BioPAX Models and Pathways as Linked Open Data

Michel Dumontier

Department of Biology, School of Computer Science, Institute of Biochemistry
Ottawa Institute for Systems Biology
Ottawa-Carleton Institute for Biomedical Engineering

Carleton University
Ottawa, Canada

BioPAX

- BioPAX is a standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data.
- Terminology is formalized as an OWL ontology
- Data instantiates the ontology and is validated via software
- 30+ resources available in BioPax format (pathguide)

BioPAX – external references

- A major feature of BioPAX data is the ability to add references that denote identity (UnificationXref), related/pertinent (RelationshipXref) or publication (PublicationXref)
- Each Xref specifies an identifier and the database that it stems from

Duplicity in database terminology

PathwayCommons (7 sources)

ARACYC, BRENDA, CABRI, CAS, CHEMICALABSTRACTS, ChEBI, CPATH, CYGD, DDBJ/EMBL/GENBANK, ECOCYC, EMBL, ENSEMBL, ENSEMBLGENOMES, **ENTREZ, ENTREZ_GENE, ENTREZGENE/LOCUSLINK,** ENZYMECONSORTIUM, EVIDENCE CODES ONTOLOGY, GENBANK, GENBANK_NUCL_GI, GENBANK_PROTEIN_GI, **GENE_ONTOLOGY**, GENE_SYMBOL, GRID, HPRD, HUMANCYC, INTACT, **COMPOUND, KEGG-LEGACY, KEGG,** IPI, INTERPRO, IOB, KNAPSACK, METACYC, MINT, NCBI TAXONOMY, NCBI_TAXONOMY, NCI, NEWT, PDB, PDBEPRIDE, PSI-MI, **PSI-MOD**, PUBCHEM, RCSB PDB, **REACTOME, REACTOME DATABASE ID,** REF_SEQ, RESID, SGD, TAXON, TAXONOMY, UMBBD-COMPOUNDS, UNIPARC, UNIPROT, WORMBASE, WWPDB, WIKIPEDIA

Biomodels (1 source)

BioModels Database, Brenda Tissue Ontology, Cell Cycle Ontology, Cell Type Ontology, ChEBI, DOI, Ensembl, Enzyme Nomenclature, FMA, **Gene Ontology**, Human Disease Ontology, ICD, IntAct, InterPro, KEGG Compound, KEGG Pathway, KEGG Reaction, NARCIS, OMIM, PATO, PIRSF, **Protein Modification Ontology**, PubMed, Reactome, Taxonomy, UniProt

BioPAX Xrefs

Pathwaycommons (level 2; download)

```
<bp:unificationXref rdf:ID="CPATH-LOCAL-653">
  <bp:ID rdf:datatype="xsd:string">9606</bp:ID>
  <bp:DB rdf:datatype="xsd:string">NCBI_TAXONOMY</bp:DB>
</bp:unificationXref>
```

