Linked Data for functional genomics

Mikel Egaña Aranguren

3205 School of Computer Science
Universidad Politécnica de Madrid (UPM)
28660 Boadilla del Monte
Spain

Ontology Engineering Group (OEG)
http://www.oeg-upm.net

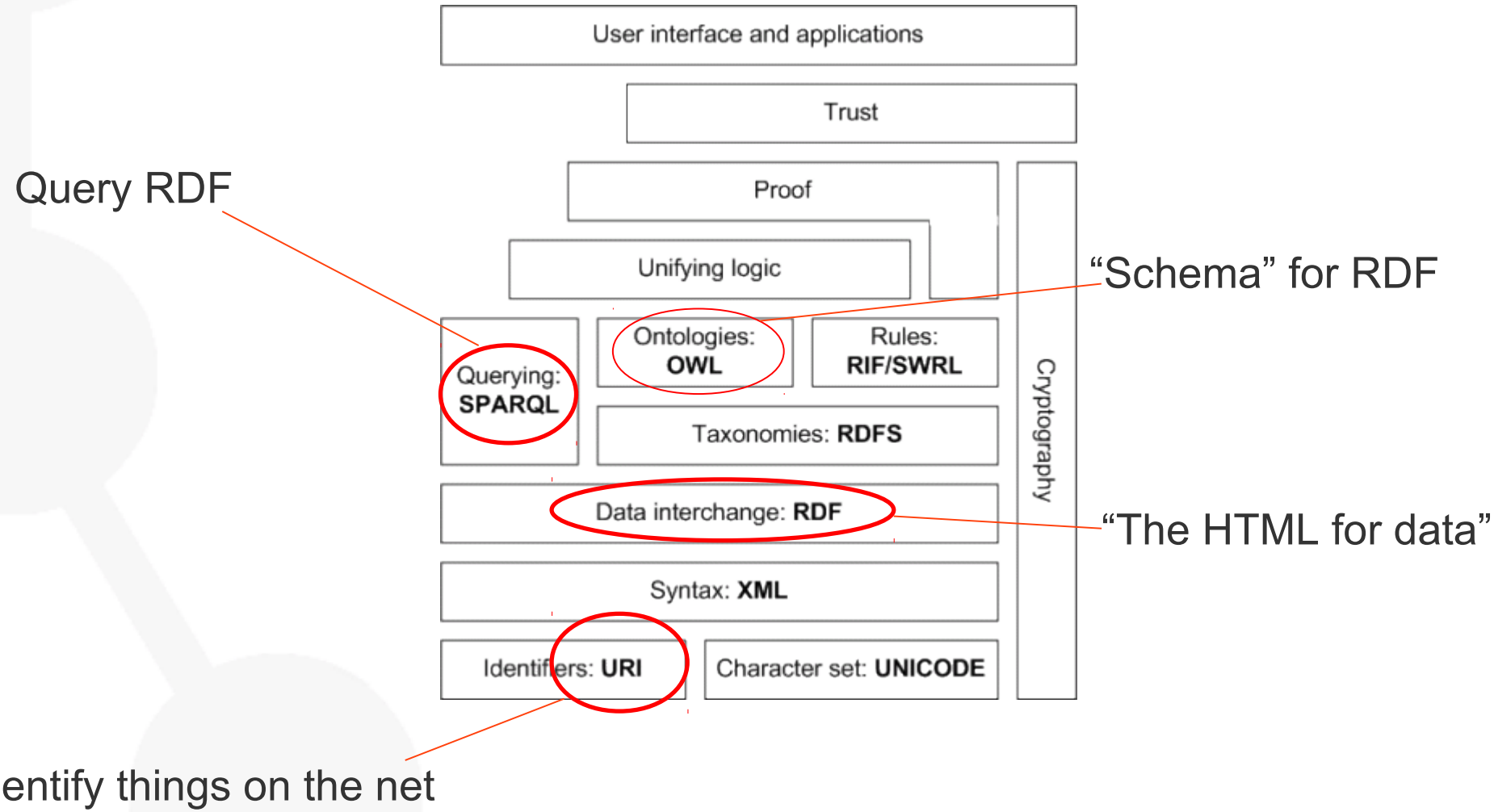
megana@fi.upm.es
http://mikeleganaaranguren.com

http://www.slideshare.net/MikelEganaAranguren/linked-data-functional-genomics

Index

What is Linked Data?

A first step towards the Semantic Web



Linked Data principles

1) Use URIs as names for things

2) Use HTTP URIs so that people can look up those names

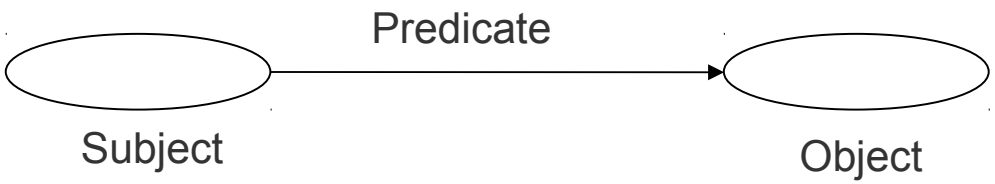
3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)

4) Include links to other URIs so that they can discover more things

http://www.w3.org/DesignIssues/LinkedData.html

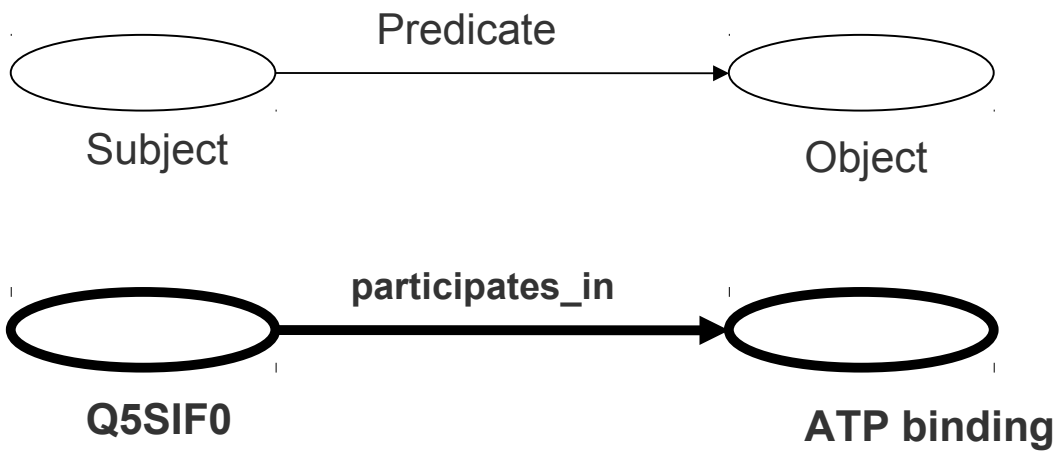
With Linked Data we publish data

Semantically: the data model is explicit for computers (RDF triple)



