MAPPING

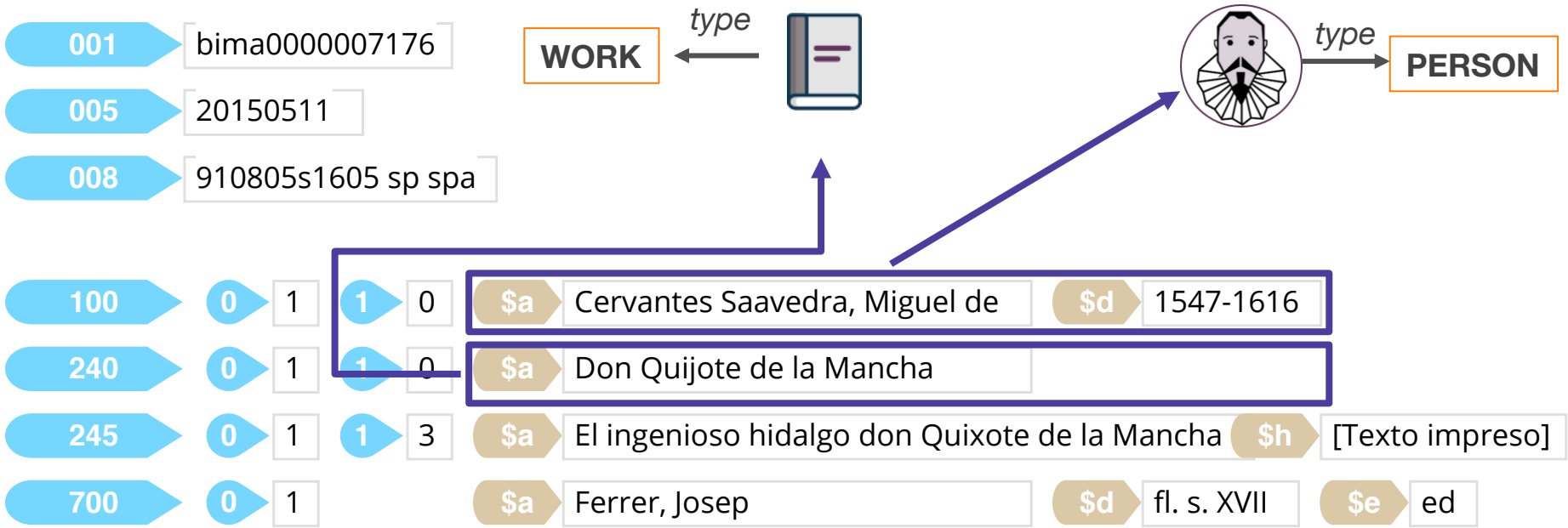
marimba-datamodel | **marimba-rml** | marimba-mapping

001	bima0000007176							
005	20150511							
008	910805s1605 sp spa							
100	0 1	1 0	$a	Cervantes Saavedra, Miguel de	$d	1547-1616		
240	0 1	1 0	$a	Don Quijote de la Mancha				
245	0 1	1 3	$a	El ingenioso hidalgo don Quixote de la Mancha	$h	[Texto impreso]		
700	0 1		$a	Ferrer, Josep	$d	fl. s. XVII	$e	ed

.

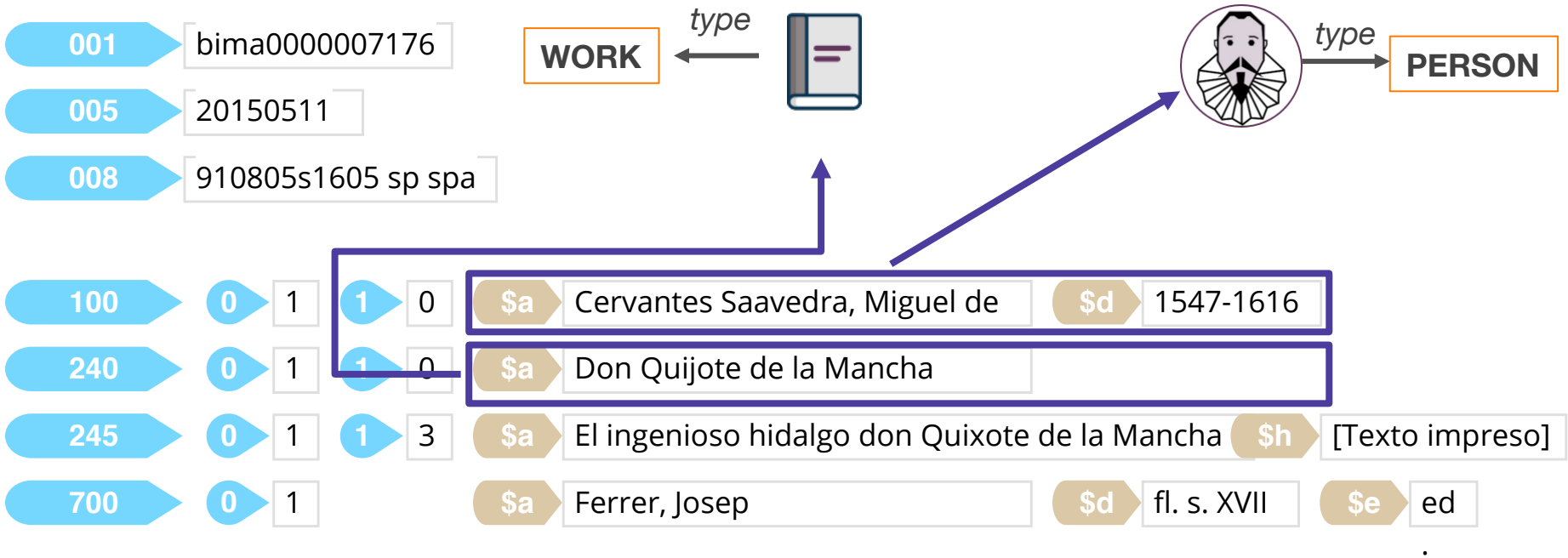
MAPPING

marimba-datamodel | **marimba-rml** | marimba-mapping



MAPPING

marimba-datamodel | **marimba-rml** | marimba-mapping



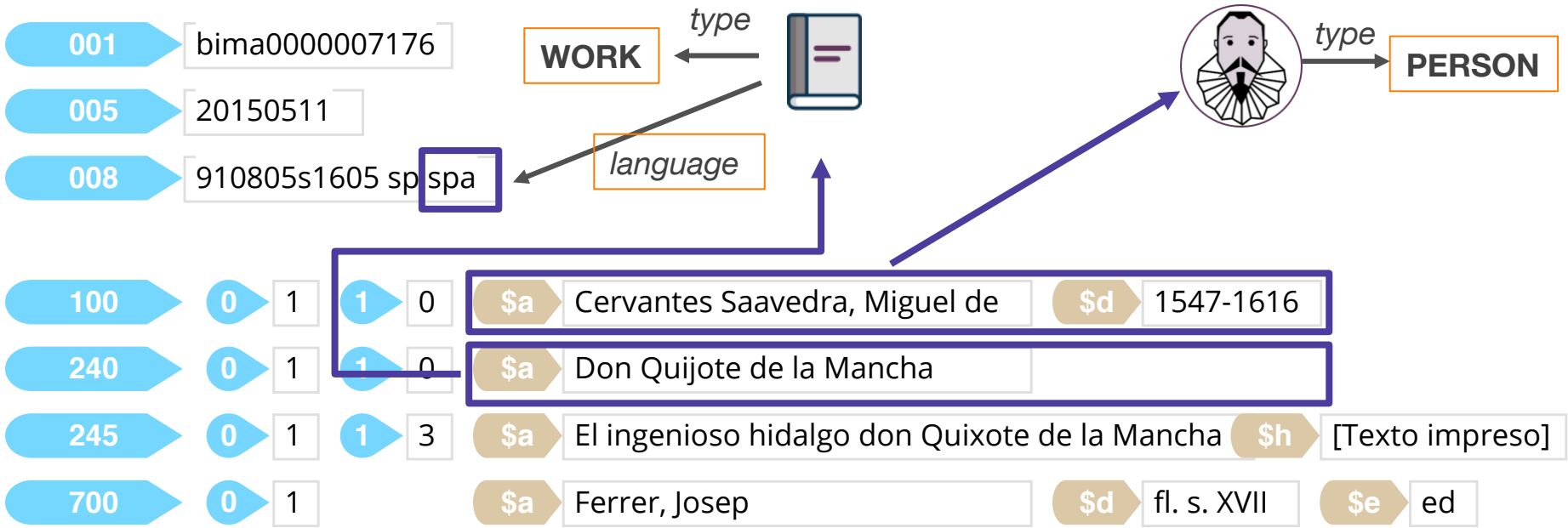
First modification of R2RML (*R2RML views)*

rr:sqlVersion → marimba-sql IRI

marimba-sql queries *for processing data in nested relations.*

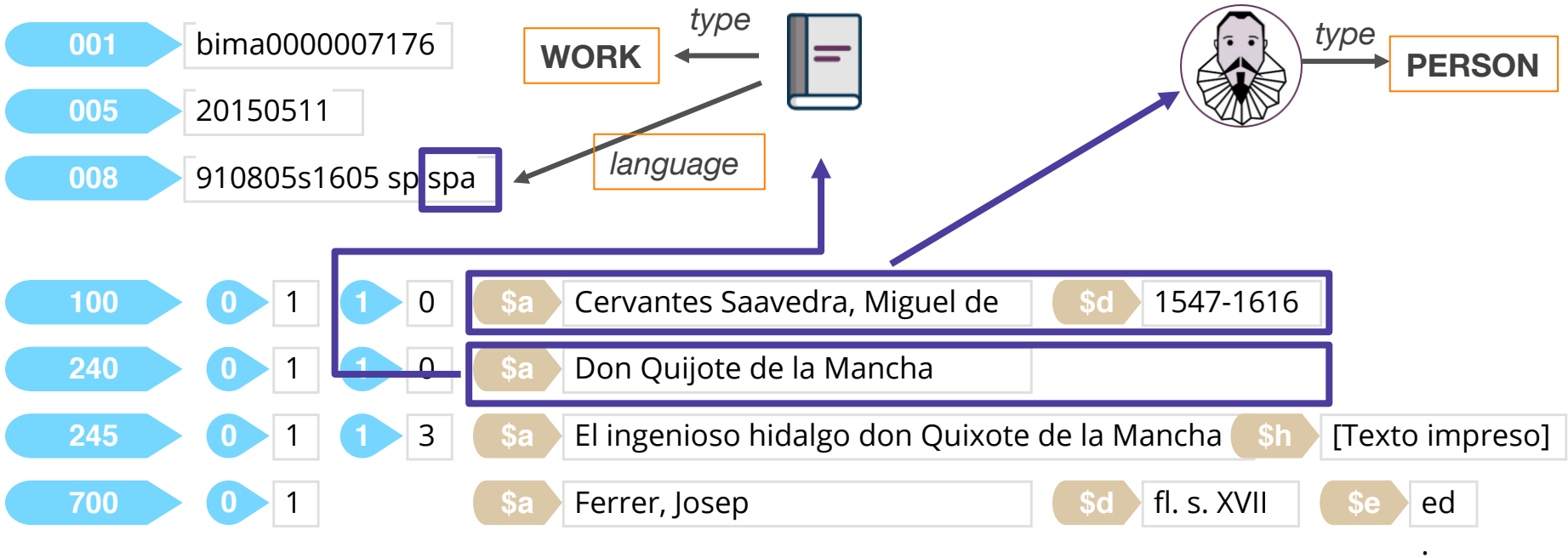
MAPPING

marimba-datamodel | **marimba-rml** | marimba-mapping



MAPPING

marimba-datamodel | **marimba-rml** | marimba-mapping



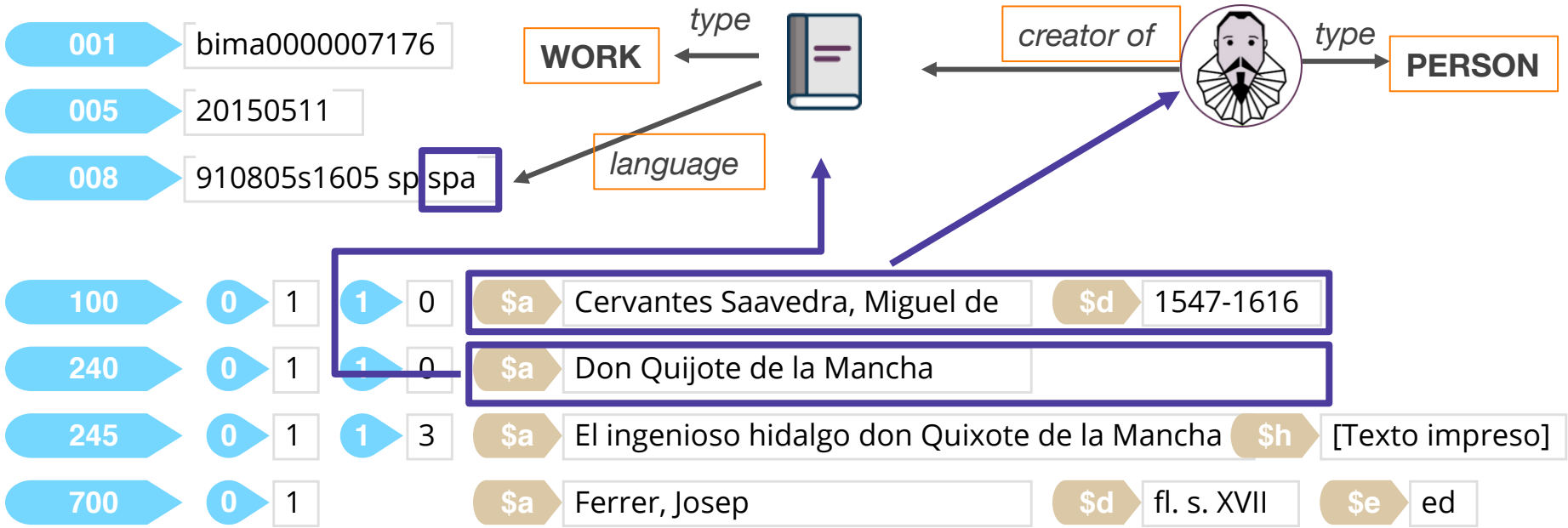
Second modification of R2RML *(rr:template and rr:column)*

