Ontology-based data access technology applied to Bio2RDF

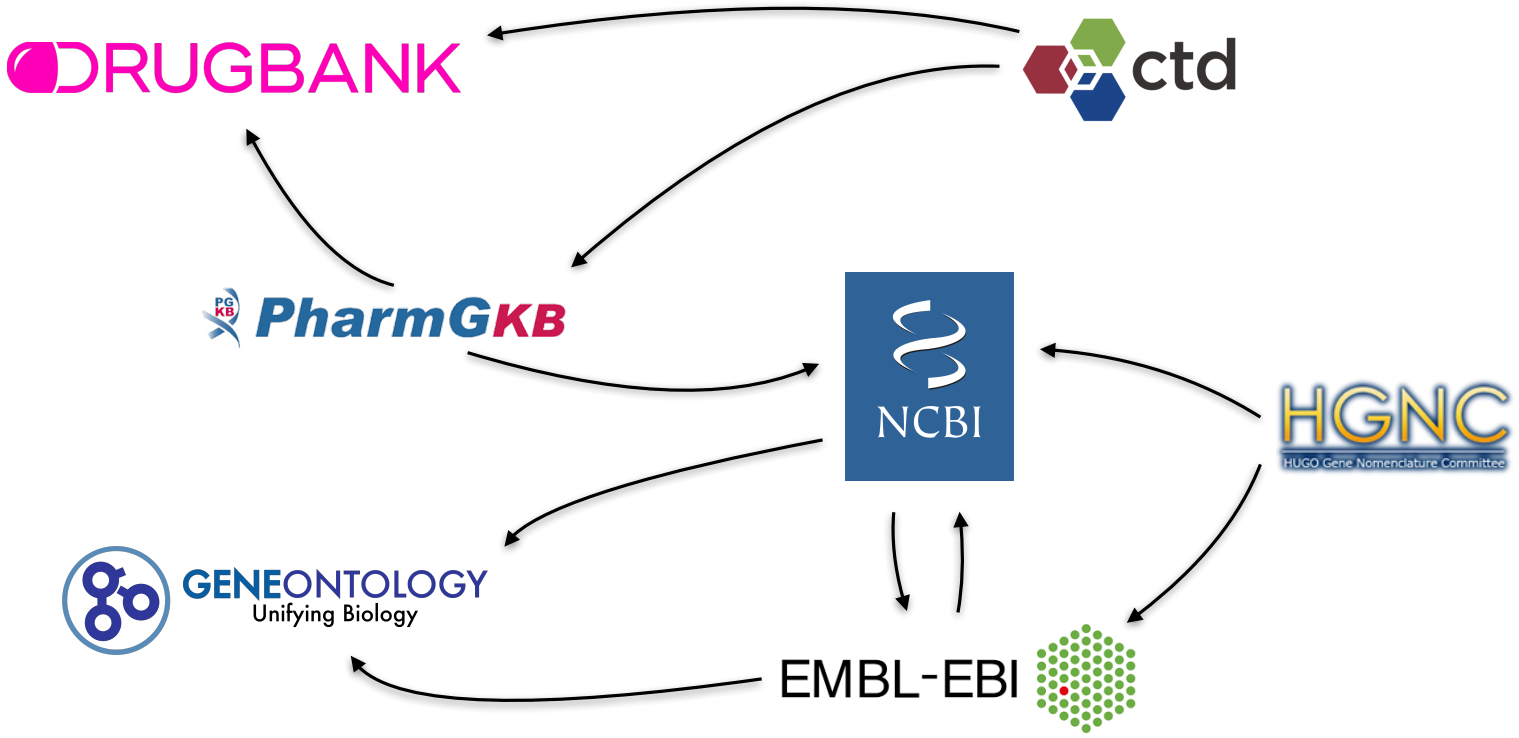
Author: Ana Iglesias Molina
Director: Oscar Corcho García

ana.igmolina@alumnos.upm.es

03/07/2019

There are more tan 1500 bioinformatic data sources, since the biological knowledge is increasing...

There are more tan 1500 bioinformatic data sources, since the biological knowledge is increasing...

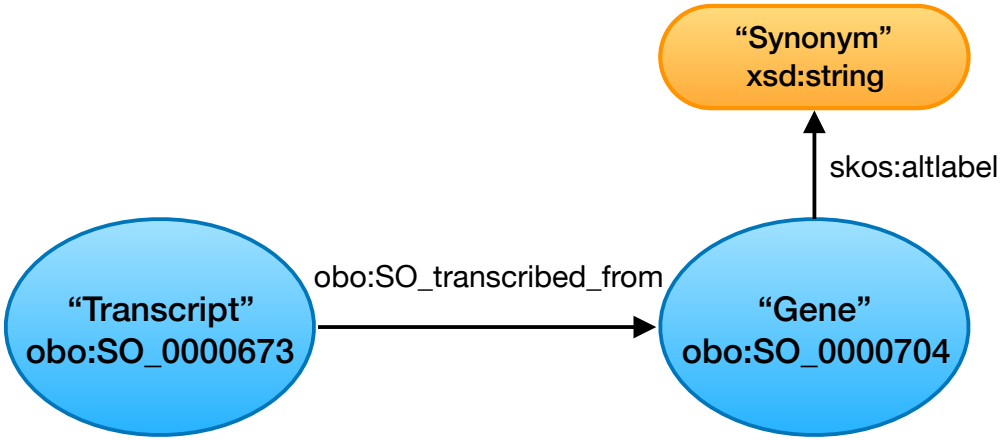


...but are they really linked?

Semantic Web

- Solution to address the data integration problem
- Published data in RDF, which is based on triples
- The schema followed is proposed as ontologies
- Queried with SPARQL