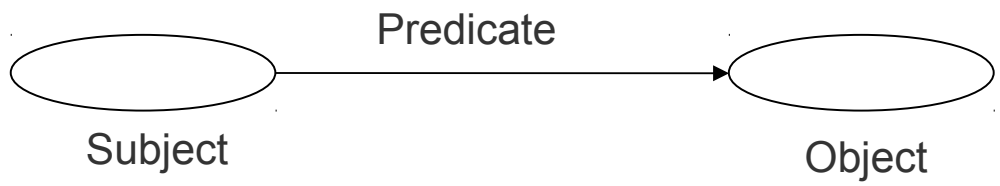
With Linked Data we publish data

Semantically: the data model is explicit for computers (RDF triple)

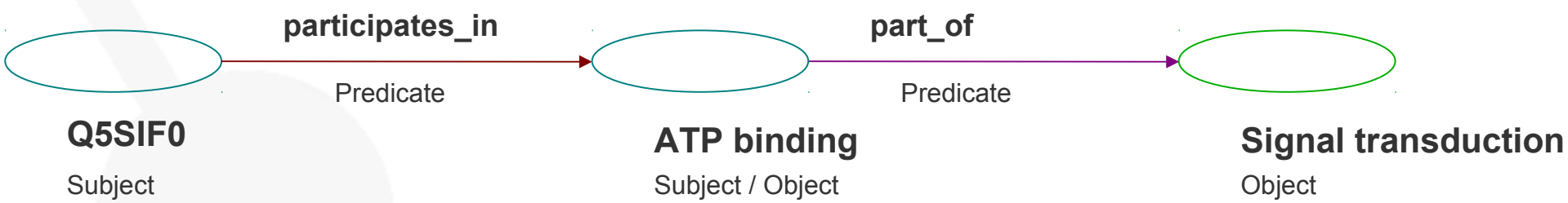


With Linked Data we publish data

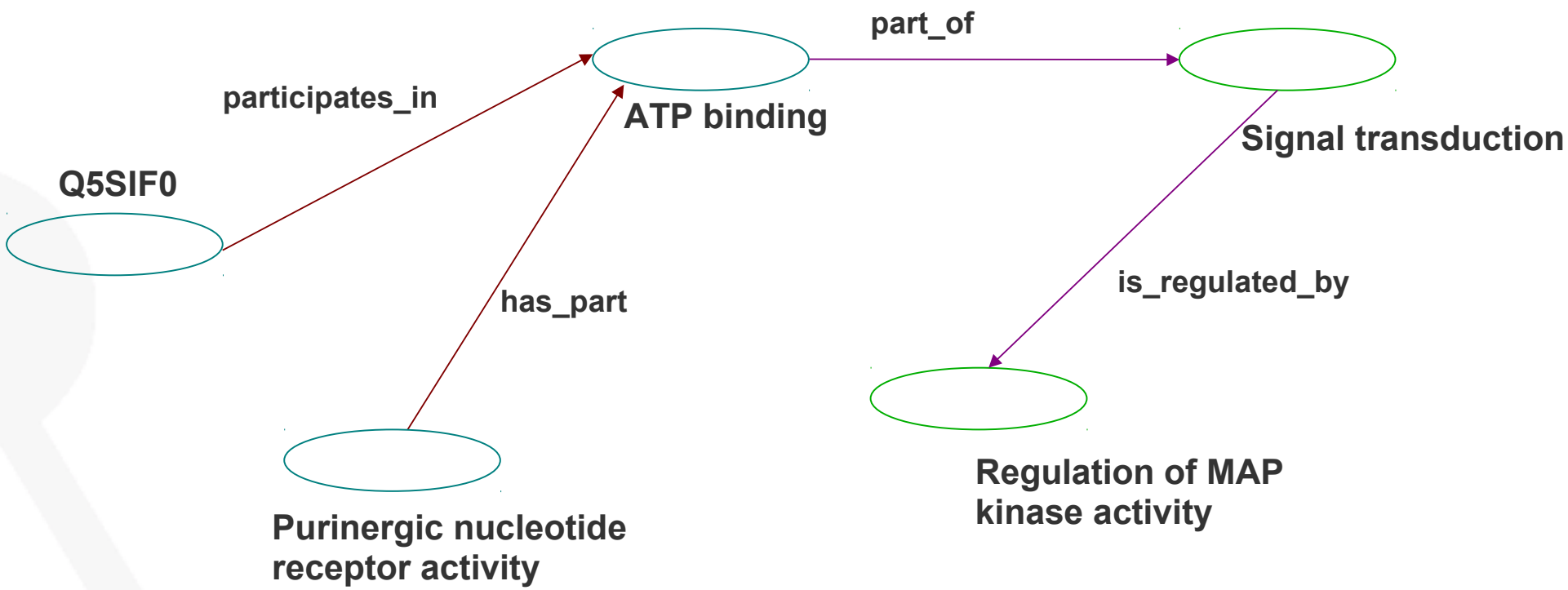
Semantically: the data model is explicit for computers (RDF triple)



Inter-linked: the data is linked to data from other resources over the web



Global network of Linked Data



What is Linked Data?