R2RML column → marimba-sql query

*Generate **values** from attributes in nested relations.*

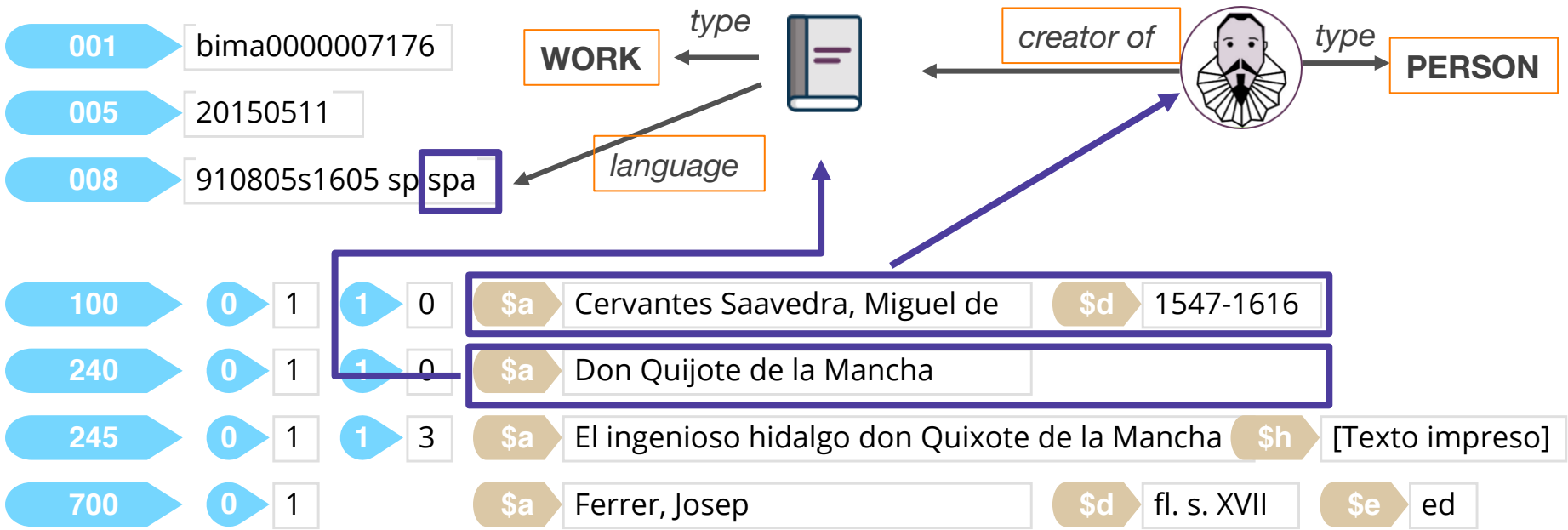
MAPPING

marimba-datamodel | **marimba-rml** | marimba-mapping



MAPPING

marimba-datamodel | **marimba-rml** | marimba-mapping



Third modification of R2RML *(rr:refObjectMap)*

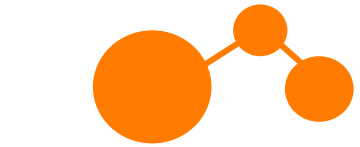
R2RML column → marimba-sql query

*Generate **joins** across nested relations*

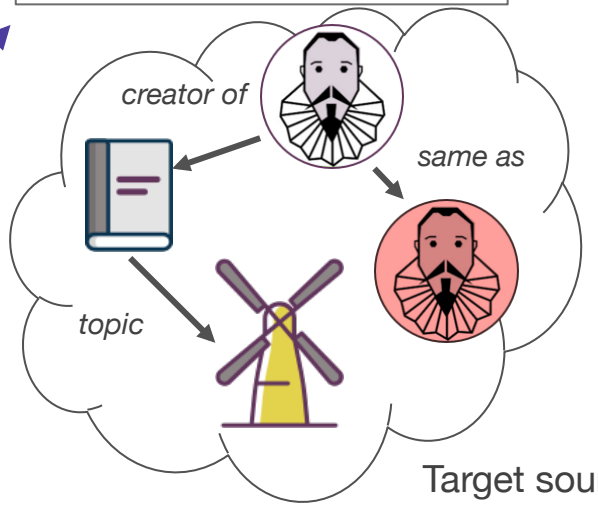
H2 Defined operational semantics
Three core modifications

marimba-mapping

P1
Mapping



GLOBAL SCHEMA



Target source
(RDF)

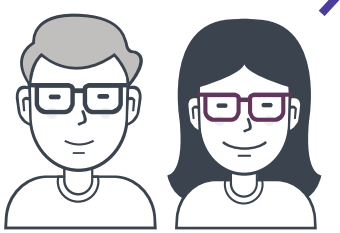
MAPPING

marimba-sql

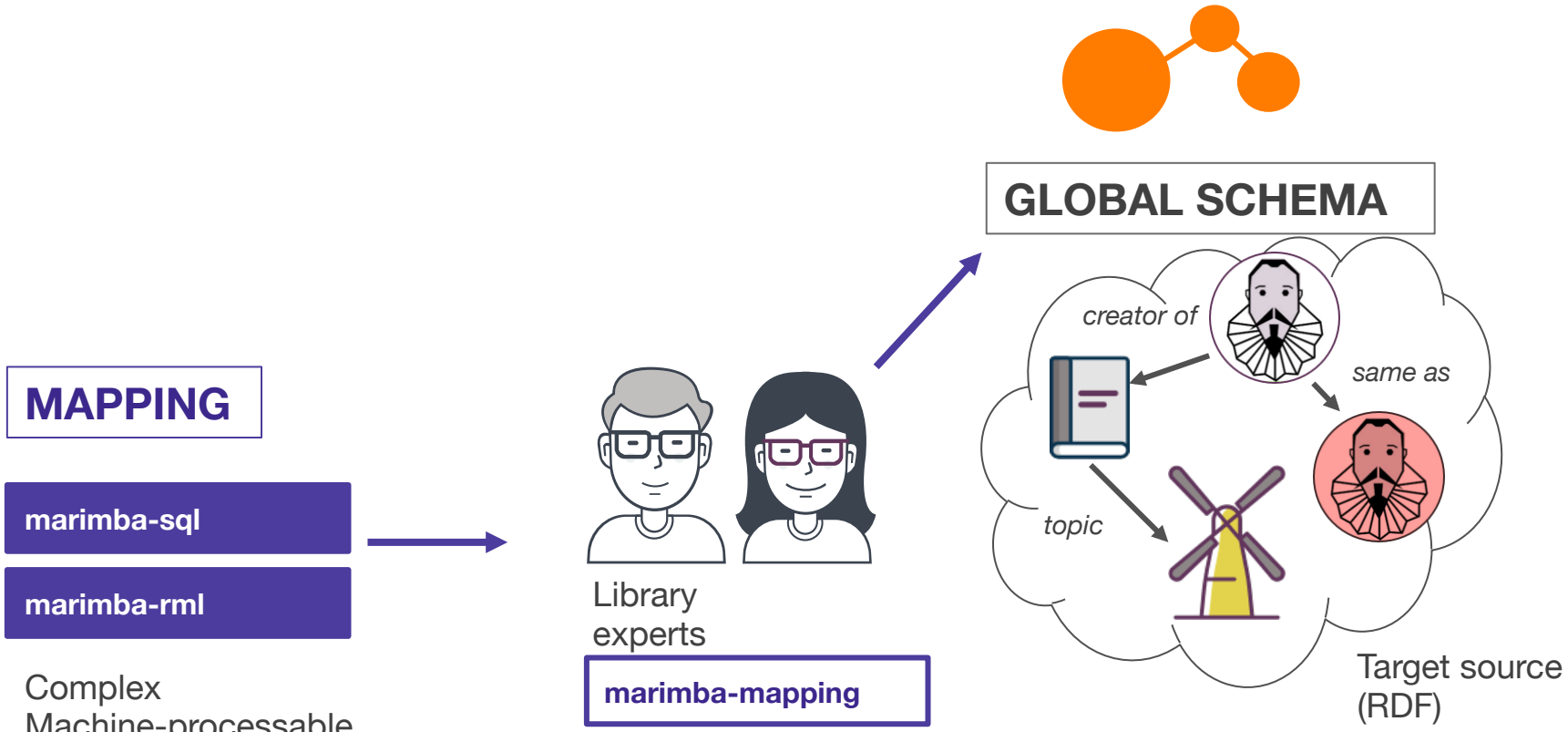
marimba-rml

Complex
Machine-processable

?



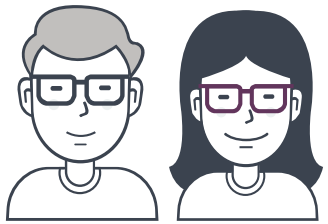
Library
experts


GLOBAL SCHEMA
creator of
same as
topic
Target source
(RDF)
MAPPING
marimba-sql
marimba-rml
Complex
Machine-processable
Library
experts
marimba-mapping

marimba-datamodel | marimba-rml | **marimba-mapping**

Generation of mapping templates

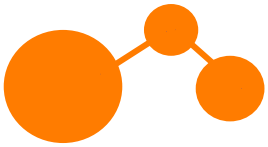
- Support the **mapping and ontology development process by library experts**
- **Full coverage** of data sources (e.g., at BNE more than 4,000 patterns)
- Library experts write simple mappings that are **translated into marimba-rml**



MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

SCHEMA STATISTICS

OUTLINE



BNE ontology

marimba-topicmodel

marimba-modelling

MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

P1
Mapping

P2
Ontology development process

P3
Library applications

Empirical studies

datos.bne.es

Library data sources

Ontology-based Library data (RDF)

Library online applications

BNE ontology

marimba-topicmodel

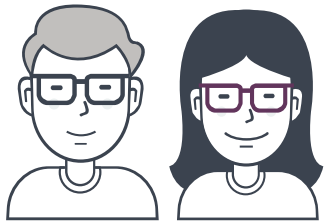
marimba-modelling

P2
Ontology development process

ONTOLOGY DEVELOPMENT

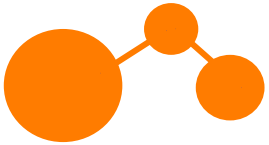
MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

Mapping templates



Library experts

?