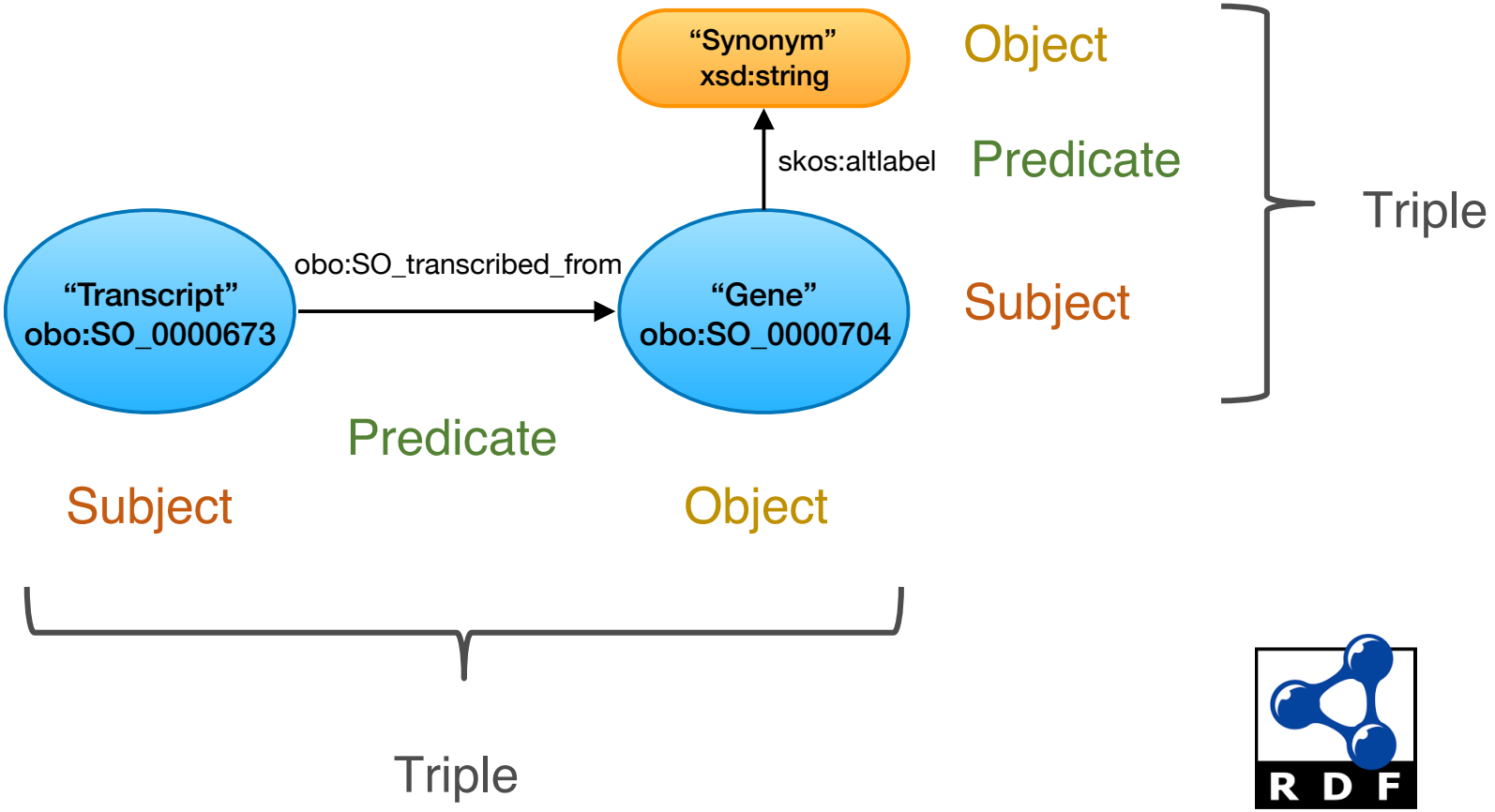
Semantic Web



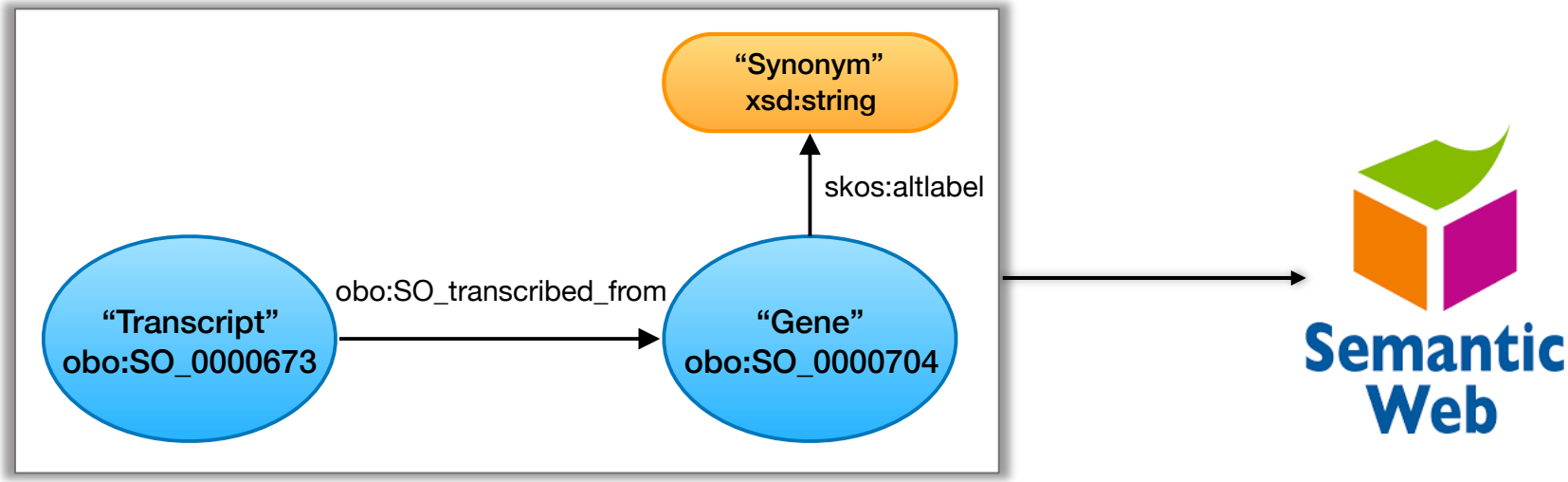
Transcript	Transcribed_from
transcript1	gene1
transcript2	gene1

Gene	Synonym
gene1	syn
gene2	syn

Semantic Web

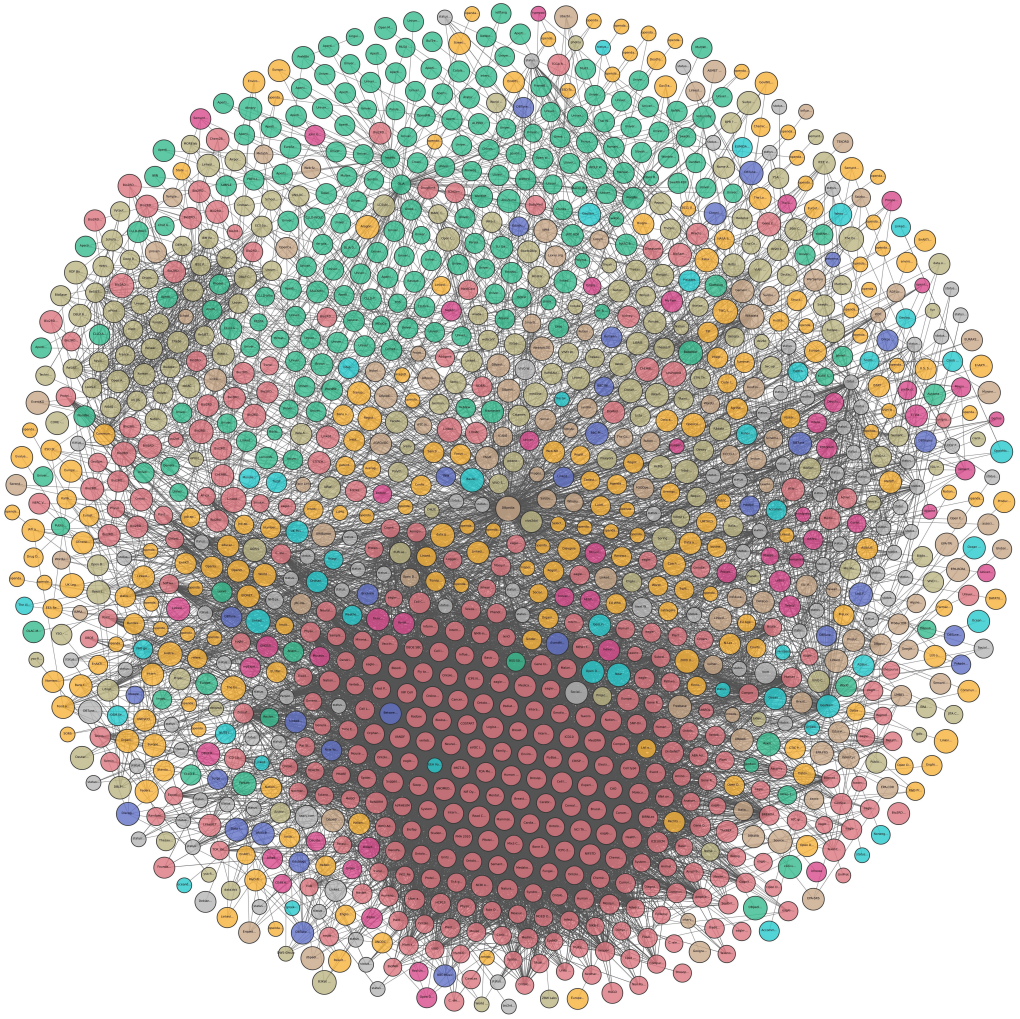


Semantic Web



@prefix obo: http://purl.obolibrary.org/obo/
@prefix skos: http://www.w3.org/2004/02/skos/core#
@prefix xsd: http://www.w3.org/2001/XMLSchema#