Internet of **data** rather than **documents**, a “universal DB”

Find precisely what we are looking for: direct queries rather than text processing (SPARQL)

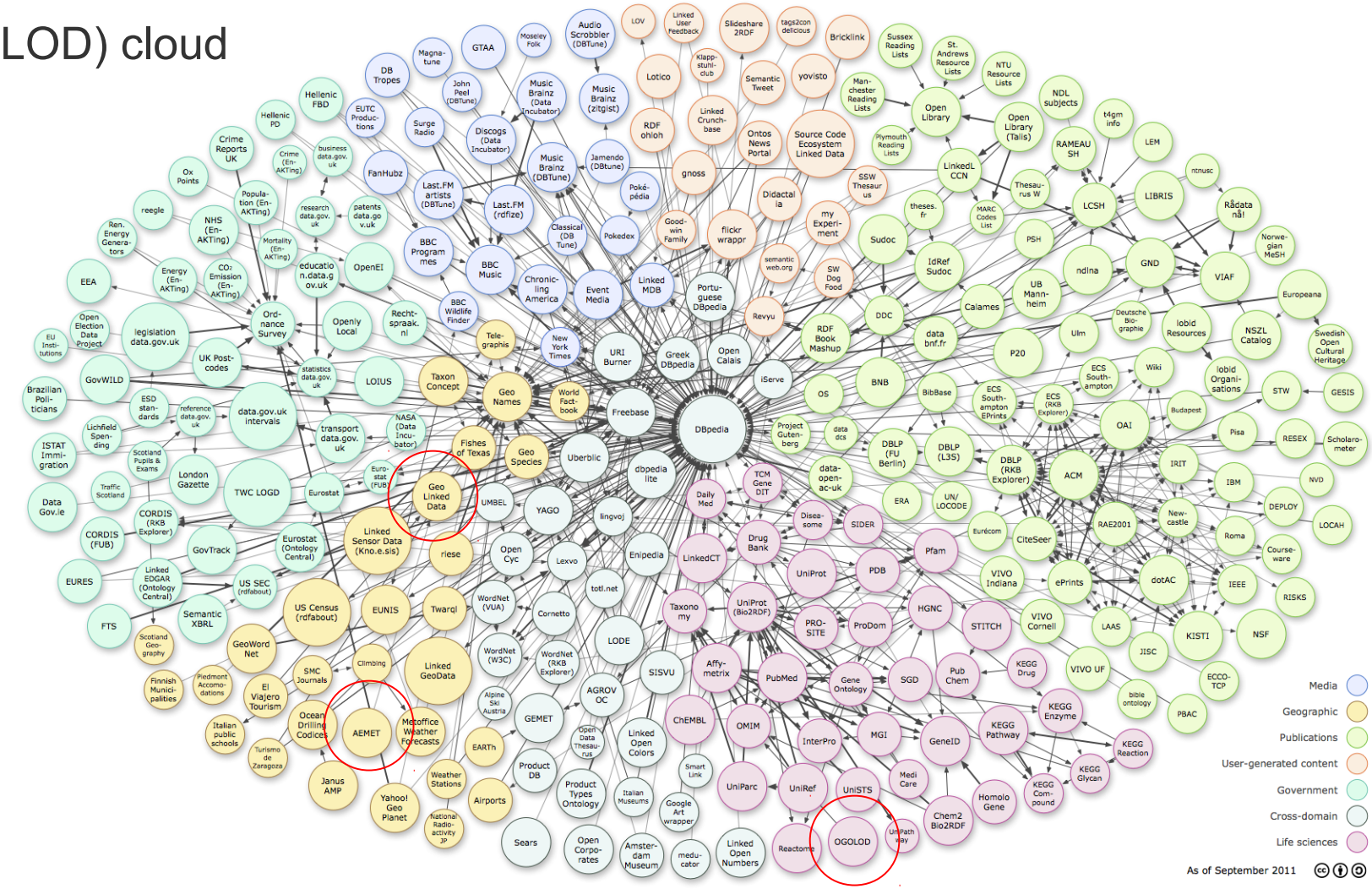
Linking new data is as easy as linking a web page

We can navigate through the data directly (RDF), rather than navigating through documents that represent data in natural language (HTML)

