

Data sources (MARC 21)

**GLOBAL SCHEMA**

Target (RDF Linked Data)

**MAPPING**

creator of

same as

topic

**CHALLENGES**

1. Library records present a **nested structure** (fields, subfields, etc.)
2. Library data sources lack **schema and query mechanisms**
3. Library standards and ontologies are **highly specialized**
INTRODUCTION

RESEARCH PROBLEM AND AREAS

PRINCIPLED TRANSFORMATION OF LIBRARY CATALOGUES INTO ONTOLOGY-BASED DATA AND THEIR PUBLICATION AND CONSUMPTION ON THE WEB.

TRANSFORMATION & MAPPING LANGUAGES

ONTOLOGY DEVELOPMENT

LIBRARY APPLICATIONS
STATE OF THE ART

TRANSFORMATION METHODS

1998

FRBR model

Experiments FRBR

[1998]

Applications of RDF

[1998]

Linked Data

[1998]

[1998]

[Hickey2002]

[Bennet2003]

[Hegna2003]

[Aalberg2006]

[Manguinhas2010]

[Takhirov2012]

[Simon2013]

[marc2bibframe2015]
INTRODUCTION

STATE OF THE ART

TRANSFORMATION METHODS

MAPPING LANGUAGES

FRBR model | Experiments FRBR | Applications of RDF | Linked Data
---|---|---|---
1998 | 2012

Relational databases into RDF | R2RML W3C standard | Extensions of R2RML for non-RDB data sources
---|---|---

[Barrasa04] [Bizer04] [Dimou2014] [Michel2015] [Slepicka2015]
INTRODUCTION

STATE OF THE ART

TRANSFORMATION METHODS

1998
FRBR model
Experiments FRBR
Applications of RDF
Linked Data

2012

MAPPING LANGUAGES

2004
Relational databases into RDF
R2RML W3C standard
Extensions of R2RML for non-RDB data sources

2014

ONTOMETRY DEVELOPMENT

1995
Initial methodologies
[Grüninger1995]
[Fernández-López1997]

2015
Ontology lifecycle models
[Staab2001]
[Pinto2004]
Neon methodology
[Suarez-Figueroa2015]

Agile approaches
[Auer2006]
[Presutti2012]
LIMITATION 1: MAPPING LANGUAGE AND METHODS

Lack of **interoperability** and reuse of mapping rules.
LIMITATION 1: MAPPING LANGUAGE AND METHODS

Lack of interoperability and reuse of mapping rules.

Format-dependent (e.g., JSONPath) [RML] [xR2RML]

Lack of queries over nested data [KR2RML]
LIMITATION 2: FEEDBACK OF LIBRARY EXPERTS

Lack of interoperability and reuse of mapping rules.

Feedback of library experts not integrated within the transformation process.
[Takhirov2012]

Format-dependent (e.g., JSONPath)

Lack of queries over nested data
INTRODUCTION

LIMITATION 2: FEEDBACK OF LIBRARY EXPERTS

Lack of interoperability and reuse of mapping rules.

Format-dependent (e.g., JSONPath)

Limited technical support for the active participation of domain experts. [NeOn]

Feedback of library experts not integrated within the transformation process. [Takhirov2012]

Lack of queries over nested data

Limited methods for understanding similarities among overlapping ontologies. [Vandenbussche2014]
LIMITATION 3: EXPERIMENTAL RESULTS

Lack of interoperability and reuse of mapping rules.

Format-dependent (e.g., JSONPath)

Limited technical support for the active participation of domain experts.

Feedback of library experts not integrated within the transformation process.

Lack of queries over nested data

Limited methods for understanding similarities among overlapping ontologies.

Lack of evaluation and experimental results of impact in end-user library applications [Simon2013]
HYPOTHESES

Research methodology, and hypotheses
HYPOTHESES

DESIGN-SCIENCE RESEARCH METHODOLOGY

RESEARCH PROBLEMS

P1 Mapping

P2 Ontology development process

P3 Library applications

HYPOTHESES

H1

H2

H3

H4

H5

EVALUATION

ARTIFACT: MARIMBA FRAMEWORK

CONTRIBUTIONS

EVALUATE HYPOTHESES

REFINEMENT

METHODS MODELS INSTANTIATIONS
MARIMBA FRAMEWORK: PROBLEMS AND CONTRIBUTIONS

Library data sources
- marimba-mapping
- marimba-rml
- marimba-datamodel

Ontology-based Library data (RDF)
- marimba-topicmodel
- BNE ontology
- marimba-modelling

Library online applications
- P1 Mapping
- P2 Ontology development process
- P3 Library applications

Empirical studies
datos.bne.es
P1: DEFINITION OF INTEROPERABLE MAPPING RULES

H1
The structure of MARC 21 records can be fully represented using a nested relational model.