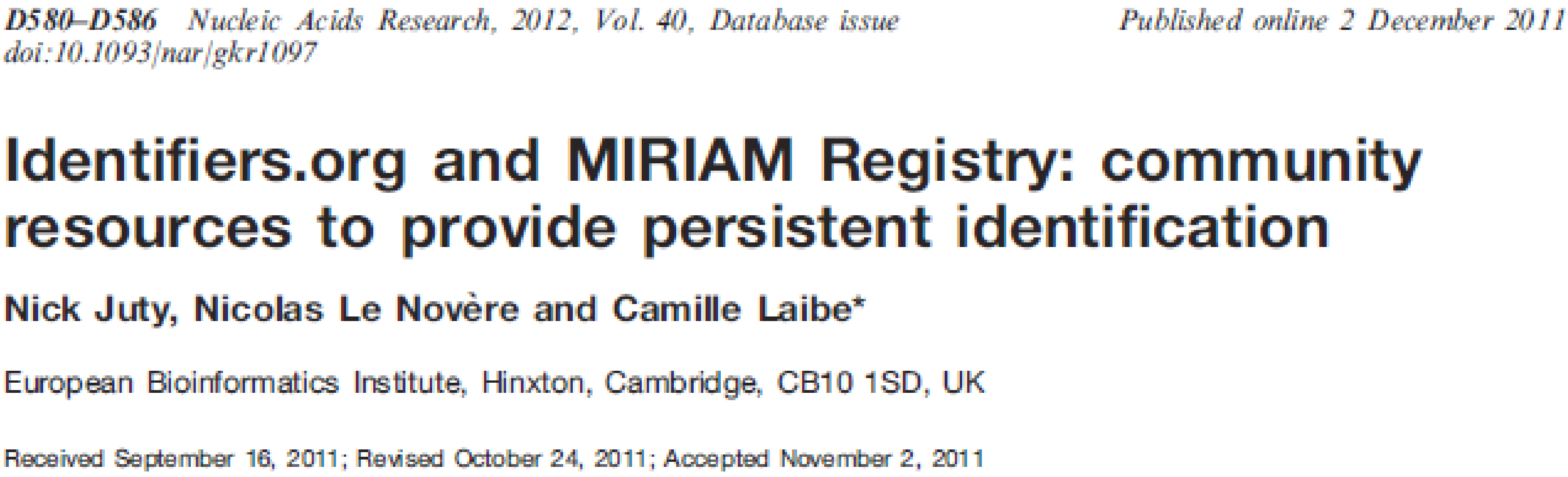
Pathwaycommons (level 3; web service)

```
<bp:UnificationXref rdf:about="urn:biopax:UnificationXref:REACTOME+DATABASE+ID_109276">
  <bp:id rdf:datatype = "http://www.w3.org/2001/XMLSchema#string">109276</bp:id>
  <bp:db rdf:datatype = "http://www.w3.org/2001/XMLSchema#string">Reactome Database ID</bp:db>
</bp:UnificationXref>
```

Biomodels (level 3)

```
<bp:UnificationXref rdf:about="http://identifiers.org/obo.go/GO:0004889">
  <bp:id rdf:datatype = "http://www.w3.org/2001/XMLSchema#string">GO:0004889</bp:id>
  <bp:db rdf:datatype = "http://www.w3.org/2001/XMLSchema#string">Gene Ontology</bp:db>
</bp:UnificationXref>
```

identifiers.org offers a way forward



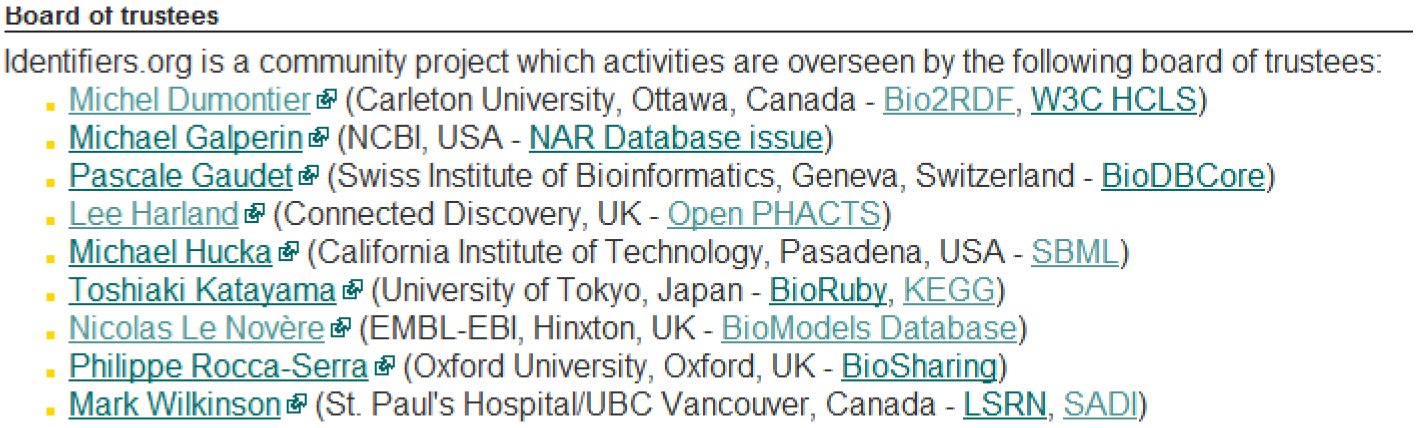
D580–D586 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 2 December 2011*
doi:10.1093/nar/gkr1097