Library ontology

ONTOLOGY DEVELOPMENT

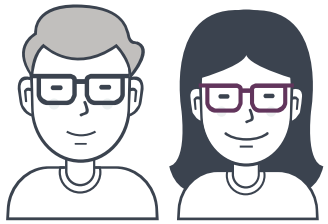
BNE ontology

marimba-topicmodel

P2
Ontology development process

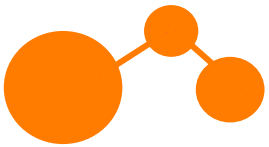
MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

Mapping templates



Library experts

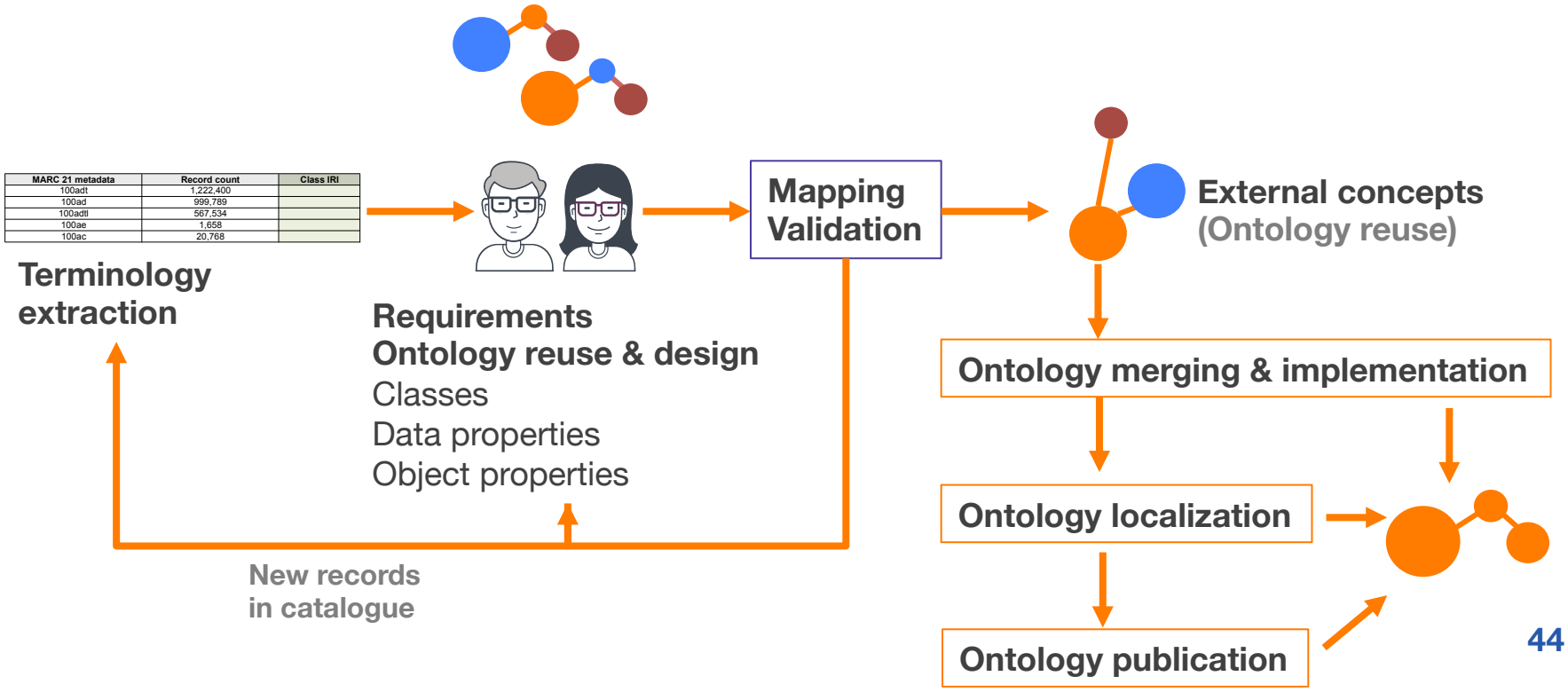
marimba-modelling



Library ontology

marimba-modelling | marimba-topicmodel | BNE ontology

- **Life-cycle model** based on **NeOn** methodology (Suarez-Figueroa et al. 2015).
- **Terminology extraction and ontology design activities** using **mapping templates**.
- **Ontology publication activity** added to NeOn.
- Two phases: **(1) Ontology reuse, and (2) ontology merging**



marimba-modelling | marimba-topicmodel | BNE ontology

ONTOLOGY DESIGN USING MAPPING TEMPLATES

MARC 21 metadata	Class IRI
100adt	frbr:C1001
100ad	frbr:C1005
100adtl	frbr:C1002
100ae	frbr:C1005
100ac	frbr:C1005

Classification template

CLASSES

WORK ← *type* 

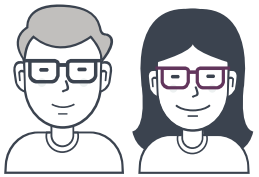
 *type* → PERSON

marimba-modelling | marimba-topicmodel | BNE ontology

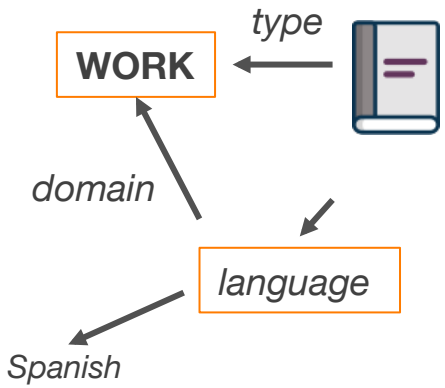
ONTOLOGY DESIGN USING MAPPING TEMPLATES

MARC 21 metadata	Datatype property IRI	Domain IRI
245n	isbd:P3033	frbr:C1003
321a	isbd:P3032	frbr:C1003
110a	frbr:P6001	frbr:C1006
400a	frbr:P5012	frbr:C1005

Annotation template



DATATYPE PROPERTIES
RDFS DOMAINS

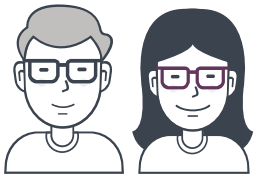


marimba-modelling | marimba-topicmodel | BNE ontology

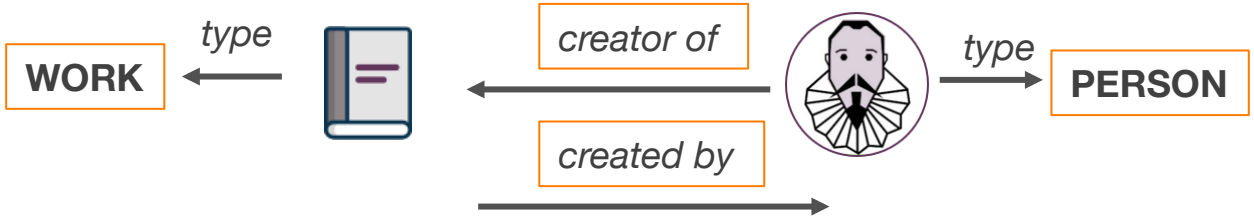
ONTOLOGY DESIGN USING MAPPING TEMPLATES

MARC 21 metadata	Object property IRI	Inverse property IRI	Domain IRI
t	frbr:P5001	frbr:OP1001	frbr:C1005
l	frbr:P1002	frbr:OP2002	frbr:C1001
n	frbr:OP1003	frbr:OP1004	frbr:C1001

Relation extraction template



OBJECT PROPERTIES
RDFS DOMAINS
INVERSE PROPERTIES

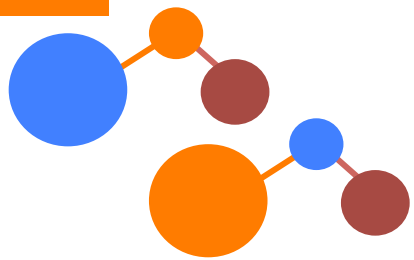


ONTOLOGY DEVELOPMENT

BNE ontology

marimba-topicmodel

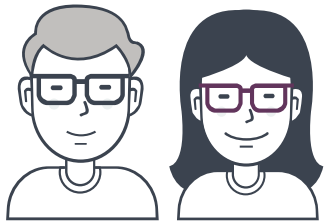
P2
Ontology development process



?

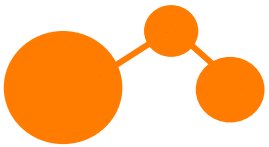
MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

Mapping templates



Library experts

marimba-modelling

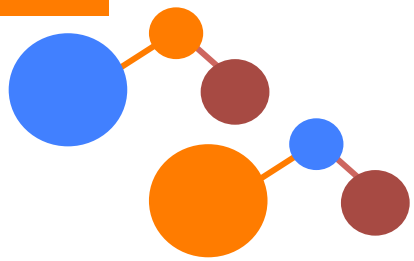


Library ontology

ONTOLOGY DEVELOPMENT

BNE ontology

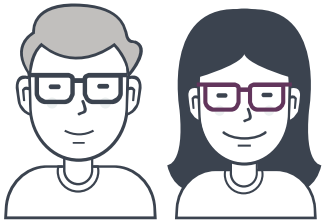
P2
Ontology development process



marimba-topicmodel

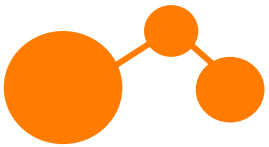
MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

Mapping templates



Library experts

marimba-modelling

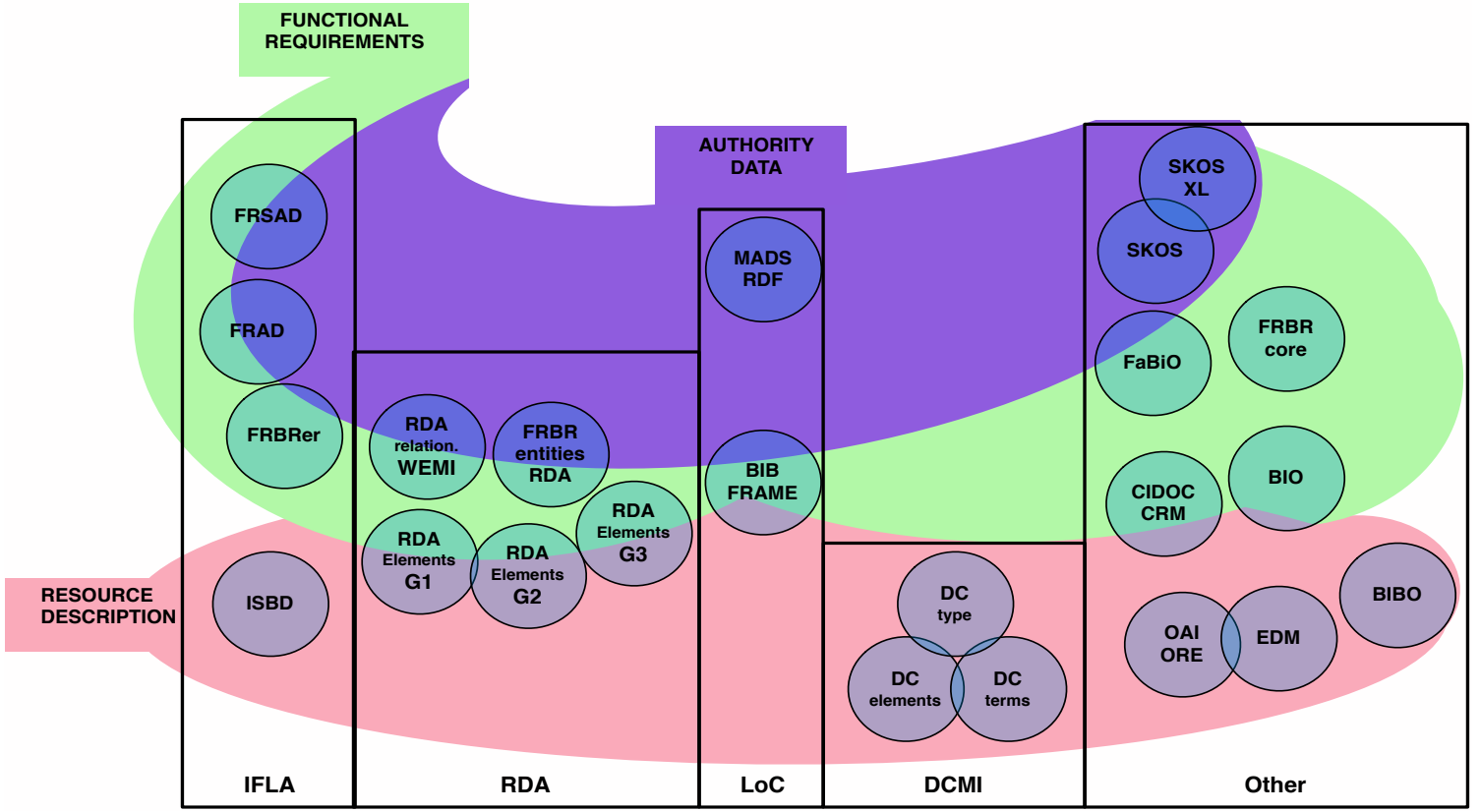


Library ontology

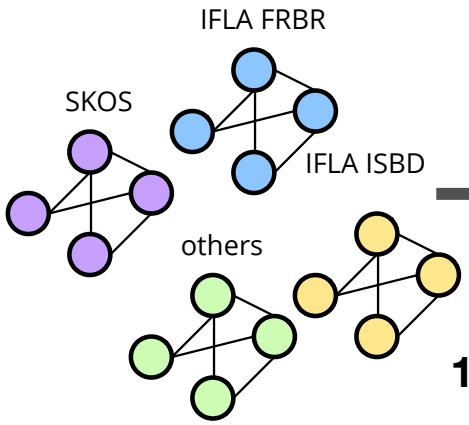
MOTIVATION: **MANUAL CLASSIFICATION OF LIBRARY ONTOLOGIES**

- Using **ontology repositories** (LOV and metadataregistry.org)
- Very **costly** and **not scalable** (e.g., new ontologies?)