Linked Open Data Project

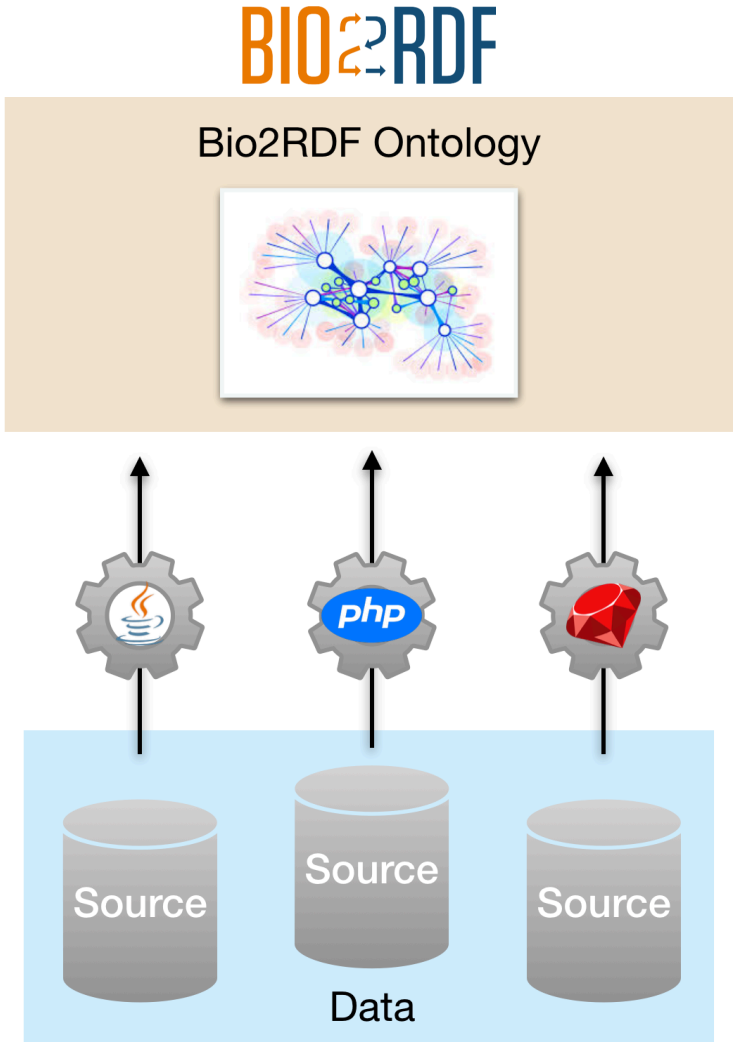


Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated

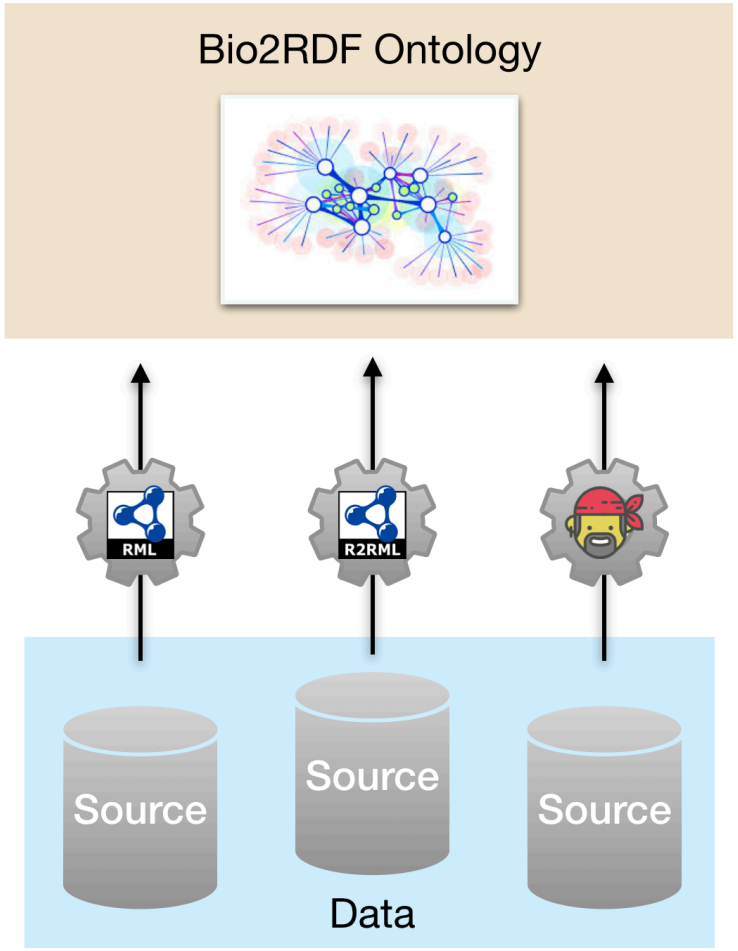
The Bio2RDF project

- It has published 43 datasets as Linked Data
- Materialization to RDF with mainly PHP scripts
- 3 Releases: last one in 2014
- Data outdated and poor maintenance since last release