H2
Minimal modifications to the W3C R2RML language to define mapping rules between MARC 21 data sources into RDF.

RESTRICTIONS:
No query optimization, only operational semantics of mapping language, authority and bibliographic formats, MARC 21 standard
HYPOTHESES

P2: LIBRARY ONTOLOGY DEVELOPMENT

H3
Analytical data and the feedback of library experts can be used to develop a library ontology with sufficient quality.

H4
Probabilistic topic modelling techniques can produce coherent descriptions of ontologies and can perform better than existing methods used in ontology search.

ASSUMPTIONS:
Participation of library experts, available library ontologies and records
P3: APPLICATION OF SEMANTIC TECHNOLOGIES IN LIBRARY APPLICATIONS

The application of semantic technologies to end-user library applications can increase user satisfaction and efficiency for finding information.

Restrictions:
Transformation in a batch process
OUTLINE

P1
Mapping

P2
Ontology development process

P3
Library applications

Library data sources
Ontology-based Library data (RDF)
Library online applications

marimba-mapping
marimba-rml
marimba-datamodel
marimba-topicmodel
BNE ontology
marimba-modelling
datos.bne.es
Empirical studies
marimba-modelling
marimba-mapping
marimba-topicmodel
marimba-modelling
Data source (MARC 21)

GLOBAL SCHEMA
- creator of
- same as
- topic

SOURCE SCHEMA

GLOBAL SCHEMA

Mapping

P1

Data source (MARC 21)
MAPPING

Source Schema

Data source (MARC 21)

Global Schema

creator of
same as
topic
Target source (RDF)

Schema extraction
marimba-datamodel

SOURCE SCHEMA

GLOBAL SCHEMA

marimba-mapping
marimba-rml

P1
Mapping
– Study of **MARC 21 standard in Nested Relational Model** (Makinouchi [1977]).

– Attributes:
  - **Atomic**: e.g., Field 001
  - **Relation-valued**: e.g., Field 100, 008

– **Complete coverage** of MARC 21 standard elements

<table>
<thead>
<tr>
<th>MARC 21</th>
<th>marimba-datamodel (NRM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td>Relation scheme $R(S)$</td>
</tr>
<tr>
<td><strong>Leader</strong></td>
<td>relation-valued attribute $R_{leader}(S_{leader}) \in S$ with $S_{leader} = (p00, p01, p02, ..., p23)$</td>
</tr>
<tr>
<td><strong>Control Field (fixed-length)</strong></td>
<td>relation-valued attribute $R_{cf}(S_{cf}) \in S$ with $S_{cf} = (p00, p01, p02, ..., pN)$ for $N$ positions</td>
</tr>
<tr>
<td><strong>Control Field (single-valued)</strong></td>
<td>Atomic attribute $cf \in S$</td>
</tr>
<tr>
<td><strong>Data Field</strong></td>
<td>relation-valued attribute $R_{df}(S_{df}) \in S$ with $S_{cf} = (i1, i2, sf1, ..., sfN)$ with $i1$ and $i2$ indicators and $sf1..sfN$ the N subfields of the data field</td>
</tr>
<tr>
<td><strong>Subfield</strong></td>
<td>Atomic attribute $sf$ in a relation scheme of a data field $\in S_{df}$</td>
</tr>
<tr>
<td><strong>Indicators</strong></td>
<td>Atomic attributes $i1$ and $i2$ in a relation scheme of a data field $\in S_{df}$</td>
</tr>
</tbody>
</table>
Novel **algorithm for schema extraction** of MARC 21 records.

- Support the **validation of queries to MARC 21 data**.
- Support the generation of **mapping templates for library experts**.
- **Complete representation** of patterns in existing MARC 21 data sources.

```
SCHEMA

TABLE bibliographic

ITEM f001 UNIQUE NOT NULL /* C.field Unique ID */

ITEM TABLE leader /* Leader Info about language, etc.*/

ITEM p01
...
ITEM p23

ITEM TABLE f100 /* Data field Author information */

ITEM a /* Subfield */
...
ITEM i0 /* Indicator */
```
Schema extraction
marimba-datamodel

SOURCE SCHEMA

Data source
(MARC 21)

GLOBAL SCHEMA

creator of
global schema

same as
topic

Target source
(RDF)

R2RML MAPPING

P1
Mapping

marimba-mapping
marimba-rml
Data source (MARC 21)
Minimal query language based on SQL/NF (Roth et al. [1987]):
- **Conciseness**
- **Orthogonality** of expressions → SFW nested expressions.

```plaintext
/* Selects bibliographic records of type Drawing*/
bibliographic  / * SELECT FROM * /
WHERE
EXISTS
  / * NESTED EXPRESSION* /
(leader WHERE p6 = "k" AND p7 != "s")
AND
  / * NESTED EXPRESSION * /
EXISTS (f007 WHERE p2 = "d")
```

*BNF grammar provided in Annex I*