Build applications that exploit the data

Apply automated reasoning on the data

Linked Open Data (LOD) cloud

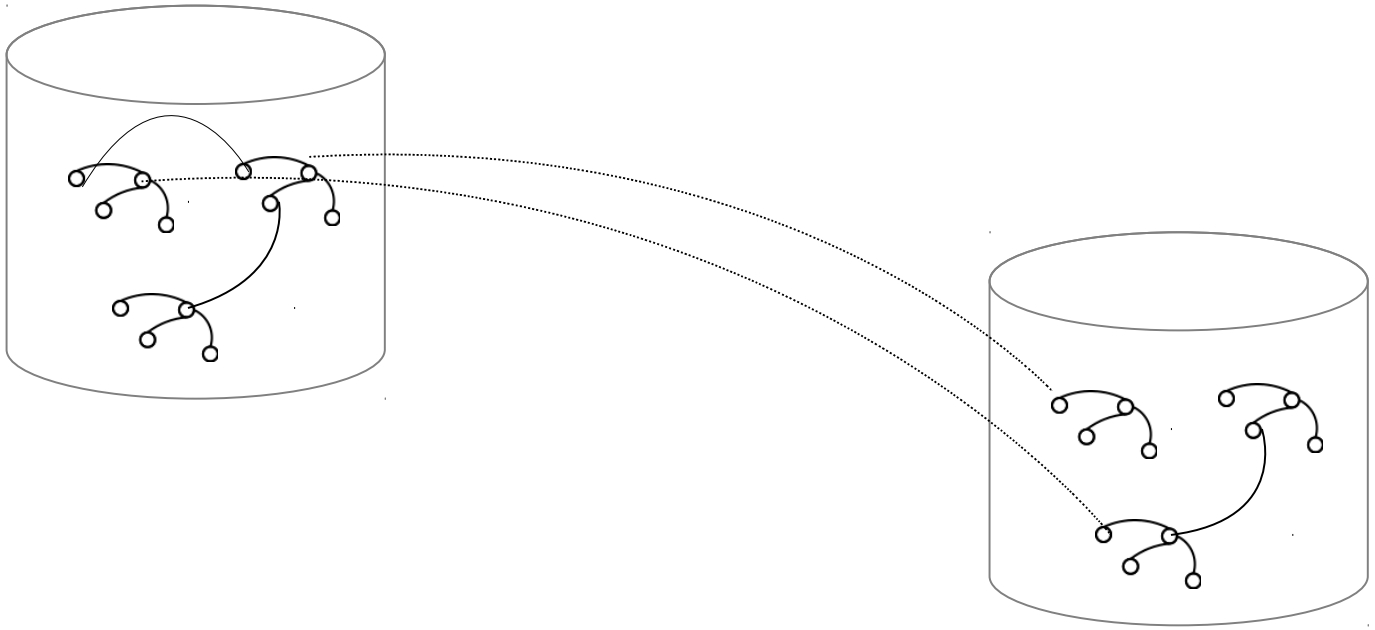


http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.html

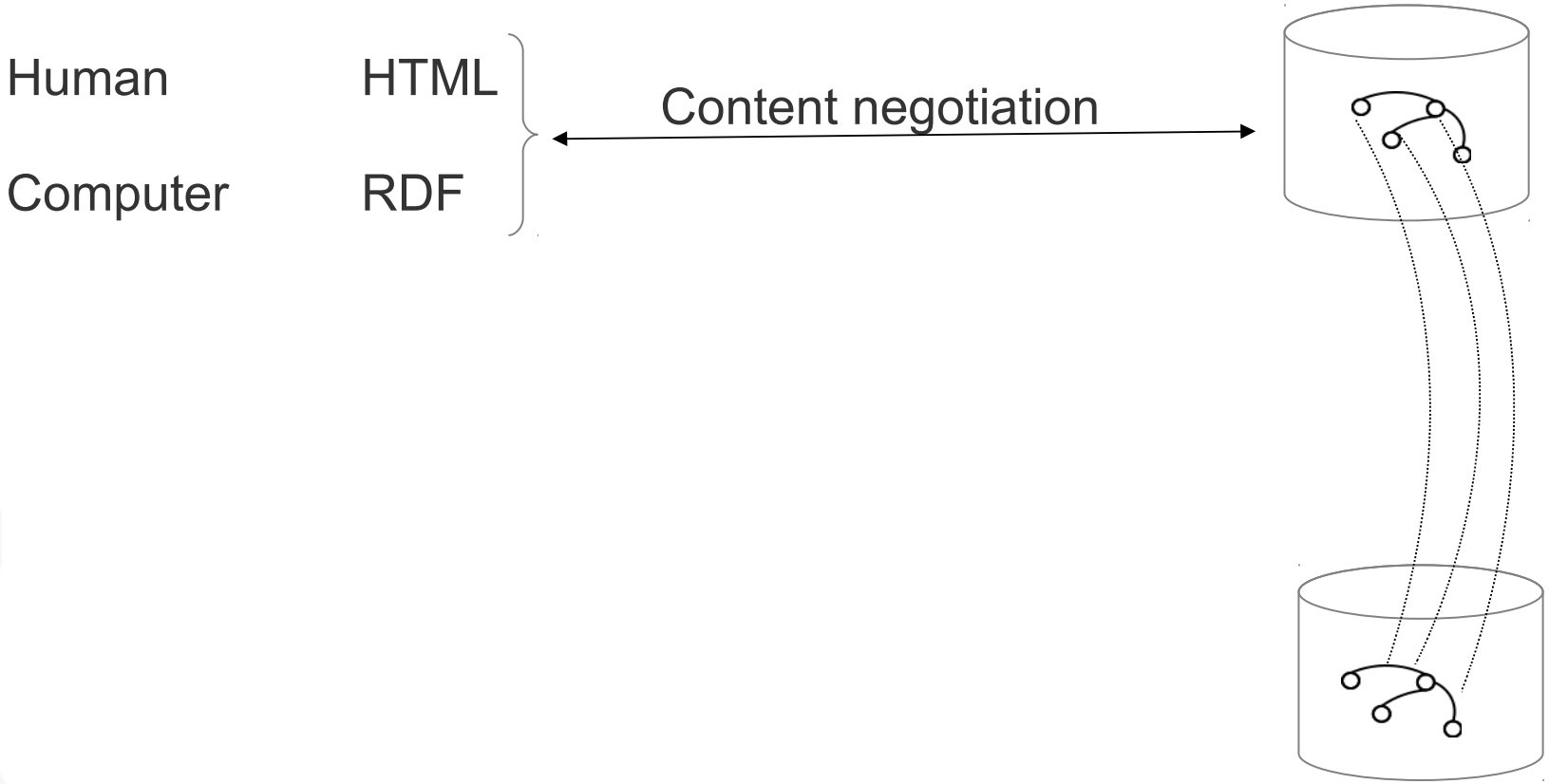
What is Linked Data?