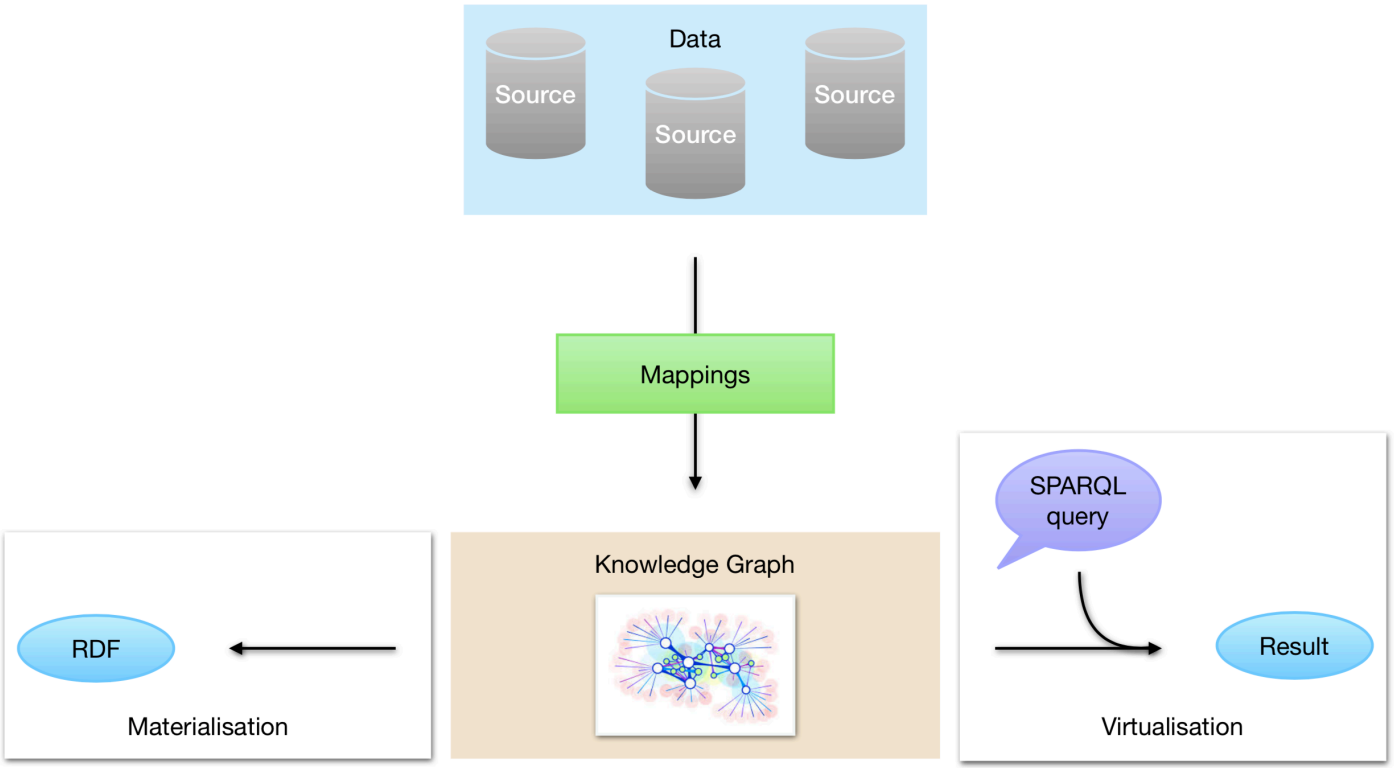
Ontology-Based Data Access (OBDA)

- Unified view and common access to a set of data sources based on ontologies and mappings
- The relationships between the knowledge graph and the source data are stablished with **mappings**
- It creates a **Virtual Knowledge Graph**, a virtual RDF view of the data



Ontology-Based Data Access (OBDA)

- Enables **materialisation** (RDF conversion) and **query translation** (SPARQL to queries supported by underlying source)



Objectives

- Change the methodology to access the data in Bio2RDF from RDF materialisation to virtualisation with OBDA. To achieve that:

  1. Create the mappings for the datasets of Bio2RDF
  2. Develop a methodology with the minimum interaction from the user that deals with the heterogeneity of the data
  3. Overcome the problems the actual methodology of Bio2RDF presents

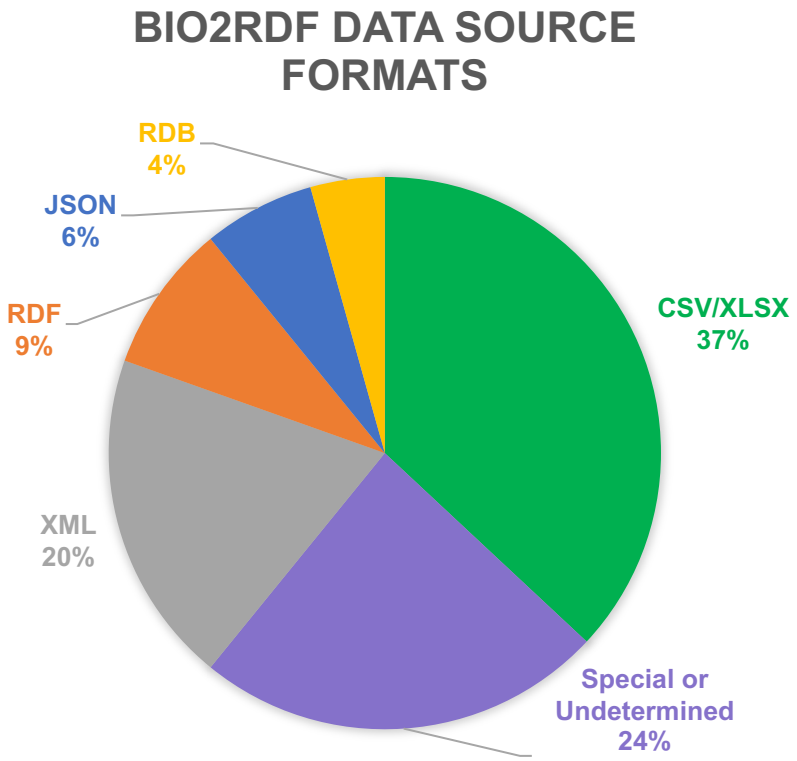
Contributions

Generation of mappings and annotations of the datasets which content is available as CSV files. These mappings are more maintainable than the original code, and reusable by other OBDA engines

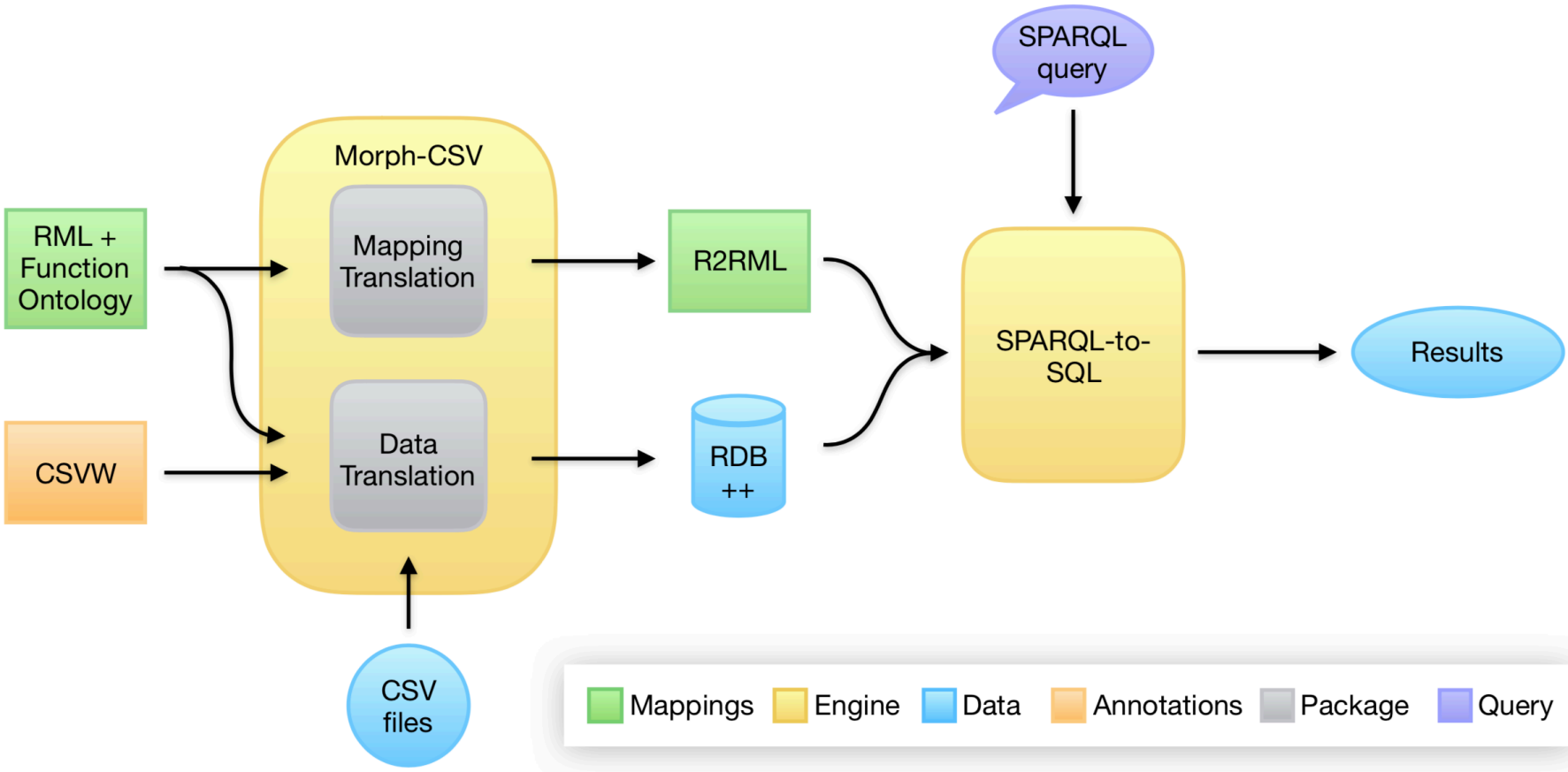
Evaluation of the process proposed with Morph, a suite of tools for OBDA, executing the mappings

