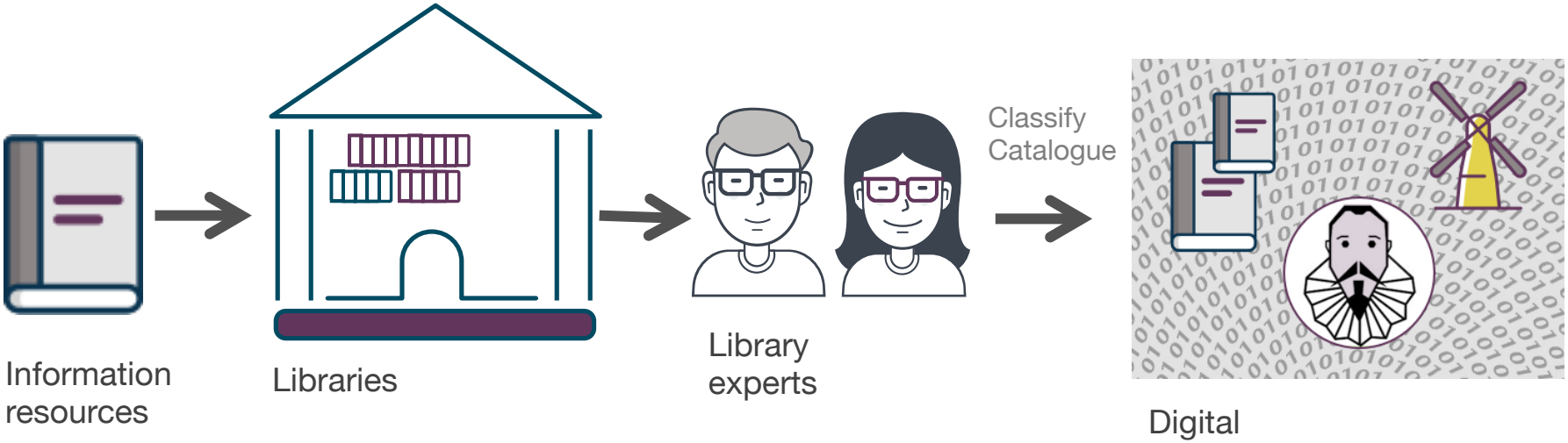
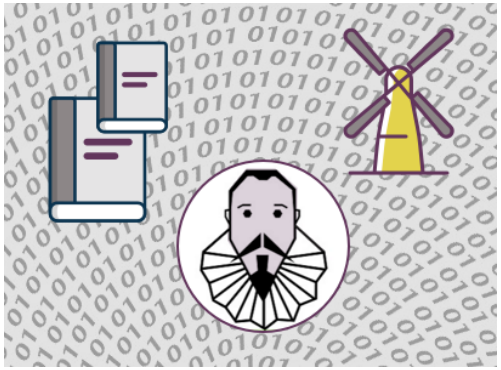
A FRAMEWORK FOR ONTOLOGY-BASED LIBRARY DATA GENERATION, ACCESS AND EXPLOITATION

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

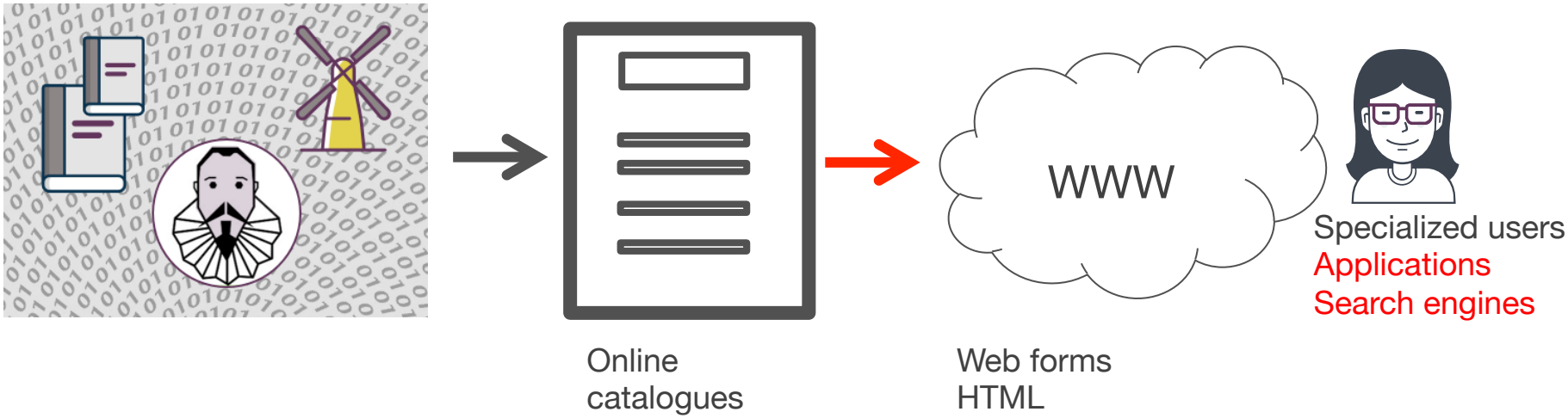




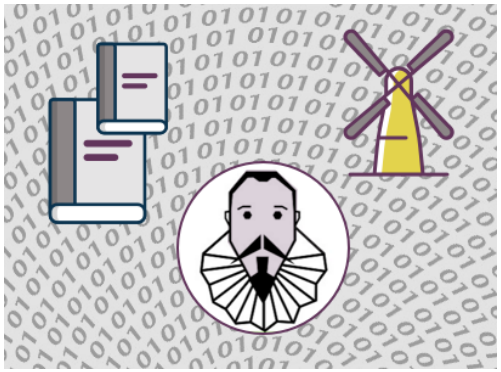
Library records describing **persons**, **concepts**, **works**, and other **entities**.