Can we do this analysis **automatically**?



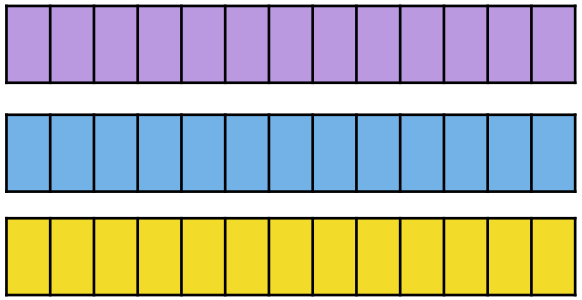
MARIMBA-TOPICMODEL: **OVERVIEW**



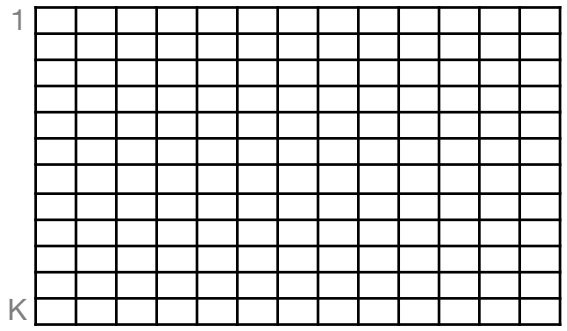
Corpus of ontologies

marimba-topicmodel

1. **Ontology documents** from textual descriptions
2. **Annotate** with **external word senses**
3. Train a **probabilistic topic model**



Ontologies as probability distributions over **K** topics



K topics as probability distributions over **S** word senses

Topic-based ontology similarity measures:

JSD: Jensen-Shannon Divergence

JSM: Jensen-Shannon Metric

marimba-topicmodel | Evaluation | Application

LOD resources
(e.g., DBpedia categories)

BabelNet sense URI	*Word(s)*
bn:s00046705n	*info*
bn:s00025314n	*data*
bn:s00029345n	*email*
bn:s00046516n	*person*
bn:s00049910n	*language*
bn:s00023236n	*Country, nation*

TOPIC EXAMPLE: PERSONAL INFORMATION

HIGH PROBABILITY: FOAF, BIBO, FABIO

EXPERIMENT 1: **TOPIC COHERENCE**

1. **External senses** improve the **coherence** of topics?
2. **Short-text oriented** topic models produce **more coherent** topics than **classical topic models**?

Measure: Mimmo et al. [2011]

METHODS

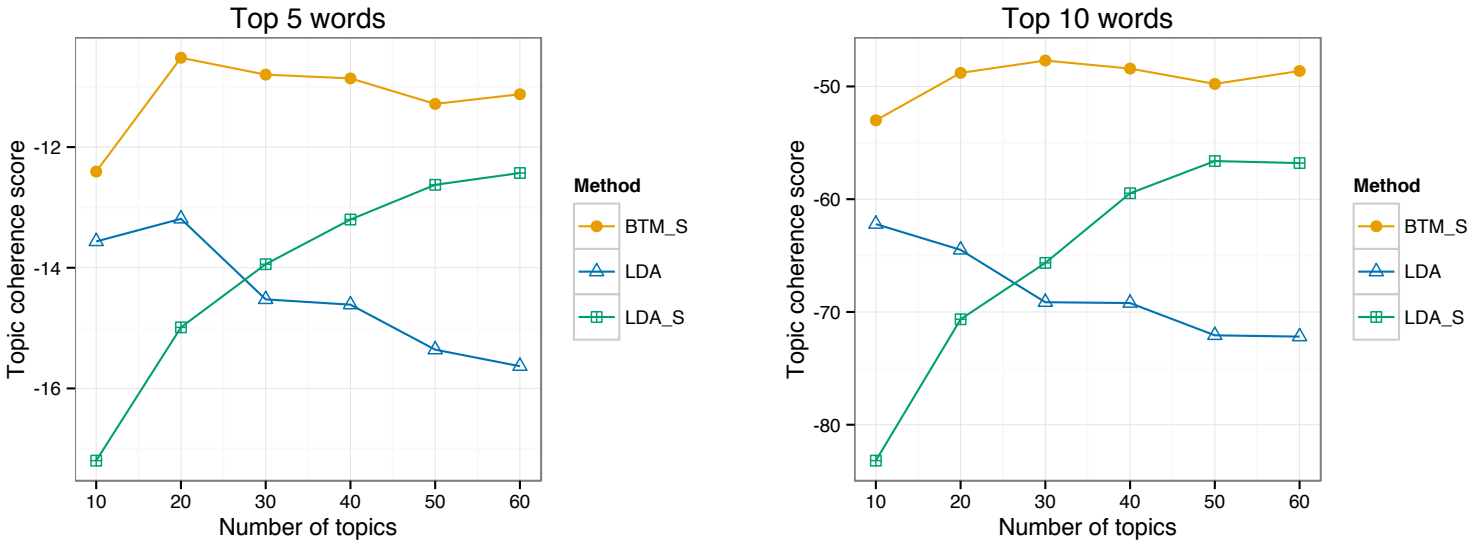
Classical LDA (Latent Dirichlet Allocation)

LDA_S: LDA with sense-annotated documents

BTM_S (marimba-topicmodel): BTM with sense-annotated documents

EXPERIMENT 1: **TOPIC COHERENCE**

1. **External senses** improve the **coherence** of topics?
2. **Short-text oriented** topic models produce **more coherent** topics than **classical topic models**?



1. Annotation with **senses increase topic coherence** (BTM_S > LDA_S >LDA)
2. **Short-text topic models increase coherence**

   (**BTM_S** consistently outperforms LDA and LDA_S).

EXPERIMENT 2: **ONTOLOGY CLUSTERING EXPERIMENT**

Does **marimba-topic model** produce **more precise clustering** than classical search methods?

- **How to measure?**
  - **H-score: IntraCluster** distance **<** **Intercluster** distance

- **Methods with different distance measures and metrics:**
  - **2 TF-IDF:** Baseline method (Cosine and Euclidean distance)
  - **2 Marimba-topicmodel** (JSD and JSM)

- **Gold-standards** (annotated by humans)
  1. **LODTHEMES:** LOD cloud themes
  2. **LOVTAGS:** Topic tags of LOV repository

marimba-topicmodel | **Evaluation** | Application

EXPERIMENT 2: **ONTOLOGY CLUSTERING EXPERIMENTAL RESULTS**

	tf-idf cosine	marimba-cosine	marimba-jsd
LOVTAGS	0.924	0.861	**0.859**
LODTHEMES	0.842	0.722	**0.683**

	tf-idf eucli	marimba-jsm
LOVTAGS	0.808	**0.998**
LODTHEMES	0.723	**0.741**

* **Lower** values mean **better precision** (low intracluster, high intercluster)

Does **marimba-topic model** produce **more precise clustering** than classical search methods

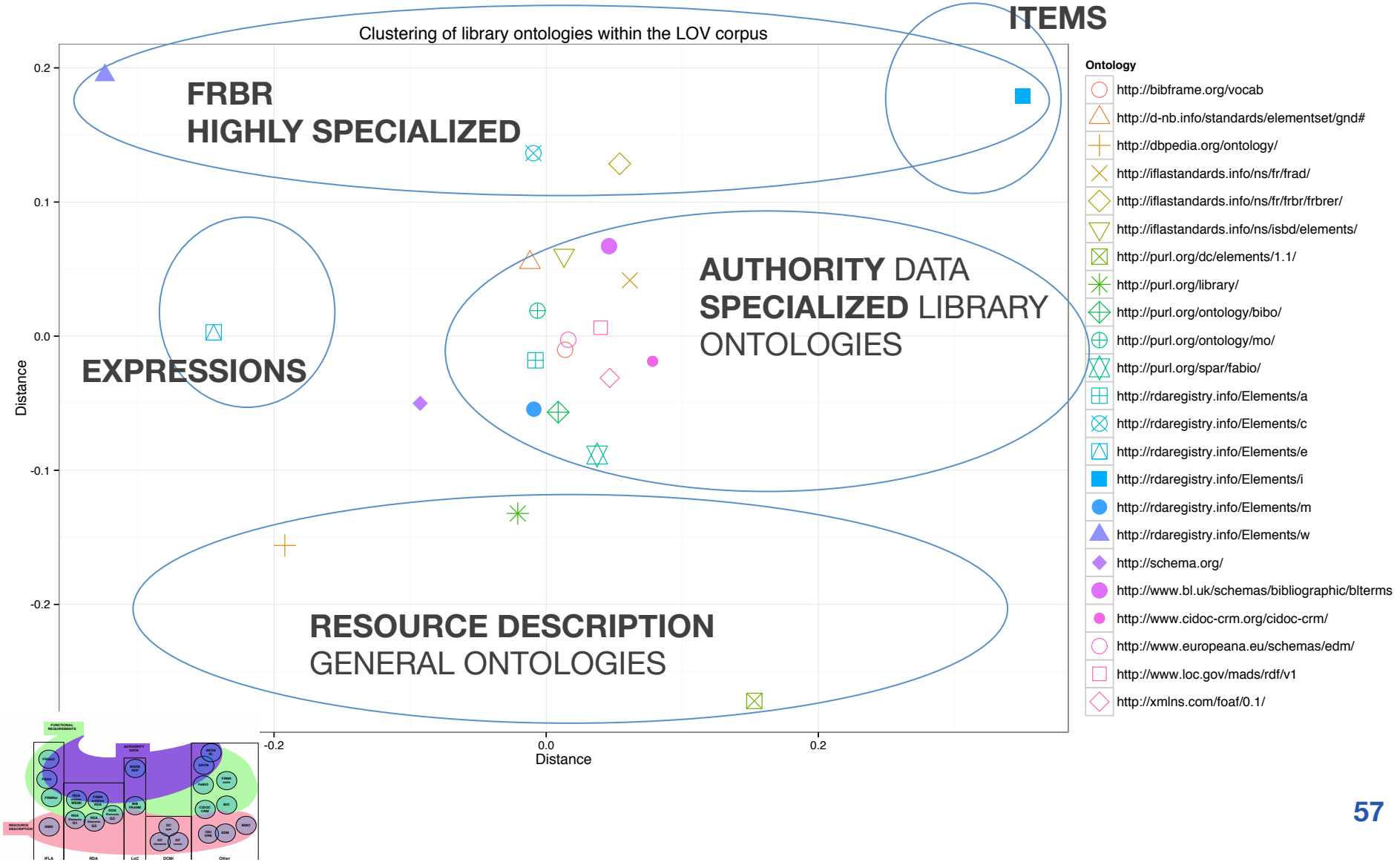
H4 **Topic-based ontology similarity measures perform better** than existing methods used in ontology search for **clustering widely used LOD ontologies** (LOV repository and LOD cloud).

But, what about **library ontologies** and **more qualitative results**?