Summary of challenges and difficulties that the management of the data presents

- **CSV** is the main format
- Focus on dealing with datasets with CSV data published

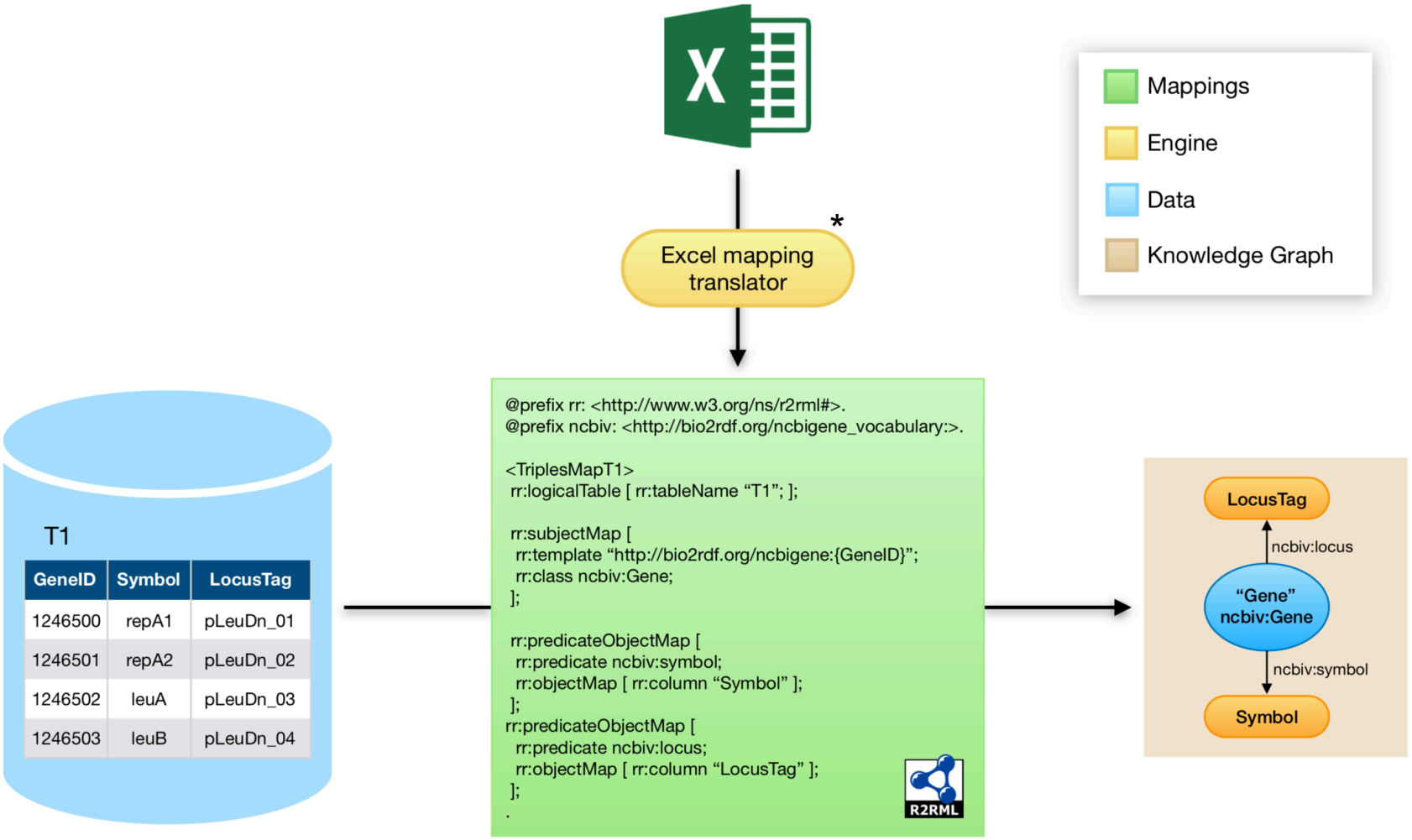


Morph-CSV



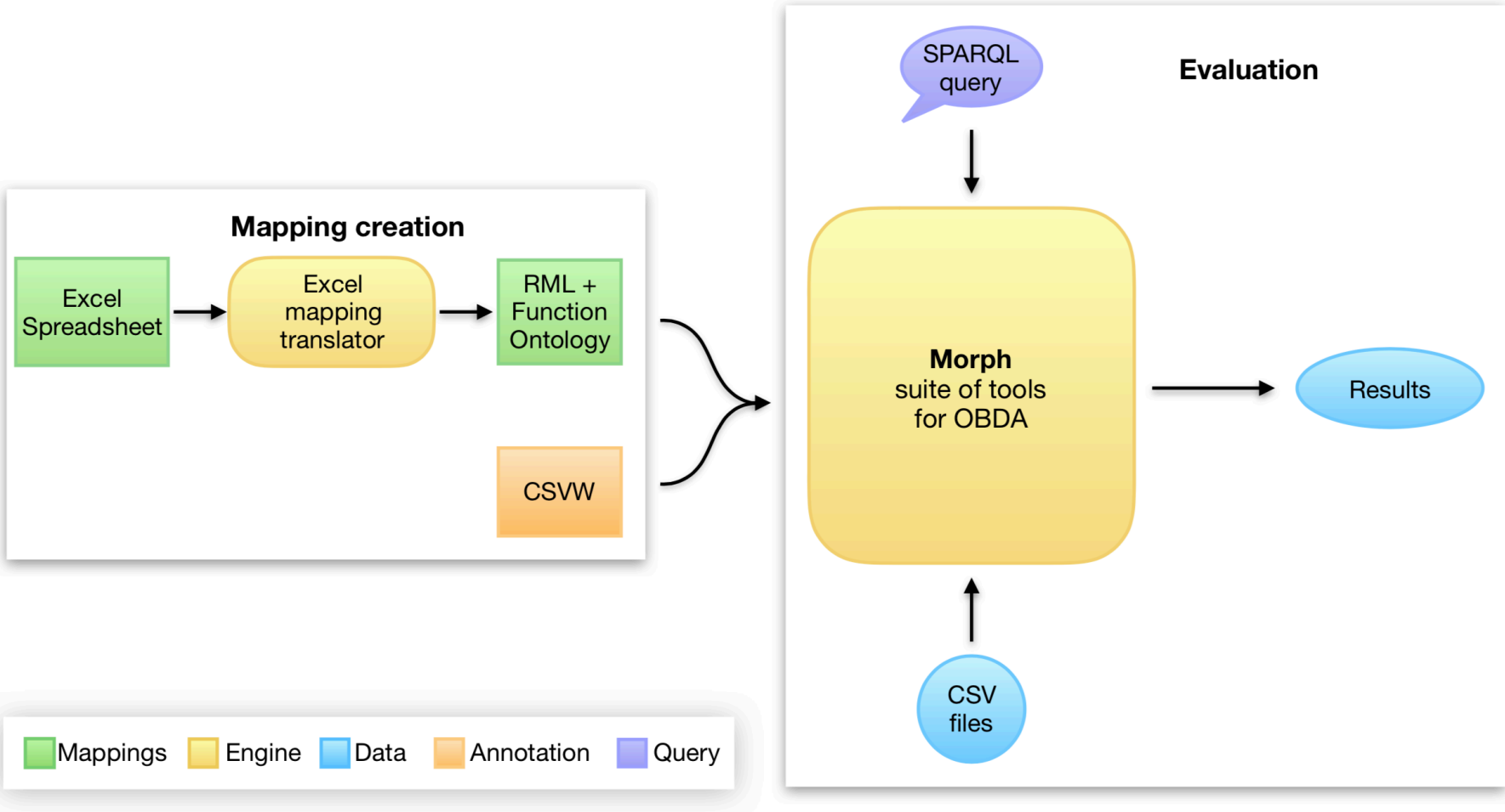
https://github.com/oeg-upm/morph-csv

Mappings



* https://github.com/oeg-upm/Excel-mapping-translator

Workflow


Mapping creation
Excel Spreadsheet
Excel mapping translator
RML + Function Ontology
CSVW
SPARQL query
Evaluation
Morph
suite of tools
for OBDA
Results
CSV files
Mappings
Engine
Data
Annotation
Query

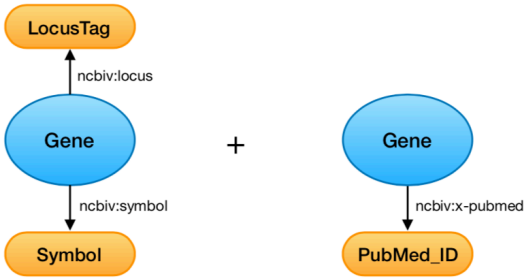
Mapping creation

- **Mappings** and **annotation** files created for 15 datasets*

Database	Number Files	Number joins	Number POM	Size	Predicate Template	SQL View
ClinicalTrials	1	30	197	1 GB	FALSE	TRUE
CTD	8	11	41	3.6 GB	TRUE	TRUE
GenAge	2	3	25	337 KB	FALSE	FALSE
GenDR	1	3	10	85 KB	FALSE	TRUE
HGNC	1	5	48	28 MB	FALSE	TRUE
Homologene	1	2	7	13.8 MB	FALSE	FALSE
iProClass	1	2	25	>8 GB	FALSE	FALSE
iRefIndex	1	11	44	2.77 GB	FALSE	TRUE
LSR	1	2	26	849 KB	FALSE	TRUE
NCBIgene	8	6	54	4.37 GB	TRUE	TRUE
NDC	2	3	27	75 MB	FALSE	TRUE
PharmGKB	3	40	2	18.8 MB	FALSE	FALSE
SIDER	3	0	23	44 MB	FALSE	FALSE
Taxonomy	4	3	25	323.9 MB	TRUE	TRUE
Wormbase	4	6	28	72 MB	FALSE	FALSE
Total	41	127	582	>20 GB		

* https://github.com/anaigmo/Bio2RDF-mappings

Subject from more than one source



+

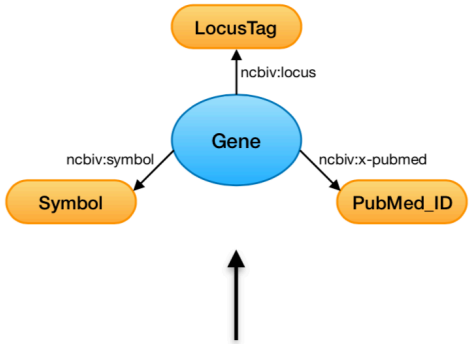
T1

GeneID	Symbol	LocusTag
1246500	repA1	pLeuDn_01
1246501	repA2	pLeuDn_02
1246502	leuA	pLeuDn_03
1246503	leuB	pLeuDn_04

T2

GeneID	PubMed_ID
1246500	9073079
1246501	9073079
1246502	20350329
1246503	129223340