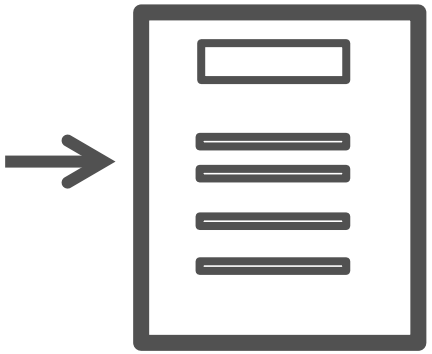
MARC 21 international standard describing **millions of records.**
(e.g., European Library +80 million records)

CATALOGUE RESOURCES ARE DISCONNECTED FROM THE WEB



CATALOGUE RESOURCES ARE DISCONNECTED FROM THE WEB

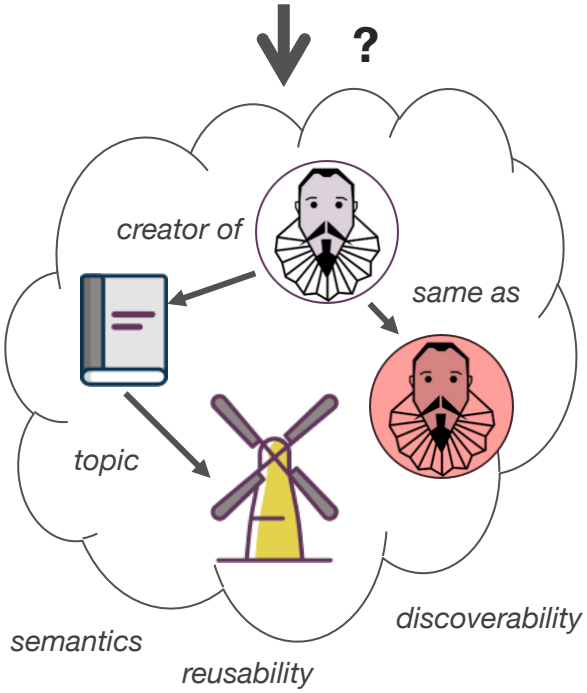




Online
catalogues

Web forms
HTML

Specialized users
Applications
Search engines



Semantic web and Linked Data technologies (**Ontologies**, **RDF**, **HTTP URIs**):