ONTOLOGY DEVELOPMENT

Ontology Engineering Group

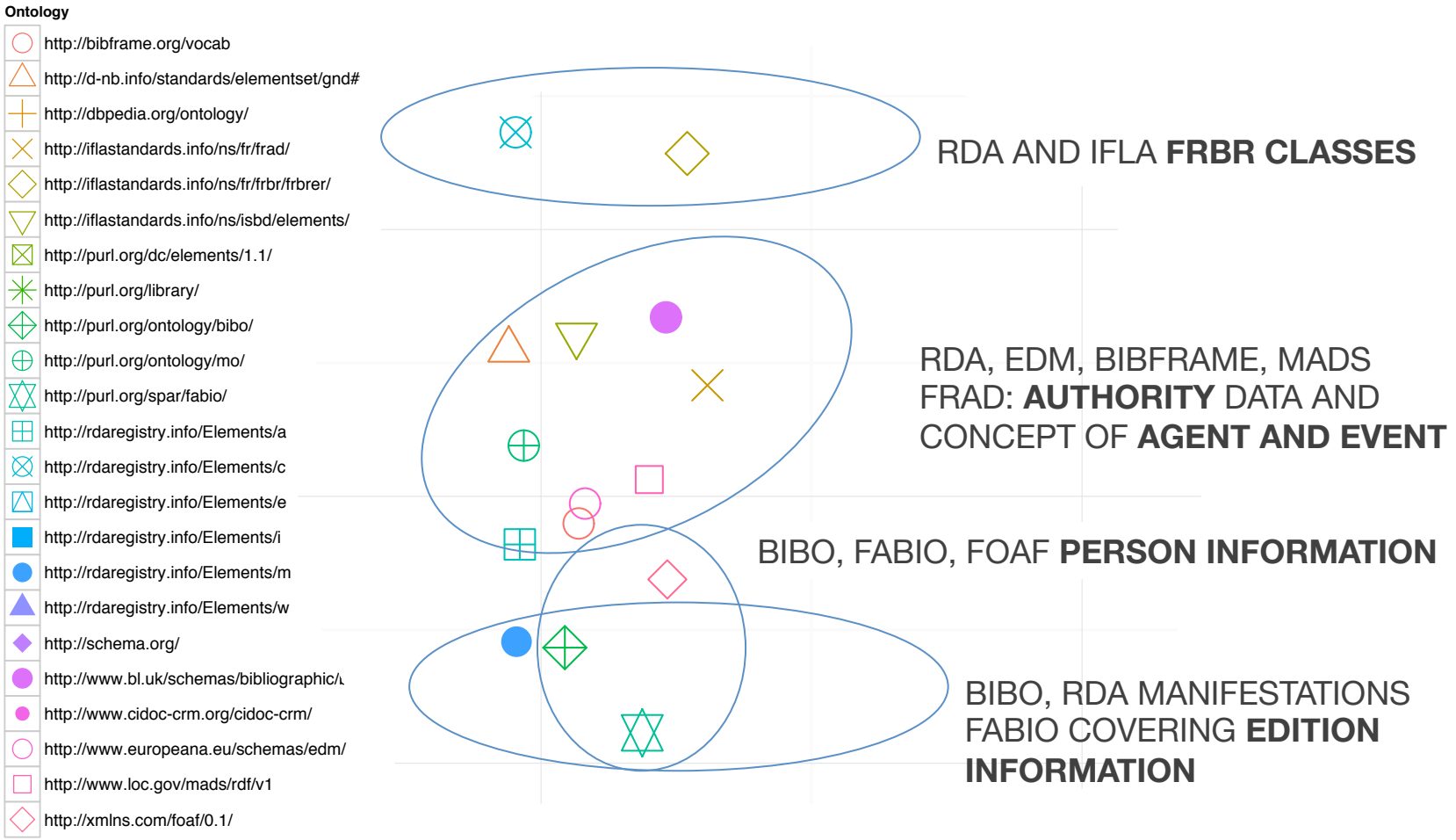
marimba-topicmodel | Evaluation | Application



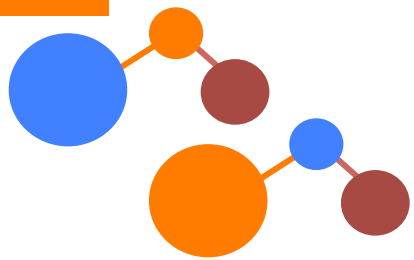
ONTOLOGY DEVELOPMENT

Ontology Engineering Group

marimba-topicmodel | Evaluation | **Application**



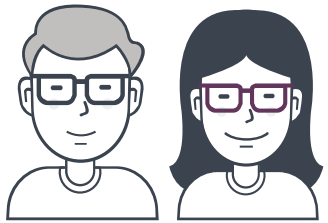
P2
Ontology development process



marimba-topicmodel

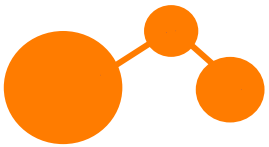
MARC 21 metadata	Record count	Class IRI
100adt	1,222,400	
100ad	999,789	
100adtl	567,534	
100ae	1,658	
100ac	20,768	

Mapping templates



Library experts

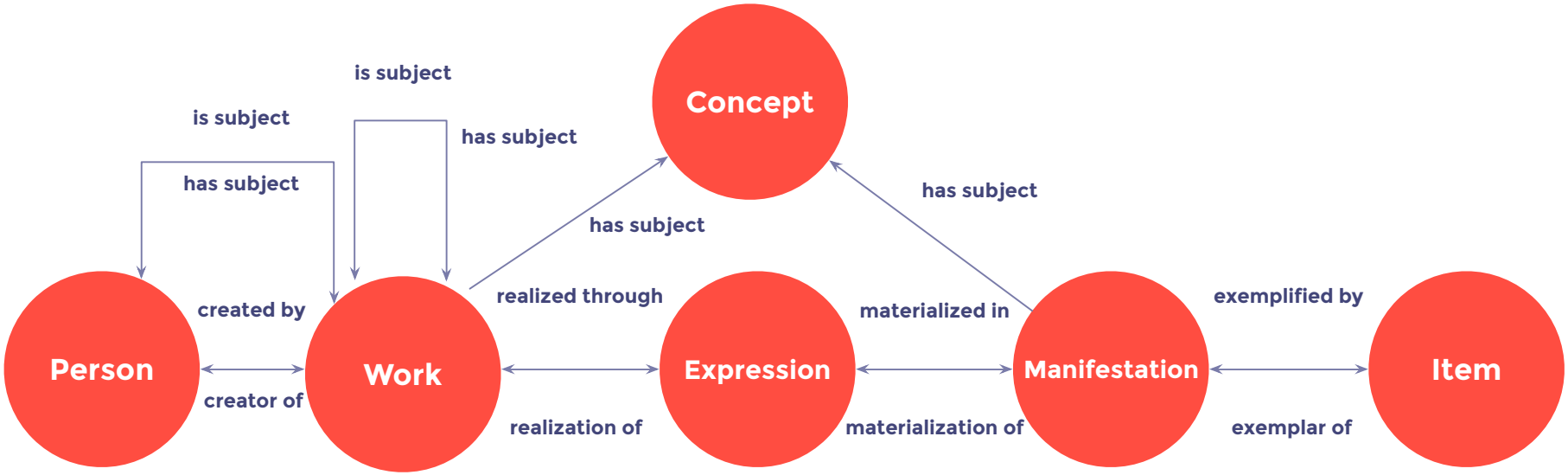
marimba-modelling



Library ontology

BNE ontology

marimba-modelling | marimba-topicmodel | **BNE ontology**



- Includes **alignments** to library ontologies (ISBD, IFLA FRBR, RDA)
- **6** classes, **33** object properties, **200** datatype properties
- **Ontology localization**: Spanish and English
- **Ontology publication**: HTML doc, RDF formats

marimba-modelling | marimba-topicmodel | **BNE ontology**

H3 Analytical data and the feedback of library experts can be used to develop a library **ontology** with sufficient **quality** with respect to **state of the art evaluation metrics.**

Evaluation from two perspectives:

1. **Topology based**: *OOPS!* (Poveda-Villalón et al. [2014]):

   0 critical pitfalls, only 3 important.

2. **Application-based:** Task-based experiment with 72 participants.

OUTLINE

Empirical studies

datos.bne.es

P1
Mapping

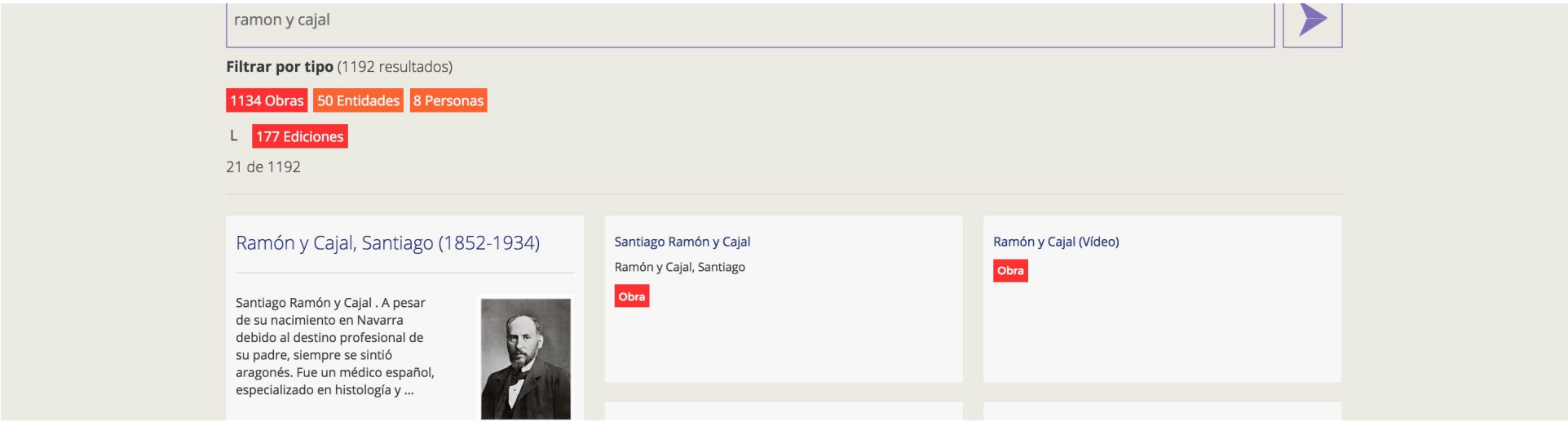
P2
Ontology development process

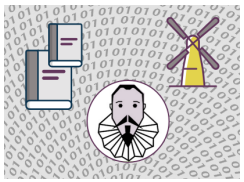
P3
Library applications

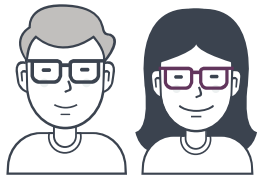
Library data sources

Ontology-based Library data (RDF)

Library online applications







marimba-rml
mappings

marimba-tool

Methodology

RDF
generation

Publication
Exploitation

datos.bne.es

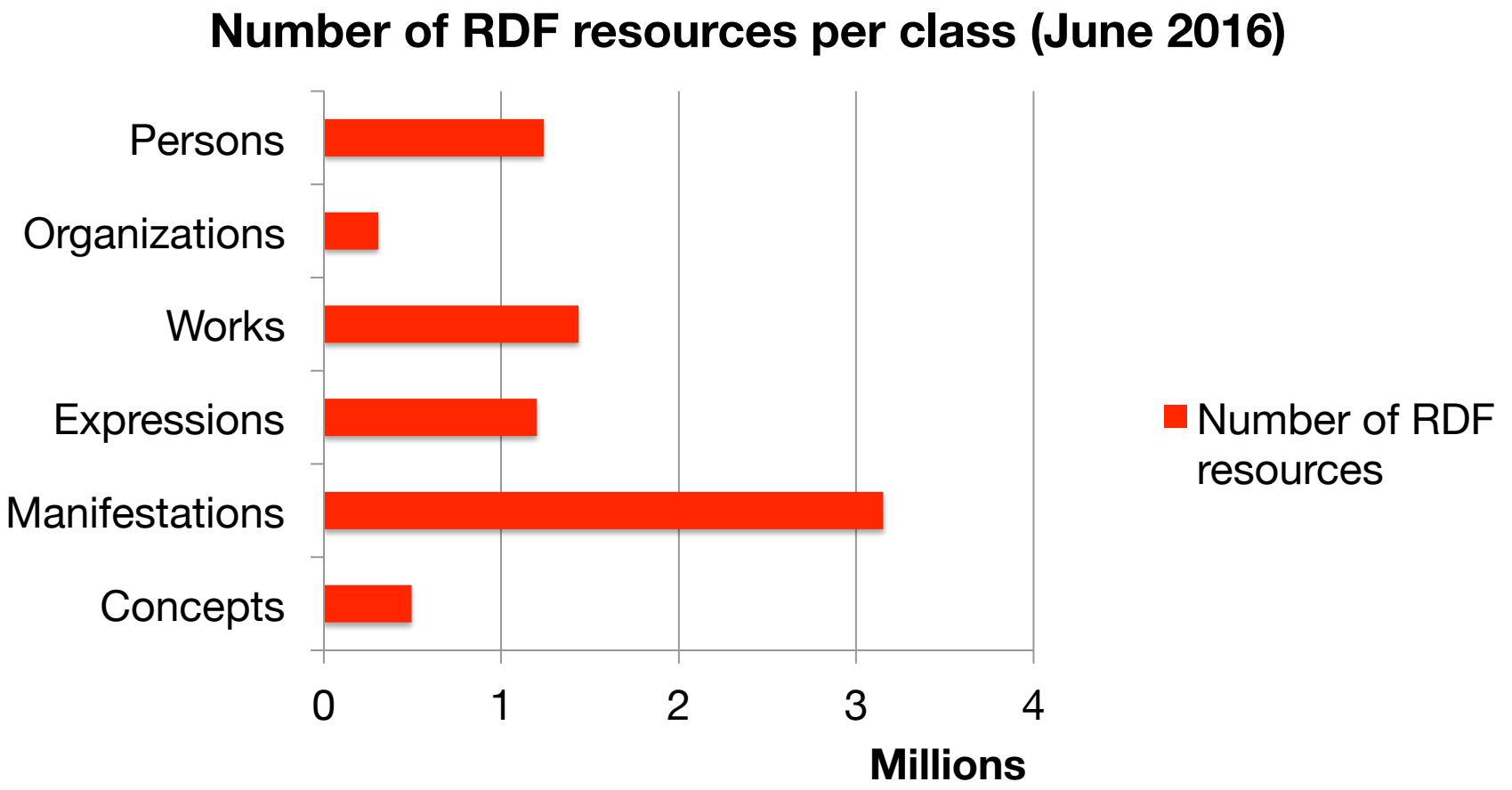
Library
data sources

Ontology-based
Library data (RDF)

Library
online applications

datos.bne.es | Empirical studies | Task-based experiment | Usability study

PUBLICATION: **+7 MILLION** RDF RESOURCES, **+1M SAMEAS** LINKS



LIBRARY ONLINE APPLICATIONS

datos.bne.es | Empirical studies | Task-based experiment | Usability study

DATOS·BNE·es *beta

Inicio Personas Entidades Obras Temas Ayuda

BIBLIOTECA NACIONAL DE ESPAÑA BNE

1. MULTILINGUAL SEARCH

aristoteles

Filtrar por tipo (663 resultados)

654 Obras | 7 Personas | 2 Entidades

907 Ediciones

21 de 663

2. ONTOLOGY ENTITIES

3. GRAPH-BASED RANKING

4. DATA **ENRICHMENT**

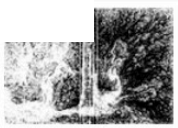
Aristóteles

Aristóteles, en griego antiguo Ἀριστοτέλης Aristotélēs, fue uno de los más influyentes filósofos de la antigüedad, de la historia de la filosofía occidental y considerado por muchos como ...