```
CREATE VIEW View-T1T2 AS
SELECT T1.*, T2.PubMed_ID
FROM T1
LEFT JOIN T2;
```



View-T1T2

GeneID	Symbol	LocusTag	PubMed_ID
1246500	repA1	pLeuDn_01	9073079
1246501	repA2	pLeuDn_02	9073079
1246502	leuA	pLeuDn_03	20350329
1246503	leuB	pLeuDn_04	129223340

+

```
<TriplesMapT1>
 rr:logicalTable [ rr:tableName “T1”; ];

 rr:subjectMap [
  rr:template “http://bio2rdf.org/ncbigene:{GeneID}”;
  rr:class ncbiv:Gene;
 ];

 rr:predicateObjectMap [
  rr:predicate ncbiv:symbol;
  rr:objectMap [ rr:column “Symbol” ];
 ];
 rr:predicateObjectMap [
  rr:predicate ncbiv:locus;
  rr:objectMap [ rr:column “LocusTag” ];
 ];
```

R2RML

```
<TriplesMapT2>
 rr:logicalTable [ rr:tableName “T2”; ];

 rr:subjectMap [
  rr:template “http://bio2rdf.org/ncbigene:{GeneID}”;
  rr:class ncbiv:Gene;
 ];

 rr:predicateObjectMap [
  rr:predicate ncbiv:x-pubmed;
  rr:objectMap [ rr:column “PubMed_ID” ];
 ];
```

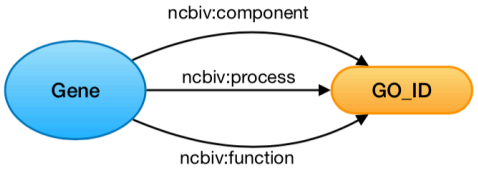
R2RML

+

```
<TriplesMapView>
 rr:logicalTable [ rr:tableName “View-T1T2”; ];
  rr:subjectMap [
  rr:template “http://bio2rdf.org/ncbigene:{GeneID}”;
  rr:class ncbiv:Gene;
 ];
 rr:predicateObjectMap [
  rr:predicate ncbiv:symbol;
  rr:objectMap [ rr:column “Symbol” ];
 ];
 rr:predicateObjectMap [
  rr:predicate ncbiv:locus;
  rr:objectMap [ rr:column “LocusTag” ];
 ];
 rr:predicateObjectMap [
  rr:predicate ncbiv:x-pubmed;
  rr:objectMap [ rr:column “PubMed_ID” ];
 ];
```

R2RML

Predicate Template



SQL

CREATE VIEW VCom

CREATE VIEW VPro

CREATE VIEW VFun

CREATE VIEW View-T1

T1

GeneID	GO_ID	Category
814629	GO:0005634	Component
814629	GO:0008150	Process
814630	GO:0003700	Function