My thesis

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

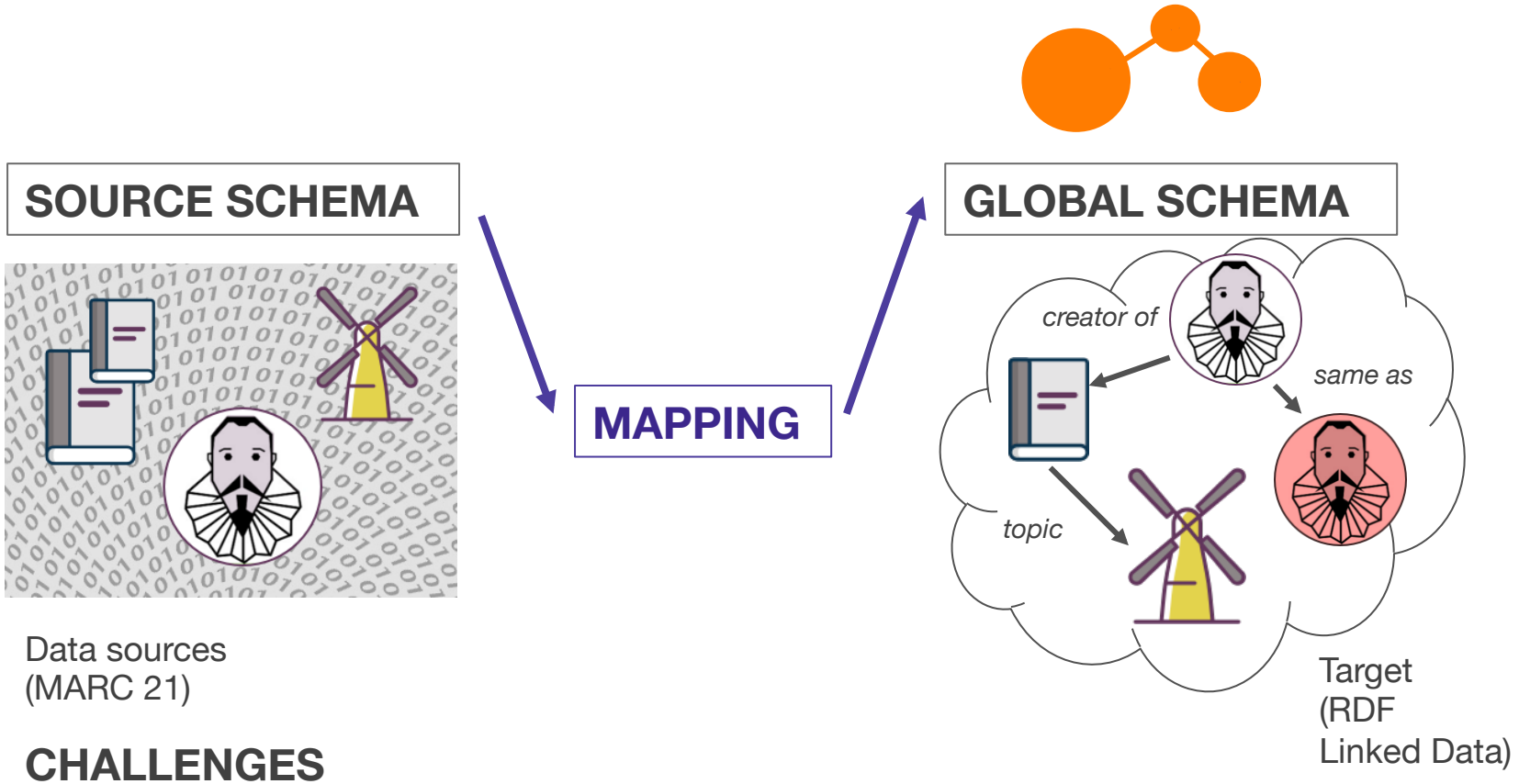
OUTLINE

- **Problem and state of the art**
- Research methodology and hypotheses
- Contributions
  - Mapping
  - Ontology development
  - Library applications
- Conclusion and future directions

RESEARCH PROBLEM

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

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



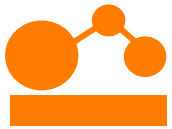
CHALLENGES

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**

RESEARCH PROBLEM AND AREAS

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

TRANSFORMATION & MAPPING LANGUAGES

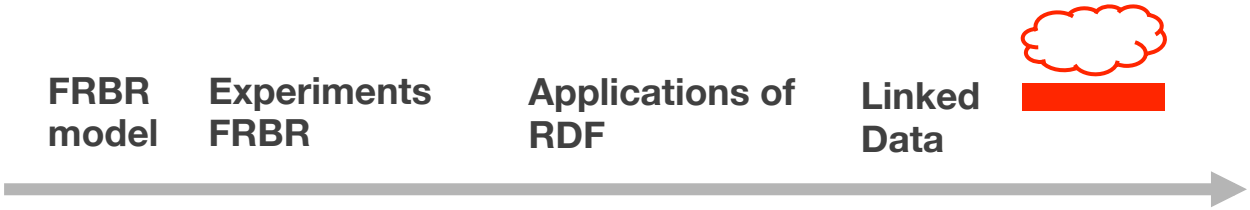


ONTOLOGY DEVELOPMENT

LIBRARY APPLICATIONS

STATE OF THE ART

TRANSFORMATION METHODS



STATE OF THE ART

TRANSFORMATION METHODS

FRBR model

Experiments FRBR

Applications of RDF

Linked Data

1998

2012

MAPPING LANGUAGES

Relational databases into RDF

R2RML W3C standard

Extensions of R2RML for non-RDB data sources

2004 [Barrasa04]
[Bizer04]

2014 RML [Dimou2014]
xR2RML [Michel2015]
KR2RML [Slepicka2015]

STATE OF THE ART

TRANSFORMATION METHODS

FRBR model (1998) → Experiments FRBR → Applications of RDF → Linked Data (2012)

MAPPING LANGUAGES

Relational databases into RDF (2004) → R2RML W3C standard → Extensions of R2RML for non-RDB data sources (2014)

ONTOLOGY DEVELOPMENT

Initial methodologies (1995)
[Grüninger1995]
[Fernández-López1997]

Ontology life-cycle models
[Staab2001]
[Pinto2004]
Neon methodology
[Suarez-Figueroa2015]

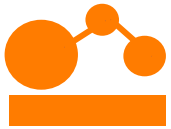
Agile approaches
[Auer2006]
[Presutti2012]

2015

LIMITATION 1: **MAPPING LANGUAGE AND METHODS**

TRANSFORMATION

MAPPING LANGUAGES



ONTOLOGY DEVELOPMENT

LIBRARY
LINKED DATA
APPLICATIONS

Lack of **interoperability** and reuse of mapping rules.

LIMITATION 1: **MAPPING LANGUAGE AND METHODS**

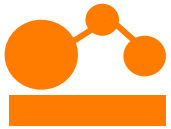
TRANSFORMATION

Lack of **interoperability** and reuse of mapping rules.

MAPPING LANGUAGES

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

Lack of **queries** over nested data [KR2RML]



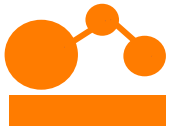
ONTOLOGY DEVELOPMENT

LIBRARY
LINKED DATA
APPLICATIONS

LIMITATION 2: **FEEDBACK OF LIBRARY EXPERTS**

TRANSFORMATION

MAPPING LANGUAGES



ONTOLOGY DEVELOPMENT

LIBRARY LINKED DATA APPLICATIONS

Lack of **interoperability** and reuse of mapping rules.