A graph is a collection of RDF triples

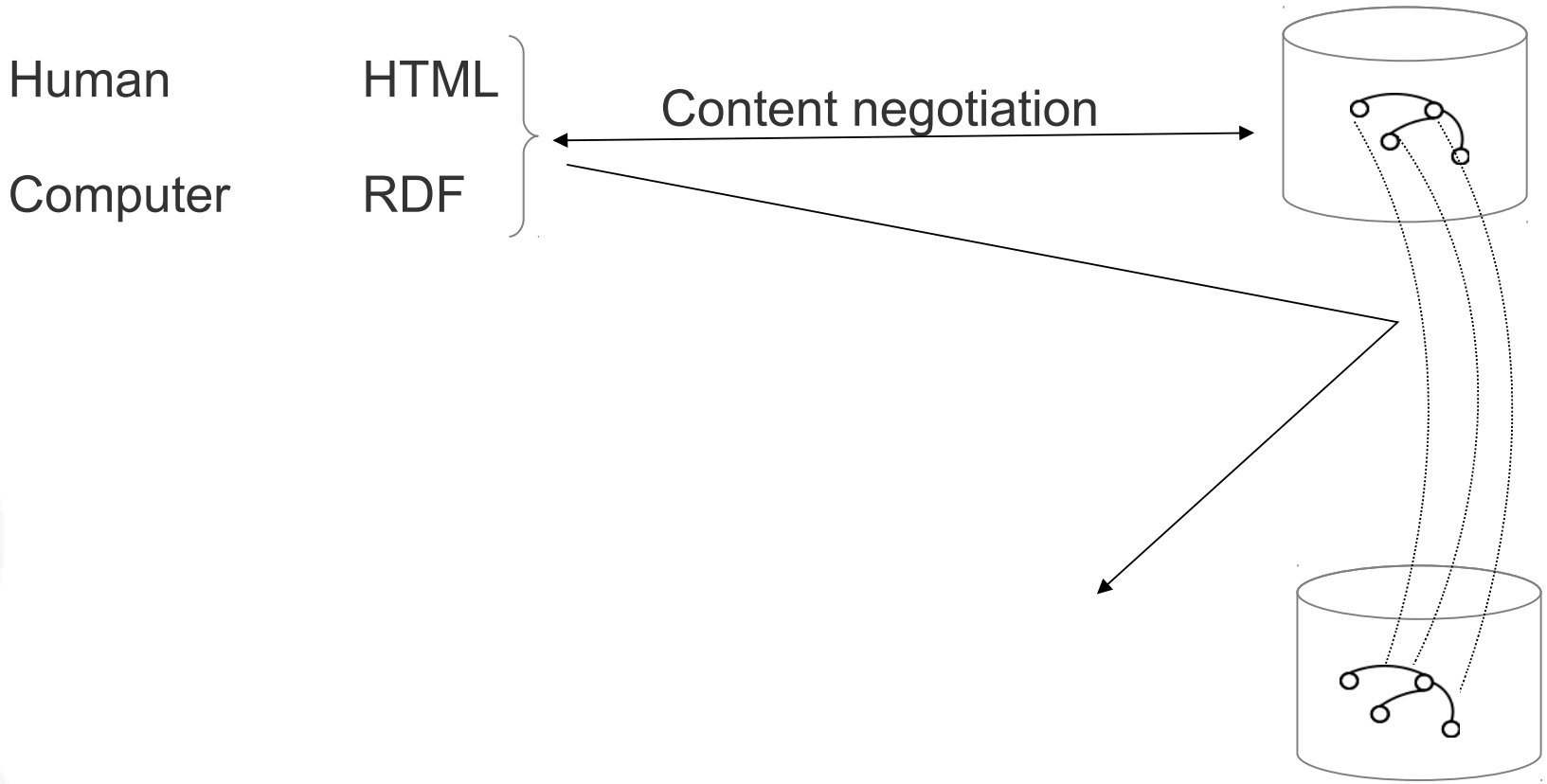
A triple store holds different graphs



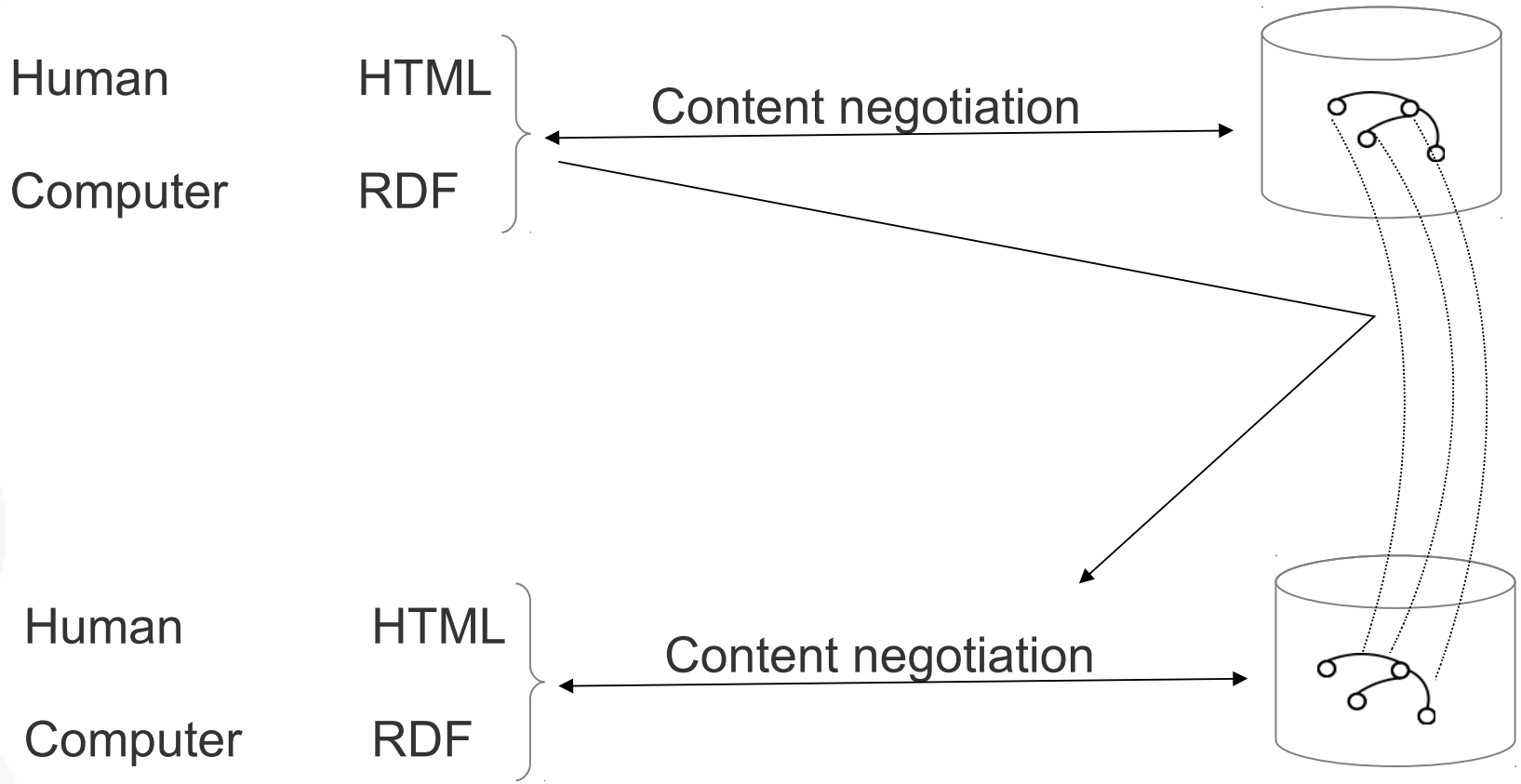
What is Linked Data?