View-T1

GeneID	Component	Process	Function
814629	GO:0005634	NULL	NULL
814629	NULL	GO:0008150	NULL
814630	NULL	NULL	GO:0003700

+

```
<TriplesMapT1>
 rr:logicalTable [ rr:tableName “T1”; ];

 rr:subjectMap [
  rr:template “http://bio2rdf.org/ncbigene:{GeneID}”;
  rr:class ncbiv:Gene;
 ];

 rr:predicateObjectMap [
  rr:predicateMap [ rr:template ncbiv:{Category} ] ;
  rr:objectMap [ rr:column “GO_ID” ];
 ];
```

R2RML

+

```
<TriplesMapView>
 rr:logicalTable [ rr:tableName “View-T1”; ];
  rr:subjectMap [
  rr:template “http://bio2rdf.org/ncbigene:{GeneID}”;
  rr:class ncbiv:Gene;
 ];
 rr:predicateObjectMap [
  rr:predicate ncbiv:component;
  rr:objectMap [ rr:column “Component” ];
 ];
 rr:predicateObjectMap [
  rr:predicate ncbiv:process;
  rr:objectMap [ rr:column “Process” ];
 ];
rr:predicateObjectMap [
  rr:predicate ncbiv:function;
  rr:objectMap [ rr:column “Function” ];
 ];
```

R2RML

Evaluation

- Subset of 7 datasets chosen for evaluation with 10 SPARQL queries with a limit of 10 results*

Query	Sources	Triple Patterns	Time (ms)	Description
Q01	1	4	93	Symbol and RefSeq reference of groups of homologous genes
Q02	2	4	244	Locus of groups of homologous genes
Q03	4	5	861	Gene's references to UniGene, STS and PubMed
Q04	3	4	1164289	Aging related genes from worm species
Q05	2	4	69	Cosmids from genes interacting in worms
Q06	3	6	68	Division, genetic code and division from taxonomy nodes
Q07	2	5	75452	Taxonomy and unique name of aging related genes
Q08	1	4	198	Description of side effects
Q09	2	5	2578	Package and start marketing date of products
Q10	5	7	860	Outcome, analysis and measurement of a clinical study

* https://github.com/anaigmo/Bio2RDF-MorphCSV

Evaluation

- Q01 and Q08 don't contain joins with other triples maps

Query	Sources	Triple Patterns	Time (ms)	Description
Q01	1	4	93	Symbol and RefSeq reference of groups of homologous genes
Q02	2	4	244	Locus of groups of homologous genes
Q03	4	5	861	Gene's references to UniGene, STS and PubMed
Q04	3	4	1164289	Aging related genes from worm species
Q05	2	4	69	Cosmids from genes interacting in worms
Q06	3	6	68	Division, genetic code and division from taxonomy nodes
Q07	2	5	75452	Taxonomy and unique name of aging related genes
Q08	1	4	198	Description of side effects
Q09	2	5	2578	Package and start marketing date of products
Q10	5	7	860	Outcome, analysis and measurement of a clinical study

Evaluation

- Q03, Q05, Q06, Q09 and Q10 require joins between triples maps in the same dataset. Some of them need the creation of SQL Views

Query	Sources	Triple Patterns	Time (ms)	Description
Q01	1	4	93	Symbol and RefSeq reference of groups of homologous genes
Q02	2	4	244	Locus of groups of homologous genes
Q03	4	5	861	Gene's references to UniGene, STS and PubMed
Q04	3	4	1164289	Aging related genes from worm species
Q05	2	4	69	Cosmids from genes interacting in worms
Q06	3	6	68	Division, genetic code and division from taxonomy nodes
Q07	2	5	75452	Taxonomy and unique name of aging related genes
Q08	1	4	198	Description of side effects
Q09	2	5	2578	Package and start marketing date of products
Q10	5	7	860	Outcome, analysis and measurement of a clinical study

Evaluation

- Q02, Q04 and Q07 link different datasets

Query	Sources	Triple Patterns	Time (ms)	Description
Q01	1	4	93	Symbol and RefSeq reference of groups of homologous genes
Q02	2	4	244	Locus of groups of homologous genes
Q03	4	5	861	Gene's references to UniGene, STS and PubMed
Q04	3	4	1164289	Aging related genes from worm species
Q05	2	4	69	Cosmids from genes interacting in worms
Q06	3	6	68	Division, genetic code and division from taxonomy nodes
Q07	2	5	75452	Taxonomy and unique name of aging related genes
Q08	1	4	198	Description of side effects
Q09	2	5	2578	Package and start marketing date of products
Q10	5	7	860	Outcome, analysis and measurement of a clinical study

Challenges

- The queries were tested with success, what proves that the **feasibility** of the methodology proposed

- However, some issues arose in the evaluation:

  1. CSV formatting
  2. Primary Keys
  3. *Ad hoc* data treatment
  4. Big Data

CSV formatting

- CSVs are not always well formatted. We have encountered 2 cases:
  - The field separators are different between fields and at the end of the line

2 | Bacteria | Bacteria <prokaryotes> | scientific name |

  - There are missing separators when there are blank fields at the end of the line

GeneID	Symbol	tax_id
1246500	repA1	
1246501		

GeneID,Symbol,tax_id
1246500,repA1
1246501

Primary Keys

- An essential step in the methodology proposed is the creation of a database from CSV files
- Sometimes, there is not a field or combination of them that serve as primary key because some of the fields contain NULL values.
- As a result, **data is lost** in the conversion

tax_id	GeneID	Ensembl_gene	Ensembl_rna	Ensembl_protein
7227	30970	FBgn0040373	NULL	FBpp0309182
7227	30970	FBgn0040373	FBtr0070108	FBpp0070103
7227	30971	FBgn0040372	NULL	NULL

Ad hoc data treatment

- The need to create SQL Views, format CSVs when their format are wrong, and finding a way to upload data to avoid primary keys related errors makes it necessary to alter the data **manually**
- Then, for each case **the workflow changes**
- The original proposal is to use this methodology is to provide a unified way to access data
- Altering the data doesn't fit with this purpose

Big data

- Just the datasets with data published as CSV have a size of more than 20 GB
- Two bottlenecks:
  1. **Creation** of the database
  2. **Querying** the database
- Databases can be created one by one. Then it becomes necessary a **federator** to query all the resulting databases