Obras destacadas

Poética
Ética a Nicómaco
Política
Metafísica

Aristóteles
Obra Disponible en digital



Ética a Nicómaco
Aristóteles
Obra Disponible en digital

Política
Aristóteles
Obra Disponible en digital

Metafísica
Aristóteles
Obra Disponible en digital

Main questions:

H5

1. Do semantic technologies increase **user satisfaction** and **efficiency?**
2. **Holistic evaluation** of contributions in the thesis

Comparing 2 large-scale services of BNE:

1. Online Public Access Catalogue: **catalogo.bne.es**
2. Linked data-based service: **datos.bne.es**

Two experiments:

1. **Task-based experiment:** 72 participants
2. **Usability and user satisfaction study**: 36 participants

EXPERIMENTAL SETTING

- **Between-group** experiment:

  72 participants of library and information sciences.

- User's goal: Complete an information retrieval **scenario of 10 tasks.**

- **Scenario:**
  Documentation about **Spanish scientist** "Santiago Ramón y Cajal"

  – **Designed** with **professors of Library and information sciences**
  – **Realistic tasks** reproducing a documentation scenario.
  – **Tasks: interrelated and with different levels of complexity. E.g.,:**

  "Find dates related to the author",
  "Find the title of three works",
  "Find works by topic"

LIBRARY ONLINE APPLICATIONS

datos.bne.es | Empirical studies | Task-based experiment | Usability study

EVALUATION APPLICATION

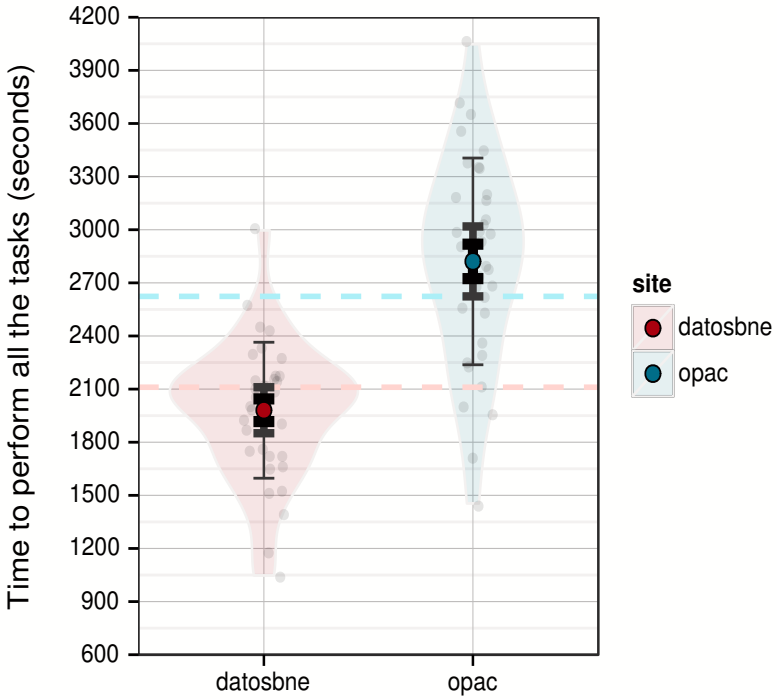
datosbne

OPAC

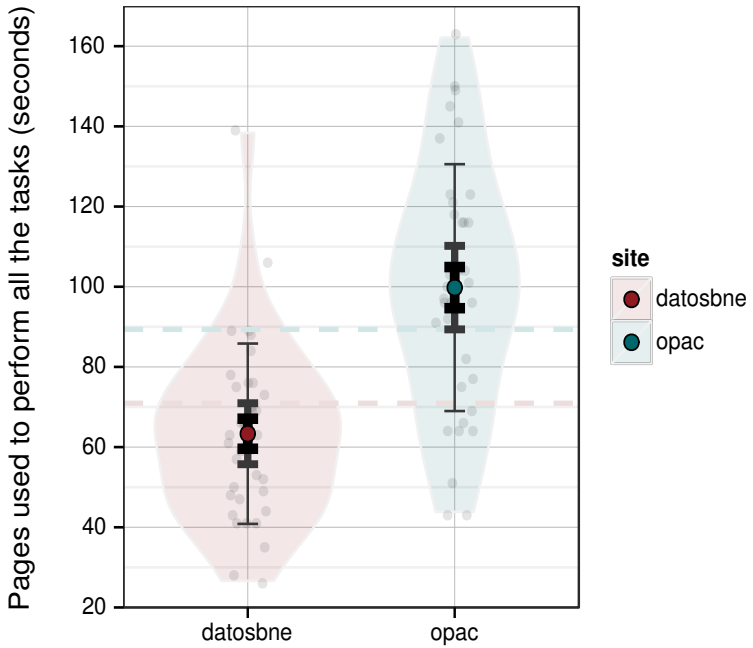
- **Measures:**
  - Task **completion time** (Kelly [2009])
  - **Visited pages** (Su [1992])
  - **User satisfaction** (1-5 scale) (Brooke [2013])

Efficiency

- **32% less time** on avg. to complete scenario (**>15 min**)



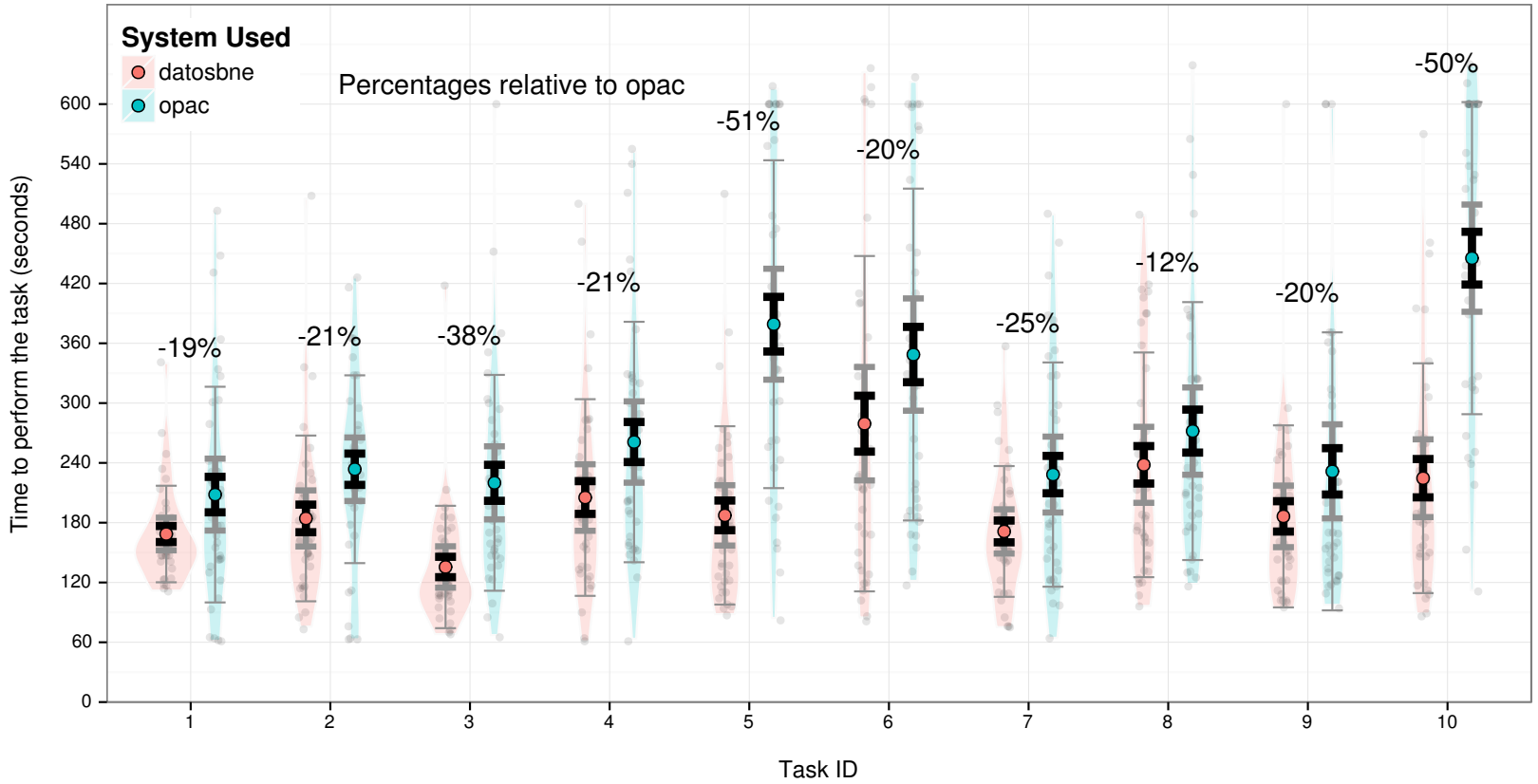
- **40% less pages** on avg. to complete scenario (**>40 pages**)



H5

Users are significantly more **efficient** in completing **information retrieval tasks**.

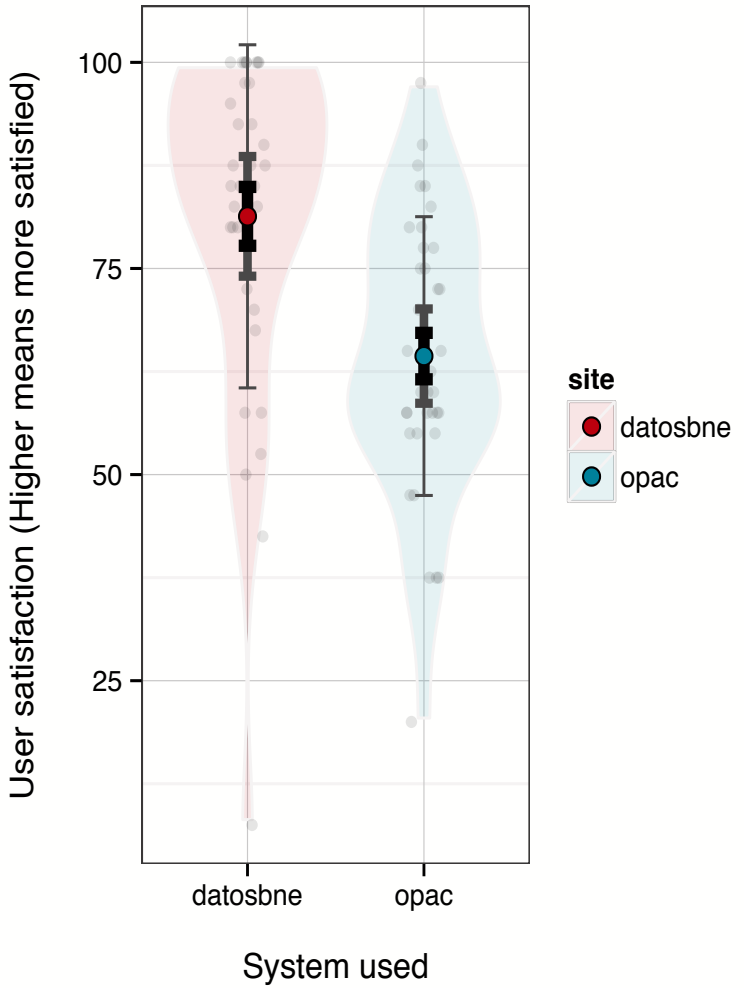
Time completion per task



- **More complex tasks: difference > 50% → Complex navigation of ontology**
  - Finding the **title** of a **translation** into English (Task 5)
  - Finding the **topic** and **author** of a work (Task 10)