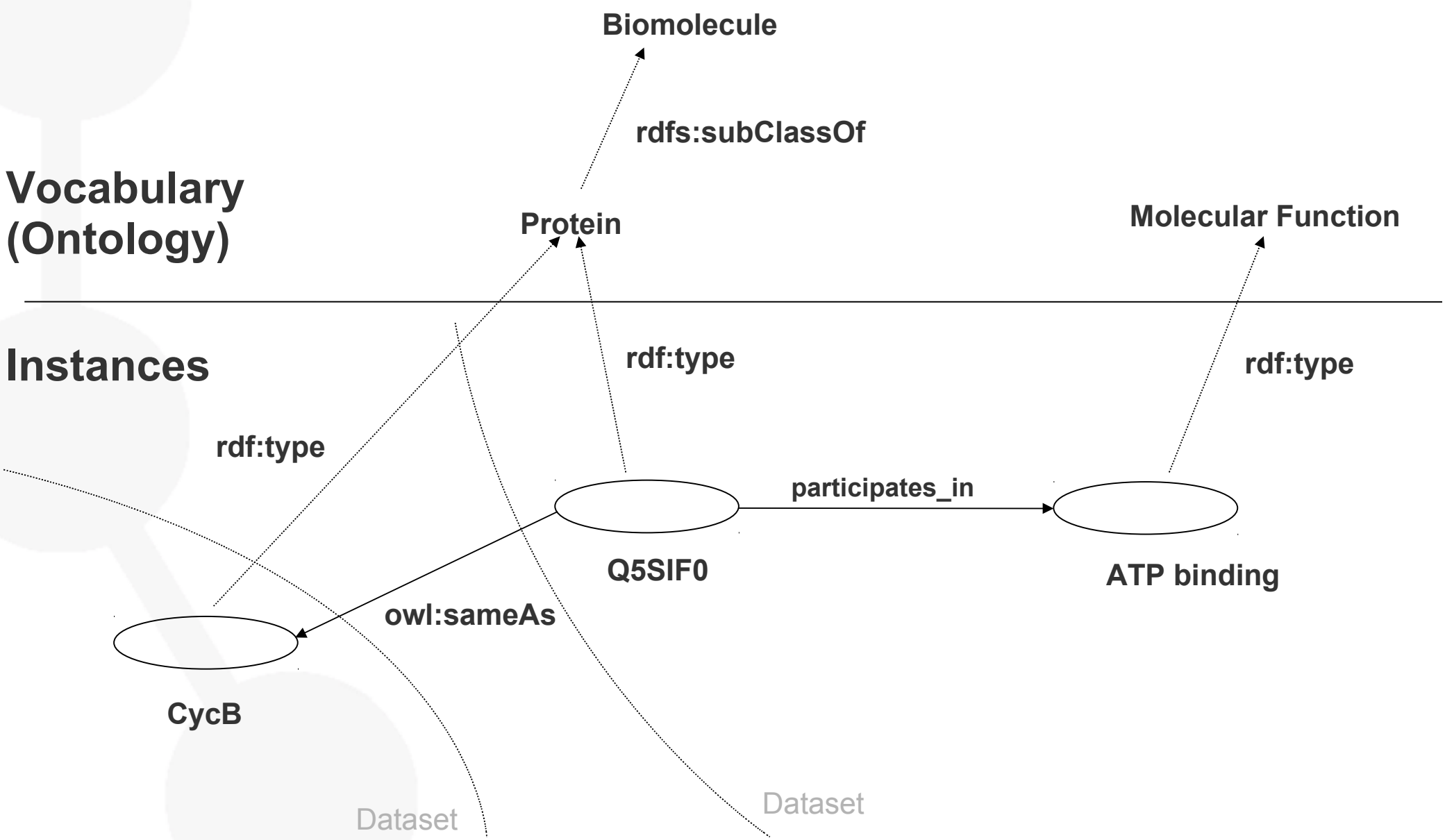
What is Linked Data?



What is Linked Data?



What is Linked Data?



Consuming Linked Data

Navigate

Query

Meshups

Consuming Linked Data

Bio2RDF (http://bio2rdf.org/)

OGOLOD (http://miuras.inf.um.es/~ogo/ogolod.html)

LinkedLifeData (http://linkedlifedata.com/)

HyQue* (http://semanticscience.org/projects/hyque/index.html)

ArrayExpress and Gene expression atlas*
(http://www.ebi.ac.uk/arrayexpress/)

* not really LD, close to LD, but (likely) soon will be full LD

Navigate

fbmlapp id: 129170987115694 [3,5]
fbmladmins: 100000304287730 [3,5]
gene symbol: CYCB [1]
identifier: chem2bio2rdf.org/rdf/resource/hgnc/CYCB [1,2]
www.reference.com/browse/CYCB [3]
likeorhate.com/thing/706106/CYCB [4]
www.reference.com/browse/CYCB%20%28disambiguation%29 [5]
bio2rdf.org/go:association-41876 [6]
bio2rdf.org/go:association-41887 [7]
bio2rdf.org/go:association-41883 [8]
bio2rdf.org/go:association-41877 [9]
bio2rdf.org/go:association-41880 [10]
bio2rdf.org/go:association-41888 [11]
bio2rdf.org/go:association-321976 [12]
bio2rdf.org/go:association-41874 [13]
bio2rdf.org/go:association-41885 [14]
bio2rdf.org/go:association-41875 [15]
bio2rdf.org/go:association-41886 [16]
bio2rdf.org/go:association-41878 [17]
bio2rdf.org/go:association-41881 [18]
bio2rdf.org/go:association-41873 [19]
bio2rdf.org/go:association-41884 [20]
label: CYCB [1,2,4]
the encyclopedic entry of CYCB [3]
Cycb encyclopedia topics | Reference.com [3]
LikeOrHate.com - Express your opinion: CYCB [4]
the encyclopedic entry of CYCB (disambiguation) [5]
Cycb (disambiguation) encyclopedia topics | Reference.com [5]
CycB [go:association-41876] [6]
CycB [go:association-41887] [7]
CycB [go:association-41883] [8]
CycB [go:association-41877] [9]
CycB [go:association-41880] [10]
CycB [go:association-41888] [11]
cycB [go:association-321976] [12]
CycB [go:association-41874] [13]
CycB [go:association-41885] [14]
CycB [go:association-41875] [15]
CycB [go:association-41886] [16]
CycB [go:association-41878] [17]
CycB [go:association-41881] [18]
CycB [go:association-41873] [19]
CycB [go:association-41884] [20]

Sources (20) | Approved (0) | Rejected (0)