Conclusions

1. This work is intended to be a first approach to change the methodology of Bio2RDF using OBDA

2. Several mappings have been created for the data sources that have their data published as CSV and relational database

3. The mappings easily maintained and can be reused in other OBDA engines

4. An evaluation of the process has been made with a subset of the data sources

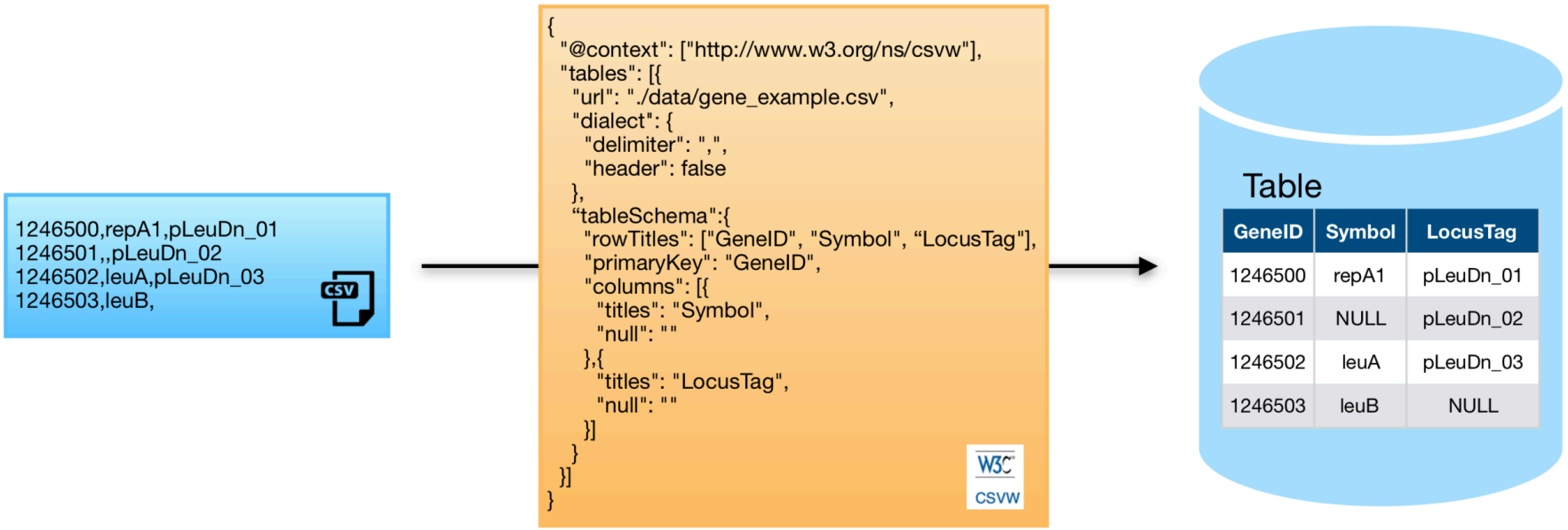
5. The issues that appeared on the evaluation are described in order to improve the methodology proposed

Future work

- A review on the methodology is necessary in order to improve it and be able to deal with the issues described from the evaluation, so that the user has to make the minimum changes to the data
  - More options have to be included in the annotations file to deal with inconsistent separators, automatic creation of SQL views and Primary Keys issues
  - New solutions have to be explored to the Big Data problem, whether using federators or data in streams
- New approaches are to be considered to treat data with other formats, such as JSON or XML

Questions?

Annotation file (CSVW)



SQL queries in Predicate Template

```
CREATE VIEW VCom AS
SELECT GeneID, GO_ID as Component
FROM T1
WHERE Category = “Component”;
```


```
CREATE VIEW ViewPro AS
SELECT GeneID, GO_ID as Process
FROM T1
WHERE Category = “Process”;
```


```
CREATE VIEW ViewFun AS
SELECT GeneID, GO_ID as Function
FROM T1
WHERE Category = “Function”;
```


```
CREATE VIEW View-T1 AS
SELECT GeneID, VCom.Component,
  VPro.Process, VFun.Function
FROM T1
LEFT JOIN VCom ON
  VCom.Component = T1.GO_ID
LEFT JOIN VPro ON
  VPro.Process = T1.GO_ID
LEFT JOIN VFun ON
  VFun.Function = T1.GO_ID;
```

Queries: Group 1

Listing 1.8. Q08

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sidv: <http://bio2rdf.org/sider_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?se ?label ?description ?comment
WHERE {
        ?se rdf:type sidv:Side-Effect ;
                rdfs:label ?label ;
                rdfs:description ?descripcion ;
                rdfs:comment ?comment.
}
```

Listing 1.1. Q01

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX homogv: <http://bio2rdf.org/homologene_vocabulary:>

SELECT ?hgroup ?label ?symbol ?refseq
WHERE {
        ?hgroup rdf:type homogv:Resource ;
                rdfs:label ?label ;
                homogv:gene-symbol ?symbol ;
                homogv:x-refseq ?refseq .
}
```

Queries: Group 2

Listing 1.3. Q03

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ncbiv: <http://bio2rdf.org/ncbigene_vocabulary:>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT ?gene ?id ?unigene ?pubmed ?unists
WHERE {
        ?gene rdf:type ncbiv:Resource ;
                dcterms:identifier ?id ;
                ncbiv:x-unigene ?unigene ;
                ncbiv:x-pubmed ?pubmed ;
                ncbiv:x-unists ?unists .
}
```

Listing 1.5. Q05

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX wbv: <http://bio2rdf.org/wormbase_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?interaction ?gene ?title ?cosmid
WHERE {
        ?interaction rdf:type wbv:Resource ;
                wbv:involves ?gene .
        ?gene dcterms:title ?title ;
                wbv:cosmid ?cosmid .
}
```

Queries: Group 2

Listing 1.6. Q06

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX taxv: <http://bio2rdf.org/taxonomy_vocabulary:>

SELECT ?tax ?div ?gencode ?mitcode ?parent ?cde
WHERE {
        ?tax rdf:type taxv:Resource ;
                taxv:division-id ?div ;
                taxv:genetic-code-id ?gencode ;
                taxv:mit-genetic-code-id ?mitcode ;
                rdfs:subClassOf ?parent .
        ?gencode taxv:translation-table ?cde .
}
```

Listing 1.9. Q09

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ndcv: <http://bio2rdf.org/ndc_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?prod ?prod_type ?date ?package ?pack_label
WHERE {
        ?prod rdf:type ndcv:Product ;
                ndcv:product-type ?prod_type ;
                ndcv:start-marketing-date ?date;
                ndcv:has-package ?package .
        ?package rdfs:label ?pack_label .
}
```

Queries: Group 2

Listing 1.10. Q10

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ctv: <http://bio2rdf.org/clinicaltrials_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?study ?label ?summary ?outcome ?measurement ?type
WHERE {
        ?study rdf:type ctv:Clinical-Study ;
                rdfs:label ?label ;
        ctv:brief-summary ?summary ;
        ctv:outcome ?outcome .
    ?outcome ctv:analysis ?analysis ;
        ctv:measurement ?measurement ;
        ctv:type ?type .
}
```

Queries: Group 3

Listing 1.4. Q04

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX wbv: <http://bio2rdf.org/wormbase_vocabulary:>
PREFIX genagev: <http://bio2rdf.org/genage_vocabulary:>

SELECT ?gene_worm ?tax ?gene_age
WHERE {
        ?gene_worm rdf:type wbv:Gene ;
                wbv:x-taxid ?tax .
        ?gene_age rdf:type genagev:Aging-Related-Gene ;
                genagev:taxon ?tax .
}
```

Listing 1.2. Q02

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX homogv: <http://bio2rdf.org/homologene_vocabulary:>
PREFIX ncbiv: <http://bio2rdf.org/ncbigene_vocabulary:>

SELECT ?hgroup ?gene ?label ?locus
WHERE {
        ?hgroup rdf:type homogv:Resource ;
                homogv:x-ncbigene ?gene .
        ?gene rdfs:label ?label ;
                ncbiv:locus ?locus .
}
```

Queries: Group 3

Listing 1.7. Q07

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX genagev: <http://bio2rdf.org/genage_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX taxv: <http://bio2rdf.org/taxonomy_vocabulary:>

SELECT ?gene ?label ?tax ?name ?uname
WHERE {
        ?gene rdf:type genagev:Aging-Related-Gene ;
                rdfs:label ?label ;
                genagev:taxon ?tax .
        ?tax rdfs:label ?name ;
                taxv:unique-name ?uname .
}
```