Format-dependent (e.g., JSONPath)

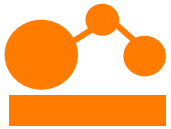
Feedback of library experts not integrated within the transformation process. [Takhirov2012]

Lack of **queries** over nested data

LIMITATION 2: **FEEDBACK OF LIBRARY EXPERTS**

TRANSFORMATION

MAPPING LANGUAGES



ONTOLOGY DEVELOPMENT

LIBRARY LINKED DATA APPLICATIONS

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**

TRANSFORMATION

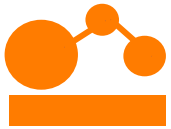
Lack of **interoperability** and reuse of mapping rules.

Feedback of library experts not integrated within the transformation process.

MAPPING LANGUAGES

Format-dependent (e.g., JSONPath)

Lack of **queries** over nested data



ONTOLOGY DEVELOPMENT

Limited **technical support for the active participation** of domain experts.

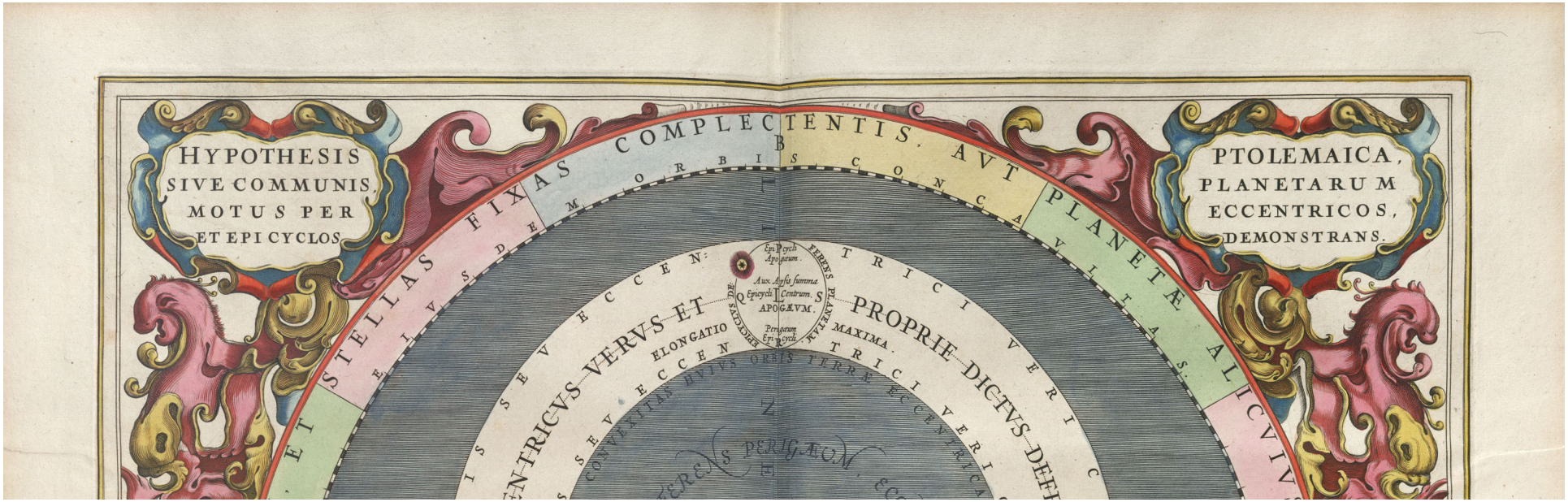
Limited **methods for understanding similarities** among overlapping ontologies.