Task User Satisfaction

- **datos.bne.es** average score = **81**
- **17% higher** than catalogo.bne.es



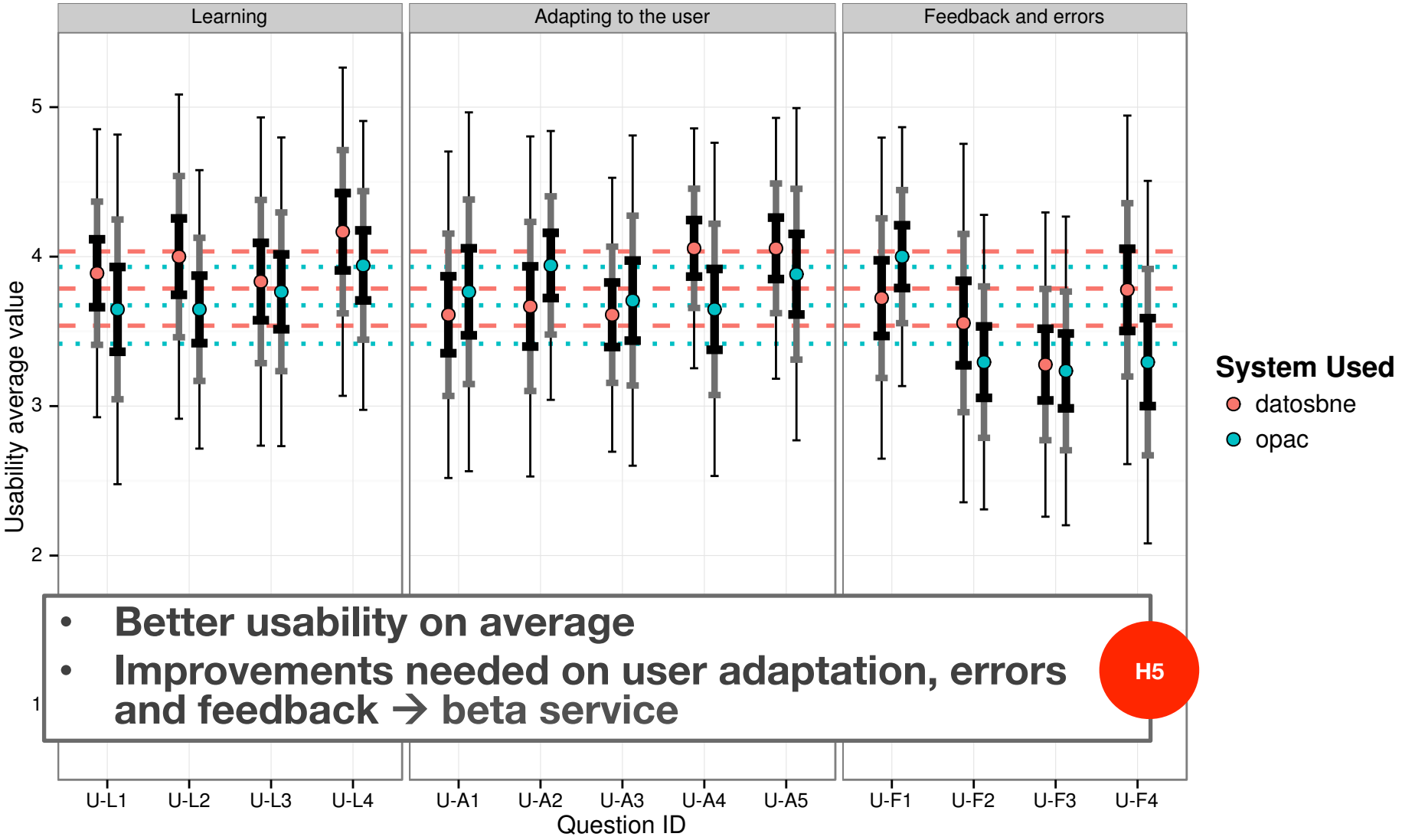
H5 Significant increase in **user satisfaction** for information retrieval tasks

USABILITY STUDY: **EXPERIMENTAL DESIGN**

- **36 participants** (18 for each system)
- **No** time measurement and **constraints**
- **No task solving**
- Usability test “Practical Heuristics for Usability Evaluation” (Perlman [1997])

LIBRARY ONLINE APPLICATIONS

datos.bne.es | Empirical studies | Task-based experiment | **Usability study**



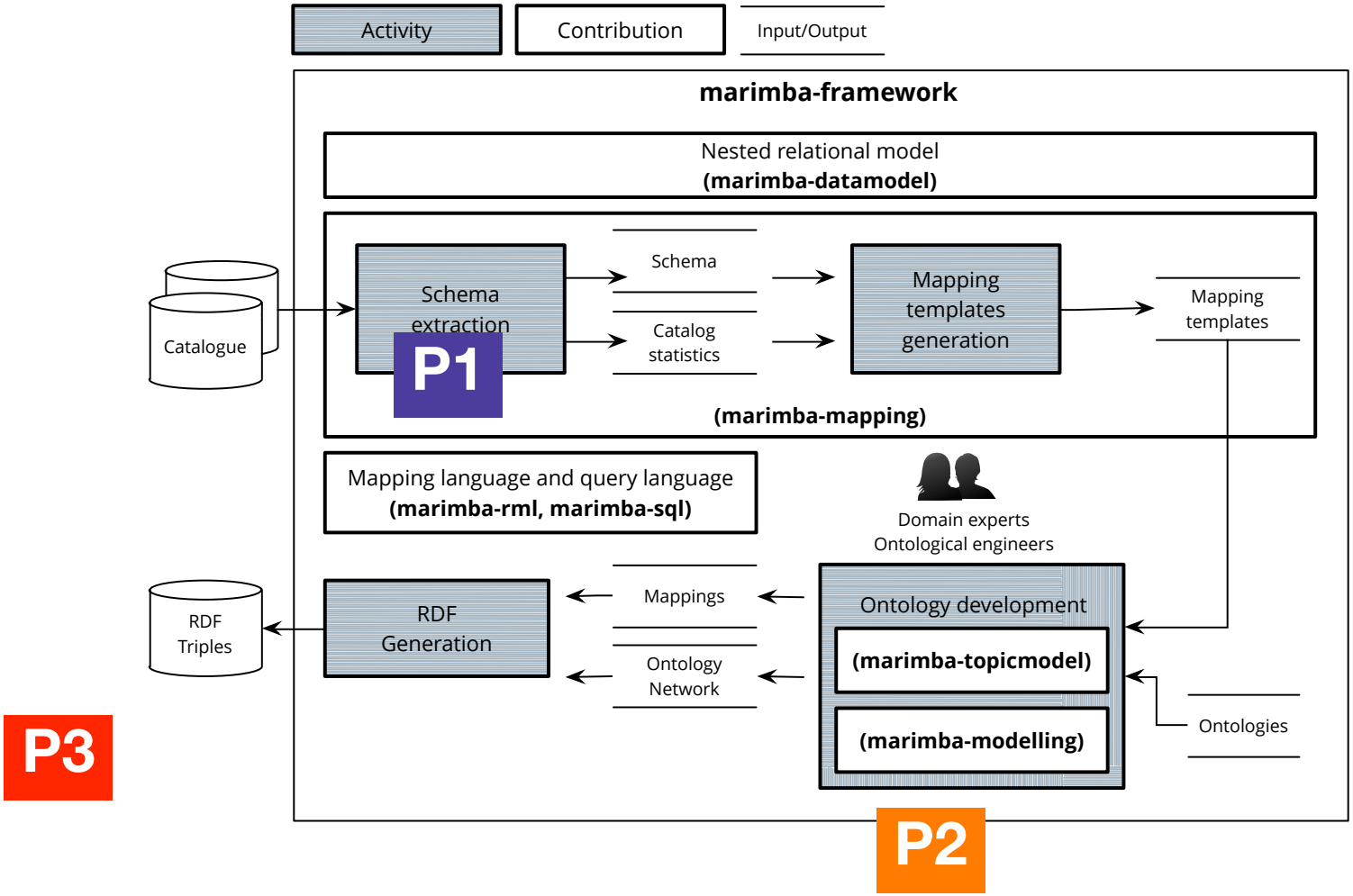
- **Better usability on average**
- **Improvements needed on user adaptation, errors and feedback → beta service**

H5

CONCLUSION AND FUTURE DIRECTIONS

THEORETICAL, TECHNOLOGICAL AND EMPIRICAL RESULTS

End-to-end solution for ontology-based library data generation, access and exploitation



OTHER RESULTS

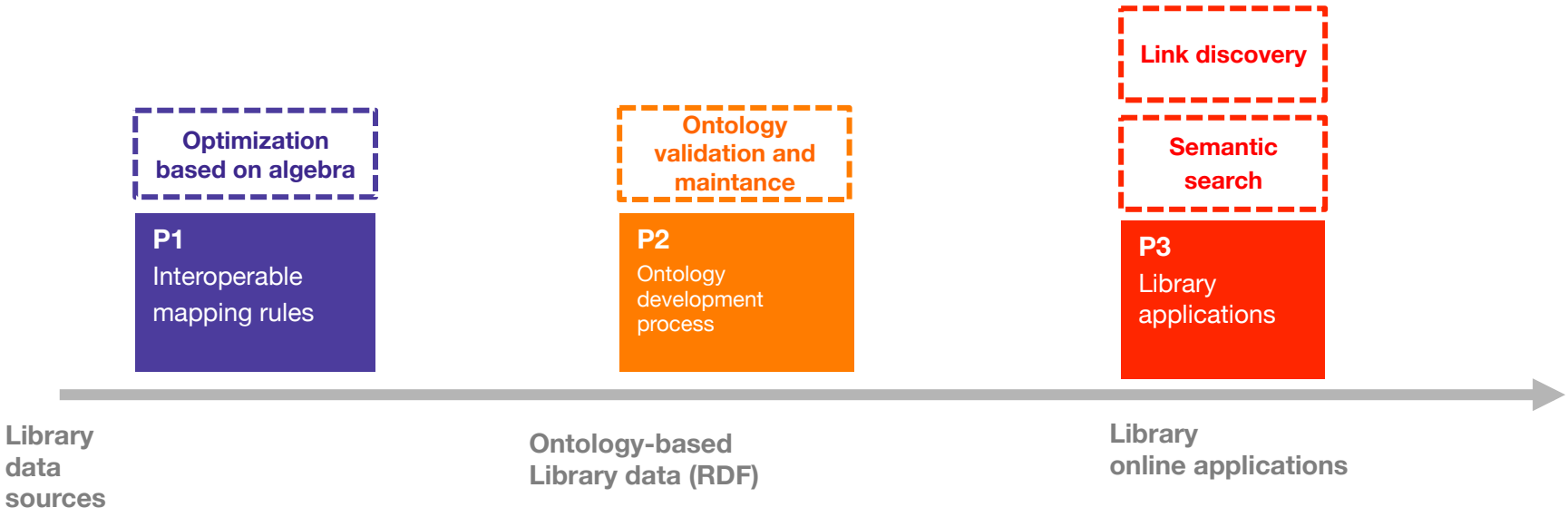
DATOS·BNE·es *beta

- **Discoverability**:
  - **45,000 visitors/month,** +1M visitors,
  - **Access** through **search engines to BNE catalogue has doubled**. Specially "**long-tail" queries**
- **Reusability**:
  - Used by **+10 third-party library projects**
  - **Wikidata**
  - **Street detective:** Fujitsu Laboratories of Europe Innovation Award 2015
- International **reference library LOD implementation**
- Library science **courses at Universities.**