1 CYCB 8 facts | 2010-06-03
indice http://chem2bio2rdf.org/uniprot/resource/...
2 CYCB 8 facts | 2011-10-21
indice http://dbpedia.org/resource/CYCB
3 Cycb encyclopedia topics... 7 facts | 2011-09-30
indice http://www.reference.com/browse/CYCB
4 LikeOrHate.com - Express... 4 facts | 2011-08-14
indice http://likeorhate.com/thing/706106/CYCB
5 Cycb (disambiguation) en... 7 facts | 2011-09-30
indice http://www.reference.com/browse/CYCB%20%2...
6 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41876
7 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41887
8 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41883
9 CycB [go:association-418... 10 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41877
10 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41880
11 CycB [go:association-418... 11 facts | 2009-08-20
indice http: solo approve ✔ ✔✔ reject ✖ ✖✖
12 cycB [go:association-321... 10 facts | 2009-08-20
indice http://bio2rdf.org/go:association-321976 (cache)
13 CycB [go:association-418... 10 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41874
14 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41885
15 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41875
16 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41886
17 CycB [go:association-418... 11 facts | 2009-08-20
indice http://bio2rdf.org/go:association-41878

<- 1 2 ->
reject all | approve all
http://example.loc/document.rdf
add source url

Navigate

at **miuras.inf.um.es**

http://miuras.inf.um.es/ogolod/resource/Gene/67440

Property	Value
ogoont:Identifier	67440
ogoont:Name	0610027A18Rik AW551379 Mtpap Papd1
ogoont:fromSpecies	ogores:NCBITaxon_10090/NCBITaxon_10090
ogoont:isTranslatedTo	ogores:Protein/21312970 ogores:Protein/Q9D0D3
ogoont:located_IDA_in	ogores:GO_0005739/GO_0005739
ogoont:located_IEA_in	ogores:GO_0005737/GO_0005737
ogoont:participates_IEA_in	ogores:GO_0000166/GO_0000166 ogores:GO_0003723/GO_0003723 ogores:GO_0004652/GO_0004652 ogores:GO_0005524/GO_0005524 ogores:GO_0006350/GO_0006350 ogores:GO_0006397/GO_0006397 ogores:GO_0016740/GO_0016740
ogoont:participates_ISO_in	ogores:GO_0071044/GO_0071044
ogoont:participates_ND_in	ogores:GO_0003674/GO_0003674
owl:sameAs	b2r:page/geneid:67440
rdf:type	ogoont:Gene

This page shows information obtained from the SPARQL endpoint at http://miuras.inf.um.es/sparql.

As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Query different resources combining the information

Select all located in Y-chromosome, human genes with known molecular interactions, which are analysed with 'Transfection'

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX gene: <http://linkedlifedata.com/resource/entrezgene/>
PREFIX core: <http://purl.uniprot.org/core/>
PREFIX biopax2: <http://www.biopax.org/release/biopax-level2.owl#>
PREFIX lifeskim: <http://linkedlifedata.com/resource/lifeskim/>
PREFIX umls: <http://linkedlifedata.com/resource/umls/>
PREFIX pubmed: <http://linkedlifedata.com/resource/pubmed/>

SELECT distinct ?genedescription ?prefLabel
WHERE {
   ?p biopax2:PHYSICAL-ENTITY ?protein .
   ?protein skos:exactMatch ?uniprotaccession .
   ?uniprotaccession core:organism <http://purl.uniprot.org/taxonomy/9606> .
   ?geneid gene:uniprotAccession ?uniprotaccession .
   ?geneid gene:description ?genedescription .
   ?geneid gene:pubmed ?pmid .
   ?geneid gene:chromosome 'Y' .
   ?pmid lifeskim:mentions ?umlsid .
   ?umlsid skos:prefLabel 'Transfection' .
   ?umlsid skos:prefLabel ?prefLabel .
}
```

(http://linkedlifedata.com/sparql)

Query different resources combining the information

We will receive only the triples of that triple store (but we can follow the links to the triples stored in other triple stores!)

For retrieving triples from other triple stores we need federated queries:

SERVICE keyword in SPARQL 1.1

```
PREFIX foaf:   <http://xmlns.com/foaf/0.1/>
SELECT ?name
FROM <http://example.org/myfoaf.rdf>
WHERE
{
  <http://example.org/myfoaf/I> foaf:knows ?person .
  SERVICE <http://people.example.org/sparql> {
    ?person foaf:name ?name . }
}
```

http://www.w3.org/TR/sparql11-federated-query/

Hypotheses evaluation with HyQue*

http://semanticscience.org/projects/hyque/

```
PREFIX hybrow: <http://bio2rdf.org/hybrow:>
PREFIX semsci: <http://semanticscience.org/resource/>
PREFIX bio2rdf: <http://bio2rdf.org/ns/bio2rdf:>