LIBRARY LINKED DATA APPLICATIONS

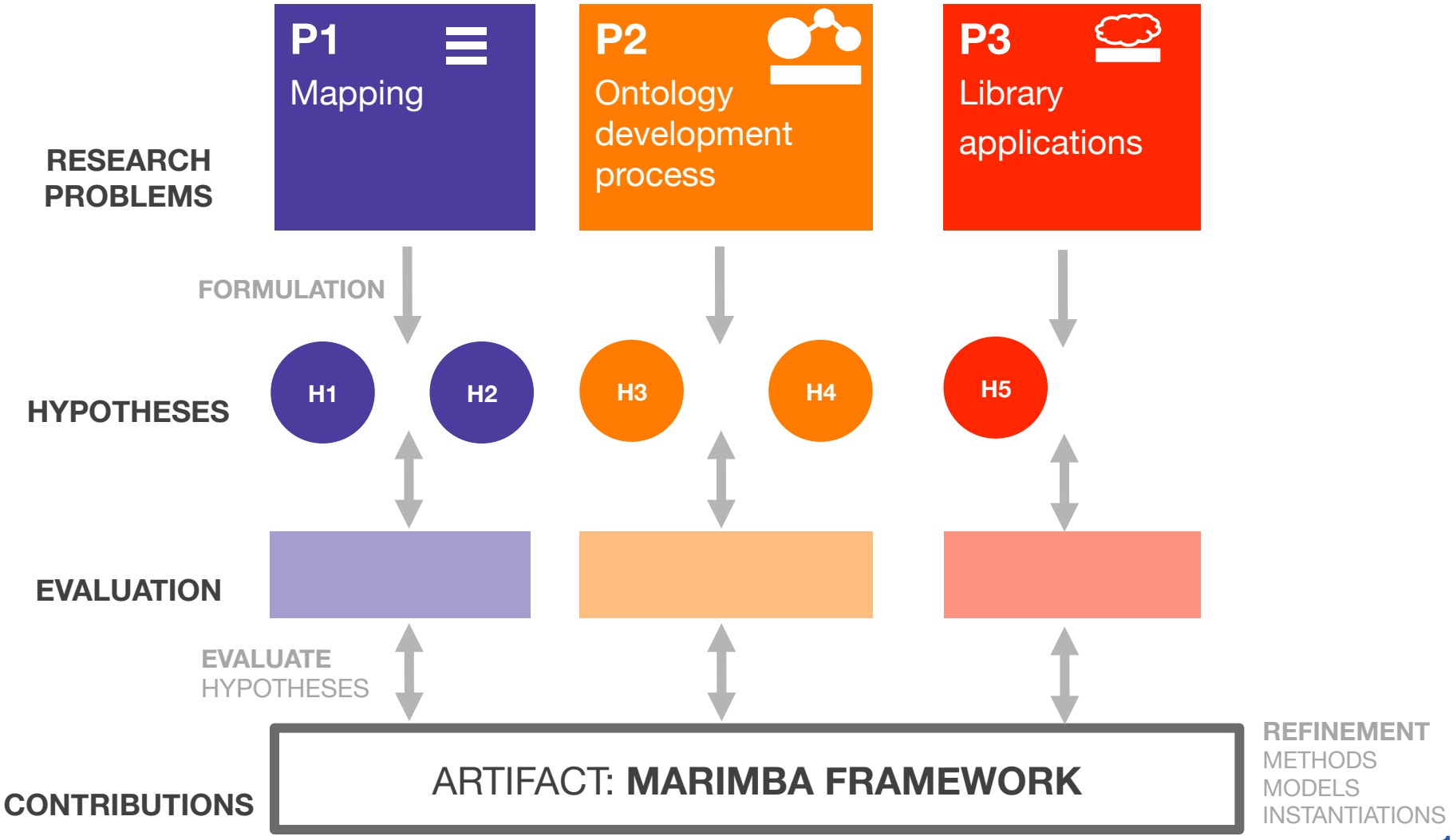
Lack of **evaluation and experimental results** of impact in end-user library applications [Simon2013]