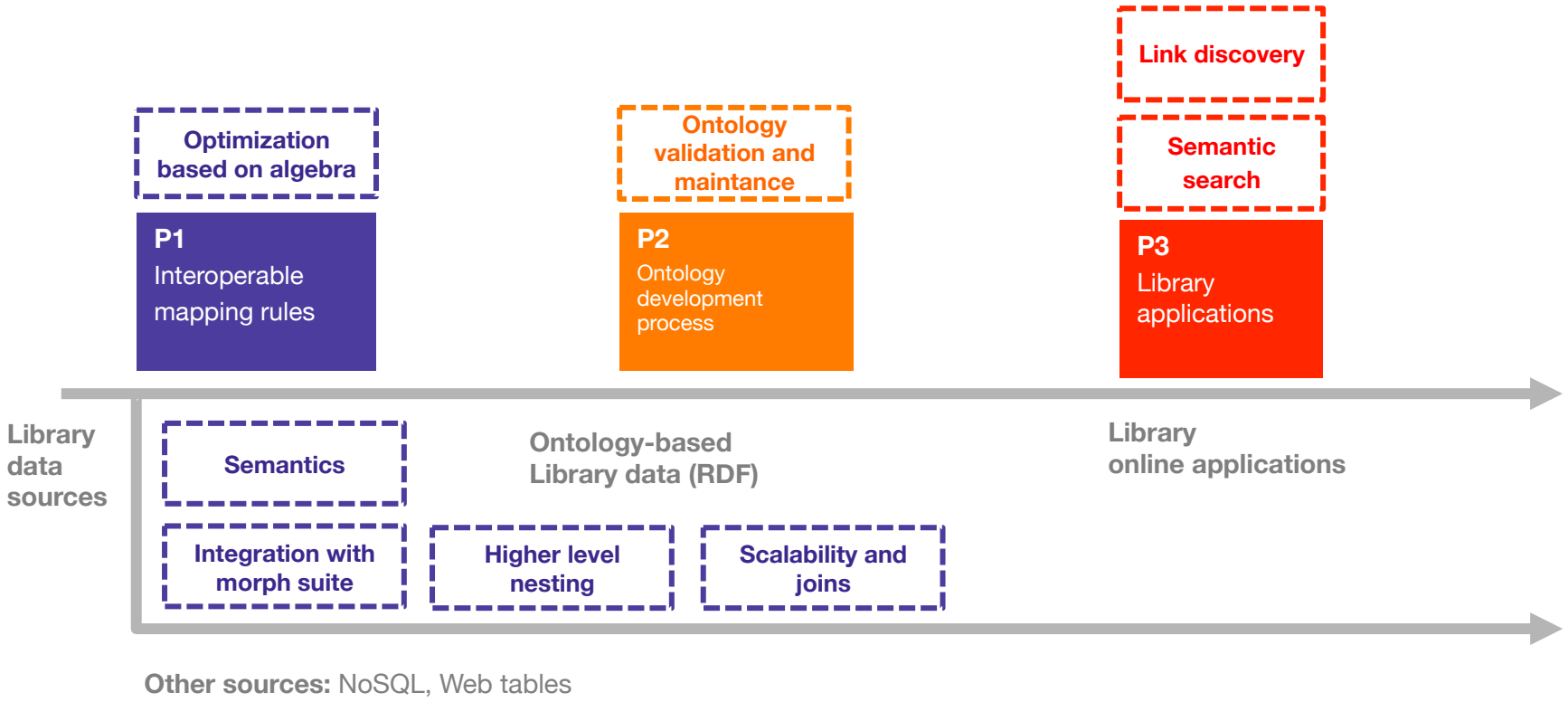
OTHER RESULTS

- **Technology transfer:**
  - **Commercial product:** deployed and licensed to BNE.
  - Requested by +**20 libraries world-wide.**
  - **Diploma** to one of the **best business ideas UPM 2016** (ActuaUPM).
- **Scientific value:**
  - Reference **methodological guidelines**.
  - Topic of **+10 invited talks** and **keynotes** at major library events (e.g., LODLAM summit, EuropeanaTech, SWIB 2012 and 2015).

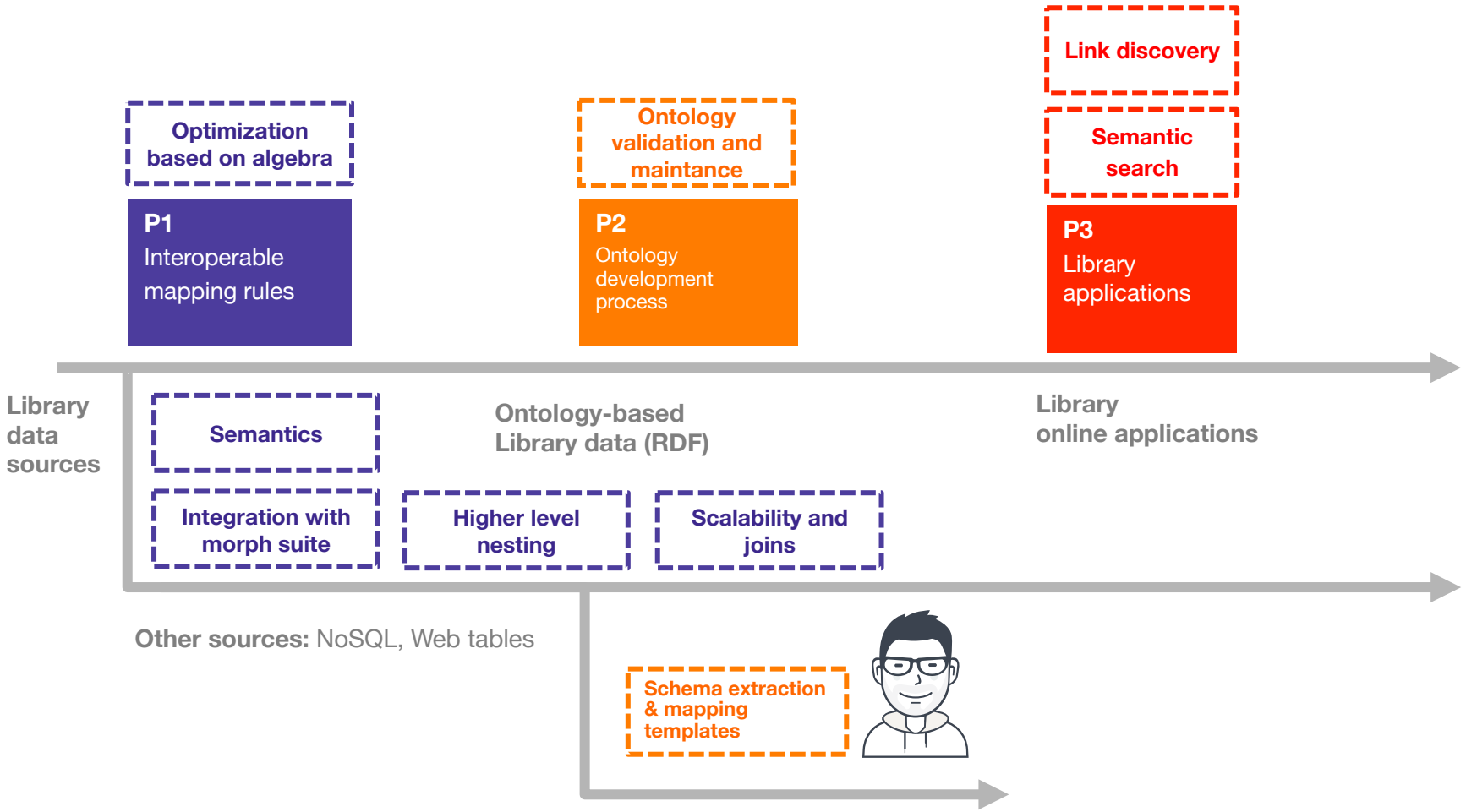
FUTURE DIRECTIONS



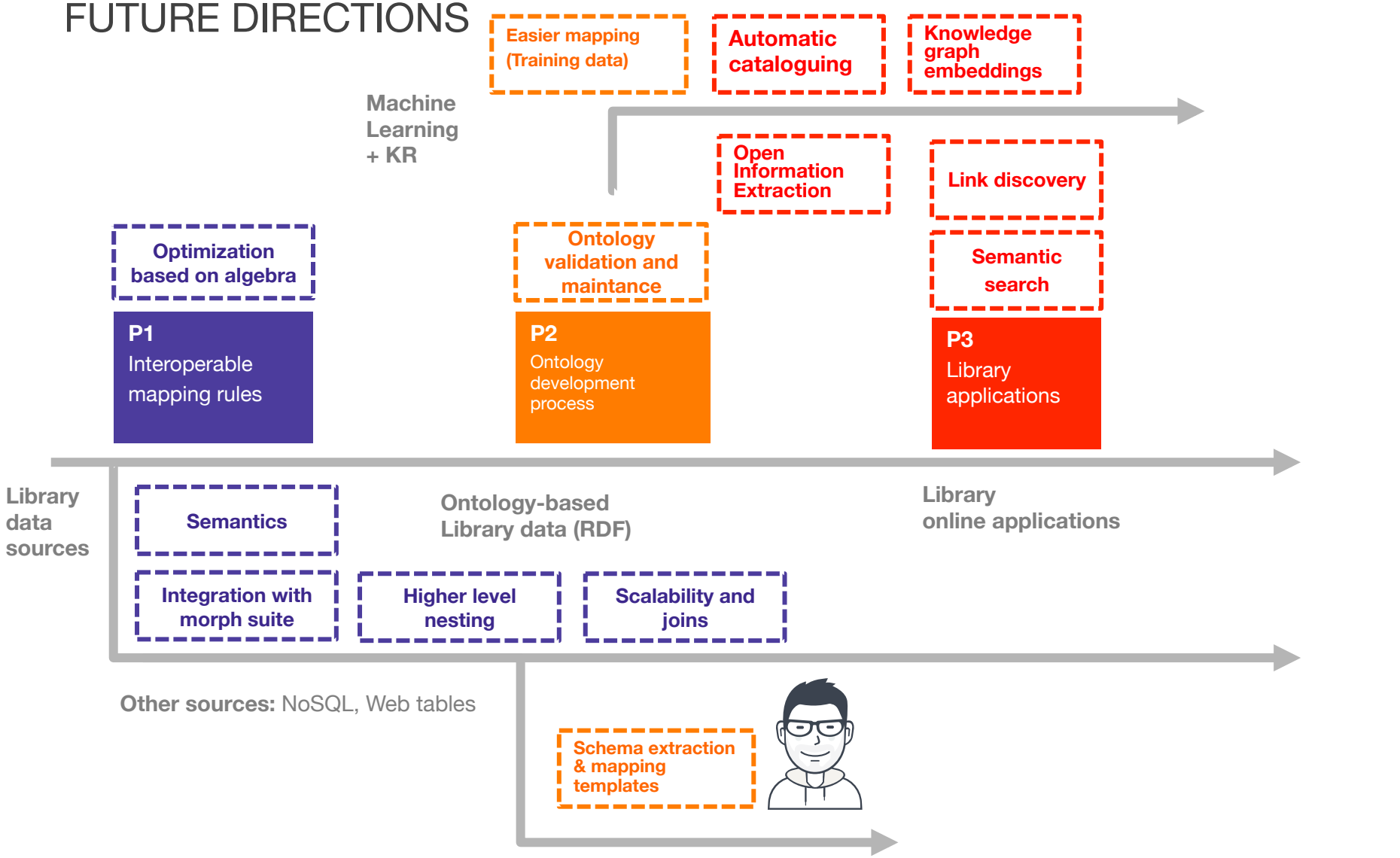
FUTURE DIRECTIONS



FUTURE DIRECTIONS



FUTURE DIRECTIONS



SELECTED PUBLICATIONS

- **JOURNAL ARTICLES**
  - **Daniel Vila-Suero** and Asunción Gómez-Pérez. datos.bne.es and marimba: an insight into library linked data. Library Hi Tech, 31(4):575–601, 2013.
  - **Daniel Vila-Suero**, Boris Villazón-Terrazas, and Asunción Gómez-Pérez. datos.bne.es: A library linked dataset. Semantic Web Journal, 4(3):307–313, 2013.
- **INTERNATIONAL CONFERENCES**
  - Ricardo Santos, Ana Manchado, and **Daniel Vila-Suero**. Datos.bne.es: a LOD service and a FRBR-modelled access into the library collections. In IFLA World Library International Conference. Cape Town, South Africa, 2015.
  - **Daniel Vila-Suero**, Víctor Rodríguez-Doncel, Asunción Gómez-Pérez, Philipp Cimiano, John McCrae, and Guadalupe Aguado-de Cea. 3ld: Towards high quality, industry-ready linguistic linked linguistic data. European Data Forum 2014, 2014.
  - Asunción Gómez-Pérez, **Daniel Vila-Suero**, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de-Cea. Guidelines for multilingual linked data. In 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13, Madrid, Spain, June 12-14, 2013, page 3.
- **OTHER**
  - **Daniel Vila-Suero**, Asunción Gómez-Pérez,Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de-Cea. Publishing linked data on the web: The multilingual dimension. In Paul Buitelaar and Philipp Cimiano, editors, Towards the Multilingual Semantic Web, pages 101–117. Book chapter. Springer Berlin Heidelberg. 2014
  - **Daniel Vila-Suero**. W3C Library Linked Data Incubator Group: Use cases Report. 2011.

RESEARCH STAYS

- **INRIA** 2012 (3 months) and 2013 (2 months)
- **Knowledge Management Institute, Open University, UK:** 2012, 1 month working on linked data applications

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