select DISTINCT * where {
        ?event rdfs:label ?label .
        ?event rdf:type ?event_type .
        ?event_type rdfs:label ?event_type_label .
        ?event hybrow:is_negated ?negated .
        ?event hybrow:physical_context ?event_location .
        ?event hybrow:physical_operator ?physical_operator .
        ?event hybrow:agent_a ?actor .
        ?event hybrow:agent_b ?target .
        OPTIONAL {
        { ?actor rdfs:subClassOf ?actor_type  } UNION { ?actor rdf:type ?actor_type }
        }
        OPTIONAL {
          { ?target rdfs:subClassOf ?target_type  } UNION { ?target rdf:type ?target_type }
        }
        ?actor semsci:isLocatedIn ?actor_gp_id_location .
        ?actor_gp_id_location rdf:type ?actor_location_type .
        ?target semsci:isLocatedIn ?target_gp_id_location .
        ?target_gp_id_location rdf:type ?target_location_type .
        ?actor semsci:hasFunction ?actor_gp_id_function .
        ?actor_gp_id_function rdf:type ?actor_function_type .
}
```

Meshups: applications that consume LOD

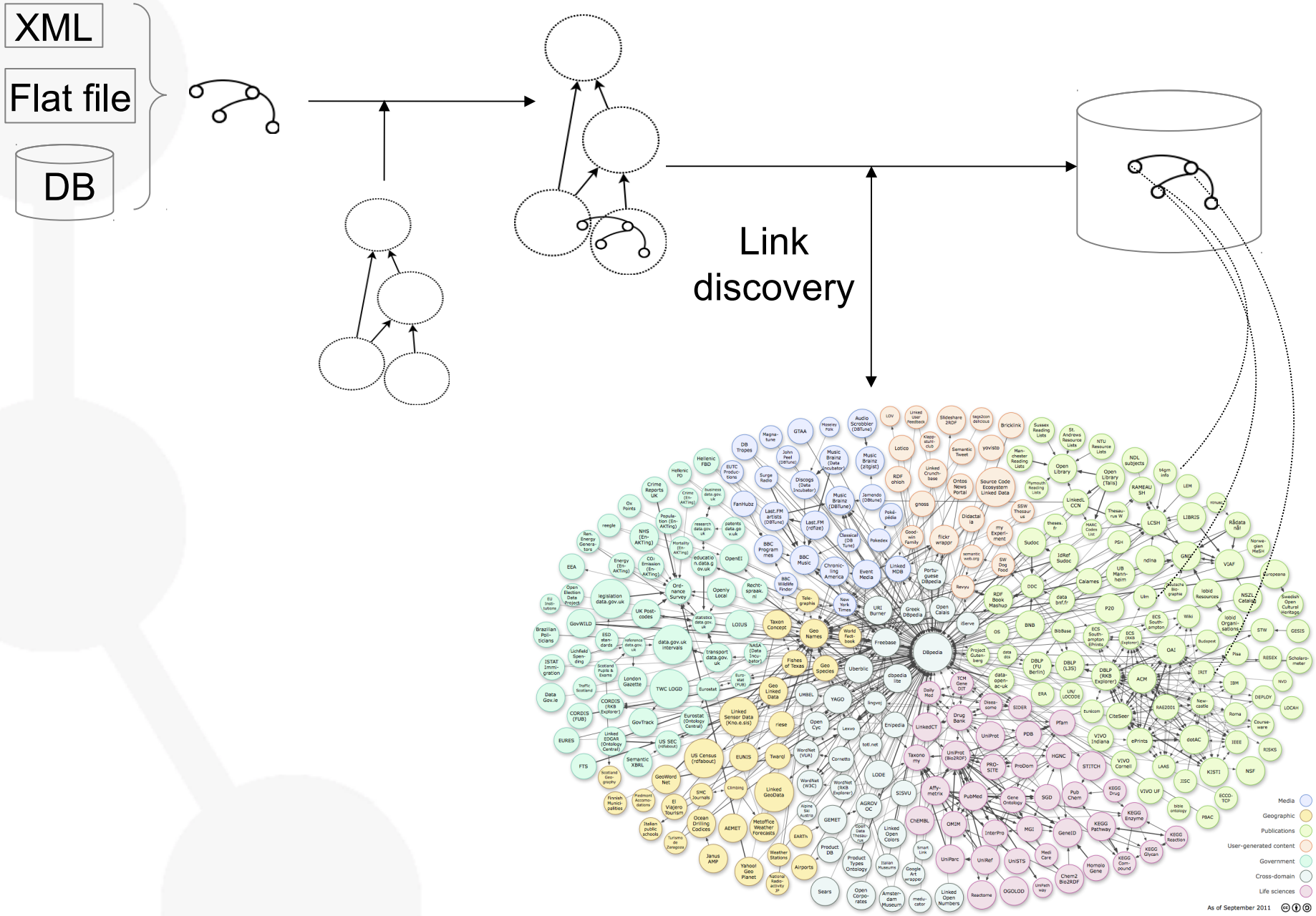
Combining information from different datasets and/or non LOD resources (e.g. Google maps)

e.g. specific visualisations

e.g. “follow your nose” applications

Publishing Linked Data

Publishing Linked Data



Announce your data

Comprehensive Knowledge Archive Network (http://ckan.org/)

Semantic Web index (http://sindice.com/)

Why publish our data in the LOD?

Why publish our data in the LOD?

It's the links, stupid

Publishing Linked Data

Why publish our data in the LOD?

It's the links, stupid

Only publish our data, reference the rest: don't need to duplicate external DB in ours

Publishing Linked Data

Why publish our data in the LOD?

It's the links, stupid

Only publish our data, reference the rest: don't need to duplicate external DB in ours

External info is updated independently and we get the benefit to our dataset because it's linked to it, without extra effort

Why publish our data in the LOD?

It's the links, stupid (II)

Why publish our data in the LOD?

It's the links, stupid (II)

By using (HTTP) URIs, others can link to us

Publishing Linked Data

Why publish our data in the LOD?

It's the links, stupid (II)

By using (HTTP) URIs, others can link to us

Increasing the potential for our data to be discovered

Why publish our data in the LOD?

It's the semantics, stupid

Publishing Linked Data

Why publish our data in the LOD?

It's the semantics, stupid

The meaning of our data is easily machine processable due to RDF (“instances”) and OWL (“schema”)

Issues with (Life Sciences) Linked Data

Issues with (Life Sciences) Linked Open Data

Provenance (e.g. For Microarray data)

Shared identifiers
http://identifiers.org/
http://sharedname.org

Dataset quality

Ontology modelling

Consensus ontologies

Lack of ontologies

Inference
To generate triples
At query time

Conclusions

Conclusions

Linked Data offers a straight method to publish data semantically in the **current** web:

Key 1: use URIs for each and every data item

Key 2: link data items to internal and **external** data

Key 3: represent data with RDF (and OWL)

Already existing web technology (URI + HTTP) will do the rest smoothly for us

Knowledge discovery

Knowledge exploitation

Conclusions

Linked Data is here to stay

Already used by many, including governments, BBC, ...

A first *usable* version of the Semantic Web with great potential

Still issues to be solved in the Life Sciences Linked Data

More information

Semantic Web Health Care and Life Sciences (HCLS) Interest Group at W3C: http://www.w3.org/blog/hcls

LD Best practices

A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In Linked Data on the Web Workshop (LDOW2010) at WWW'2010, 2010.

http://patterns.dataincubator.org/book/

Ontology Engineering Group (oeg-upm.net)

- Gov. Information on LOD (GeoLinkedData, Aemet, ...)
- OGOLOD
- Linked Data tools (ODEmapster, ...)

Acknowledgements

I'm funded by the Marie Curie Cofund programme (FP7)

I unashamedly recycled stuff from presentations by Marc-Alexandre Nolin and Eric Prud'hommeaux

I'm learning a lot at the HCLS IG W3C

NTNU provided the travel/accomodation due to Martin Kuiper's invitation