HYPOTHESES

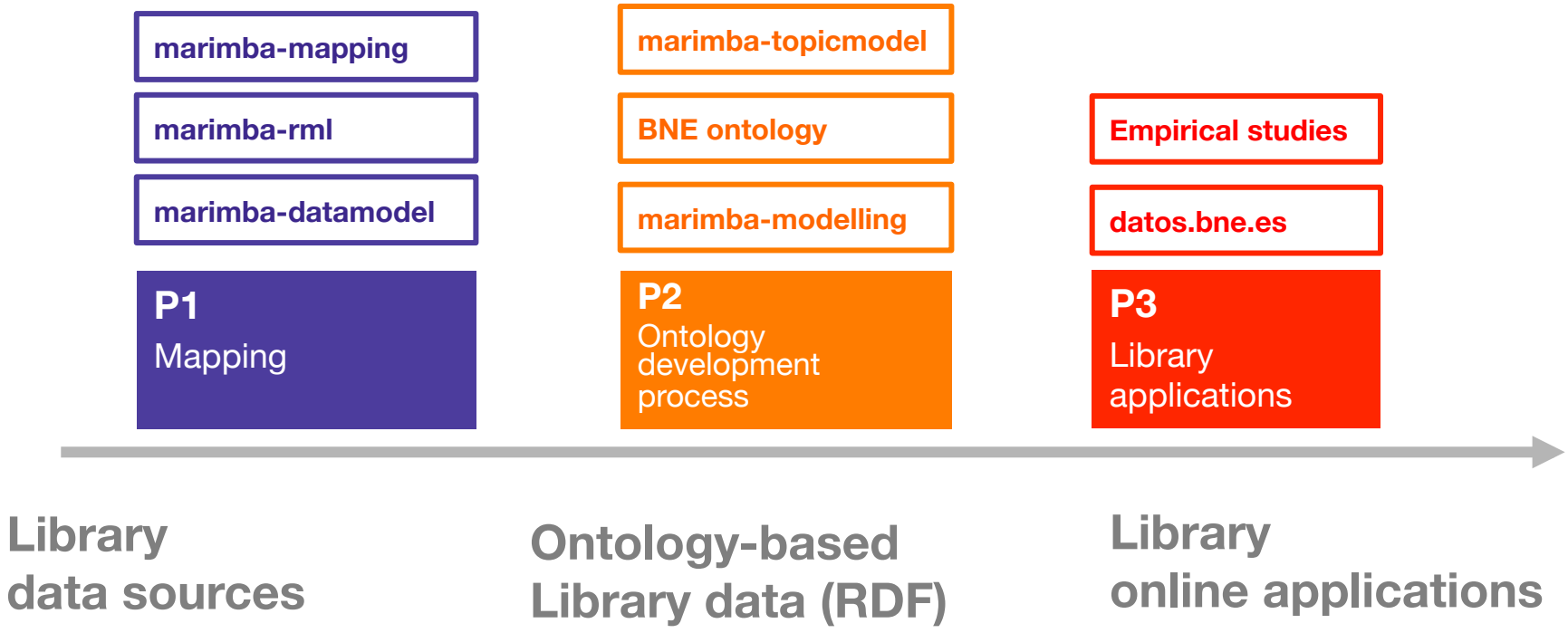
Research methodology, and hypotheses



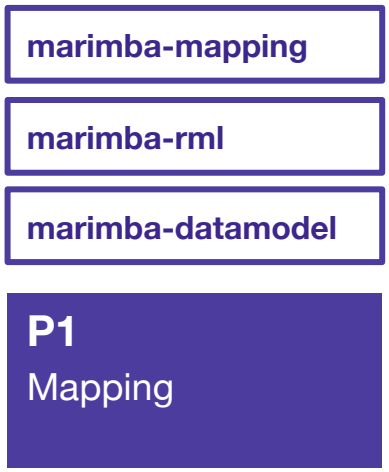
DESIGN-SCIENCE RESEARCH METHODOLOGY



MARIMBA FRAMEWORK: **PROBLEMS AND CONTRIBUTIONS**



P1: DEFINITION OF INTEROPERABLE MAPPING RULES



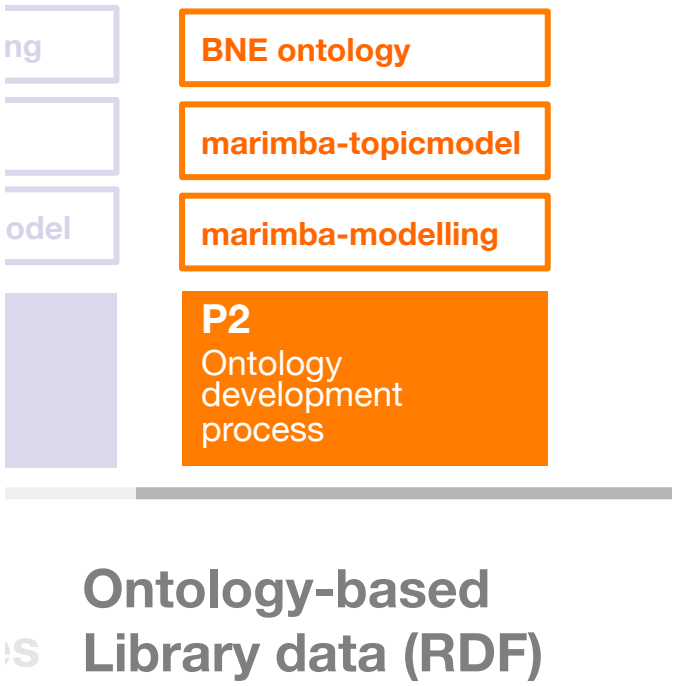
H1 The structure of MARC 21 records can be **fully represented** using a **nested relational model.**

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

Library data sources

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

P2: LIBRARY ONTOLOGY DEVELOPMENT

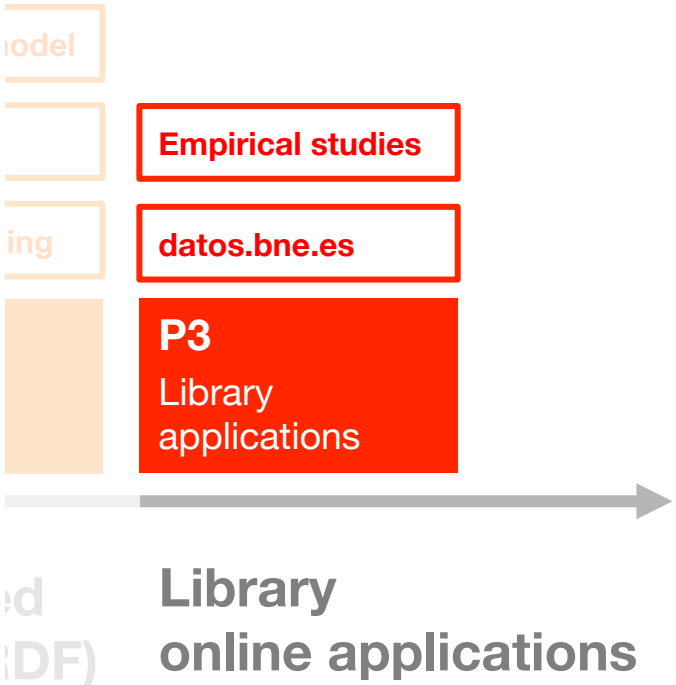


H3 Analytical data and the **feedback of library experts** can be used to develop a library **ontology** with sufficient quality

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



H5 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

marimba-mapping

marimba-rml

marimba-datamodel

P1
Mapping

marimba-topicmodel

BNE ontology

marimba-modelling

P2
Ontology development process

Empirical studies

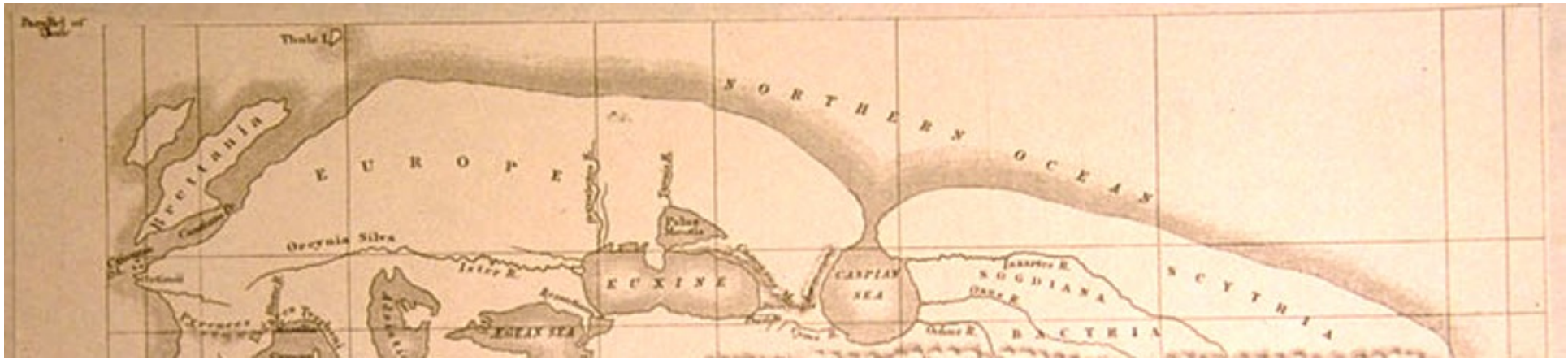
datos.bne.es

P3
Library applications

Library data sources

Ontology-based Library data (RDF)

Library online applications



MAPPING

marimba-datamodel

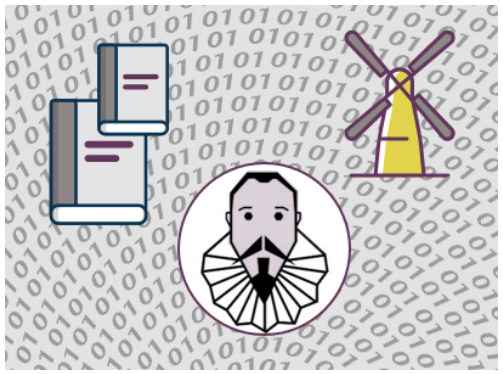
marimba-mapping

marimba-rml

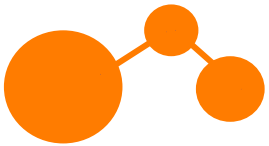
P1
Mapping

?

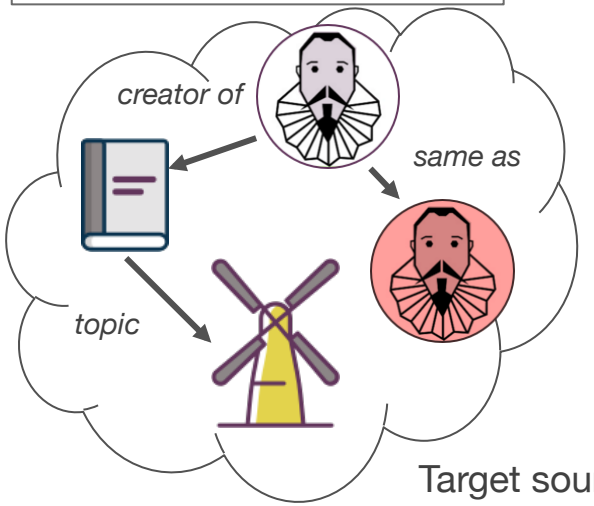
SOURCE SCHEMA



Data source
(MARC 21)



GLOBAL SCHEMA



Target source
(RDF)

marimba-mapping

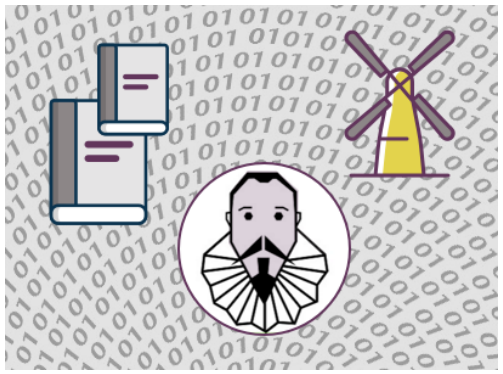
marimba-rml

P1
Mapping

Schema extraction

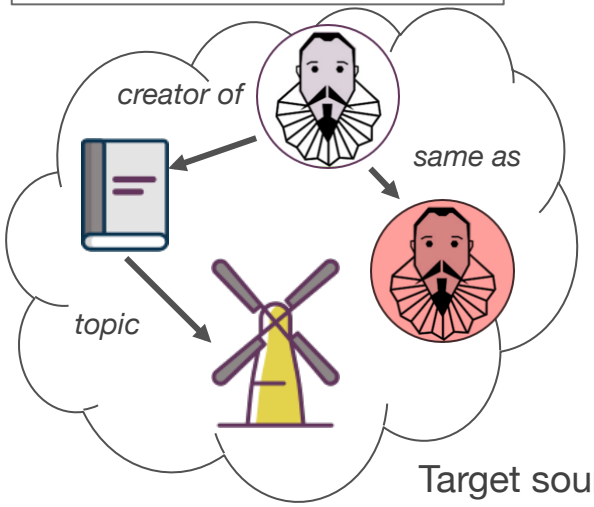
marimba-datamodel

SOURCE SCHEMA



Data source
(MARC 21)

GLOBAL SCHEMA



Target source
(RDF)

marimba-datamodel | marimba-rml | marimba-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

H1

MARC 21	marimba-datamodel (NRM)
Format	Relation scheme $R(S)$
Leader	relation-valued attribute $R_{leader}(S_{leader}) \in S$ with $S_{leader} = (p00, p01, p02, ..., p23)$
Control Field (fixed-length)	relation-valued attribute $R_{cf}(S_{cf}) \in S$ with $S_{cf} = (p00, p01, p02, ..., pN)$ for N positions
Control Field (single-valued)	Atomic attribute $cf \in S$
Data Field	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
Subfield	Atomic attribute sf in a relation scheme of a data field $\in S_{df}$
Indicators	Atomic attributes $i1$ and $i2$ in a relation scheme of a data field $\in S_{df}$

marimba-datamodel | marimba-rml | marimba-mapping

Schema extraction

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 */
```

marimba-mapping

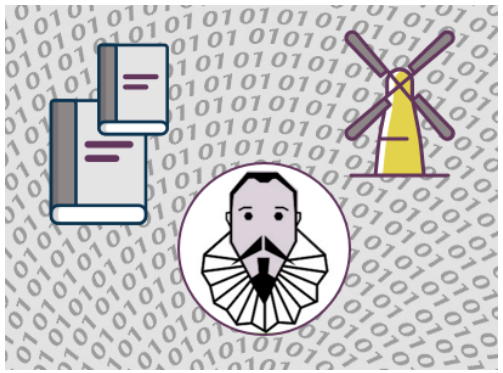
marimba-rml

P1
Mapping

Schema extraction

marimba-datamodel

SOURCE SCHEMA

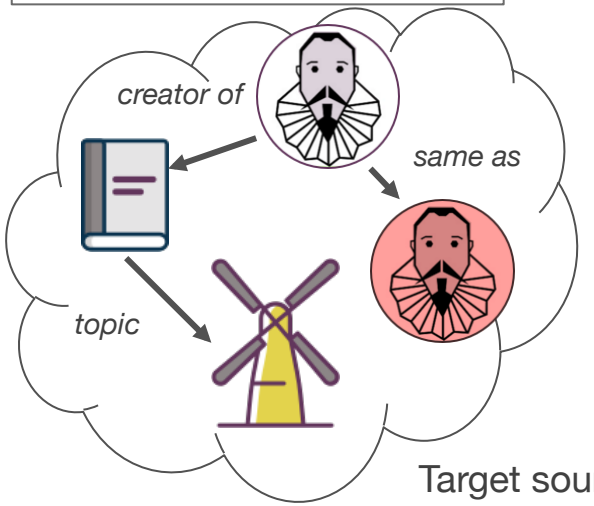


Data source
(MARC 21)

R2RML MAPPING

?

GLOBAL SCHEMA



Target source
(RDF)

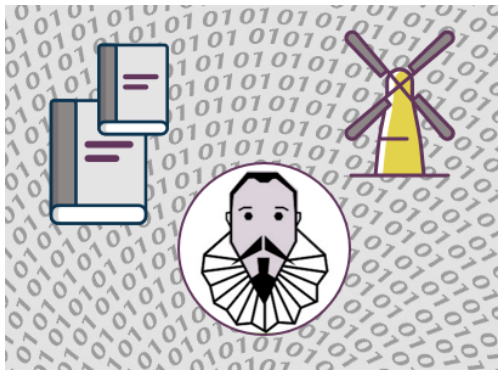
marimba-mapping

P1
Mapping

Schema extraction

marimba-datamodel

SOURCE SCHEMA



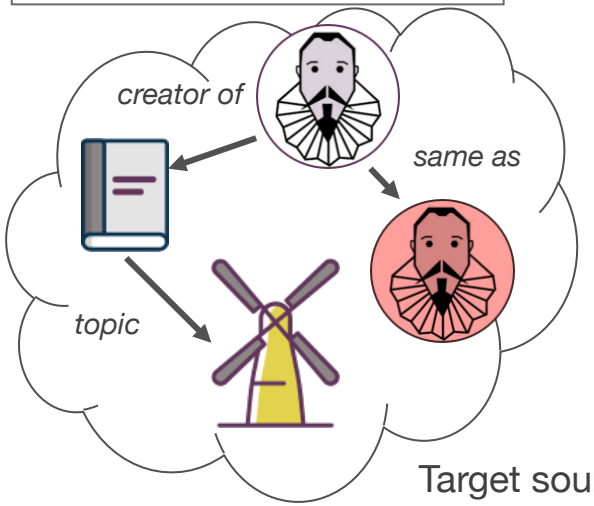
Data source
(MARC 21)

R2RML MAPPING

marimba-sql

marimba-rml

GLOBAL SCHEMA



Target source
(RDF)

marimba-datamodel | marimba-rml | marimba-mapping

marimba-sql

Minimal **query language** based on SQL/NF (Roth et al. [1987]):

- **Conciseness**
- **Orthogonality** of expressions → SFW nested expressions.

```
/* 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