Identifiers.org and MIRIAM Registry: community resources to provide persistent identification

Nick Juty, Nicolas Le Novère and Camille Laibe*

European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

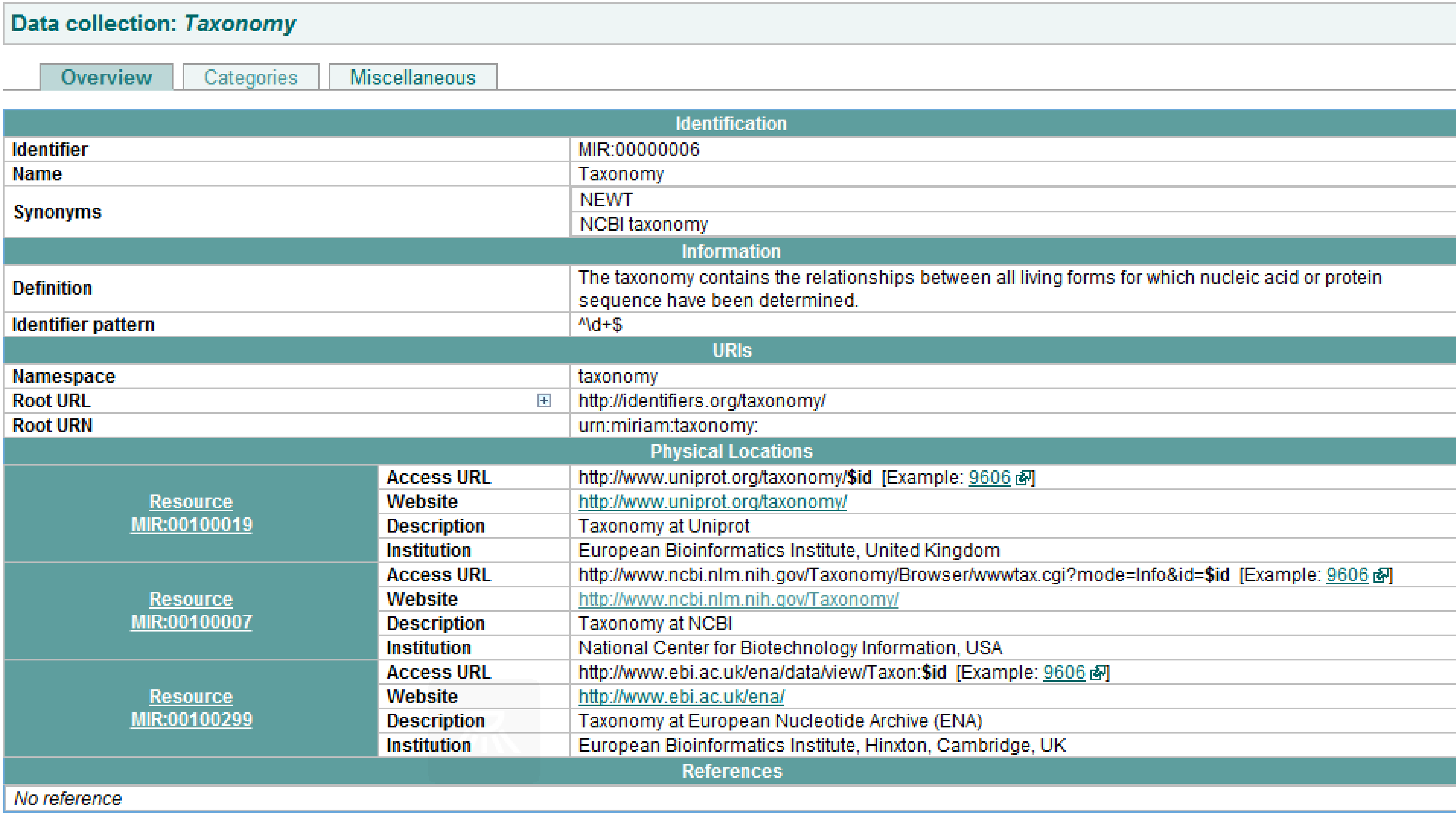
Received September 16, 2011; Revised October 24, 2011; Accepted November 2, 2011



Board of trustees

Identifiers.org is a community project which activities are overseen by the following board of trustees:

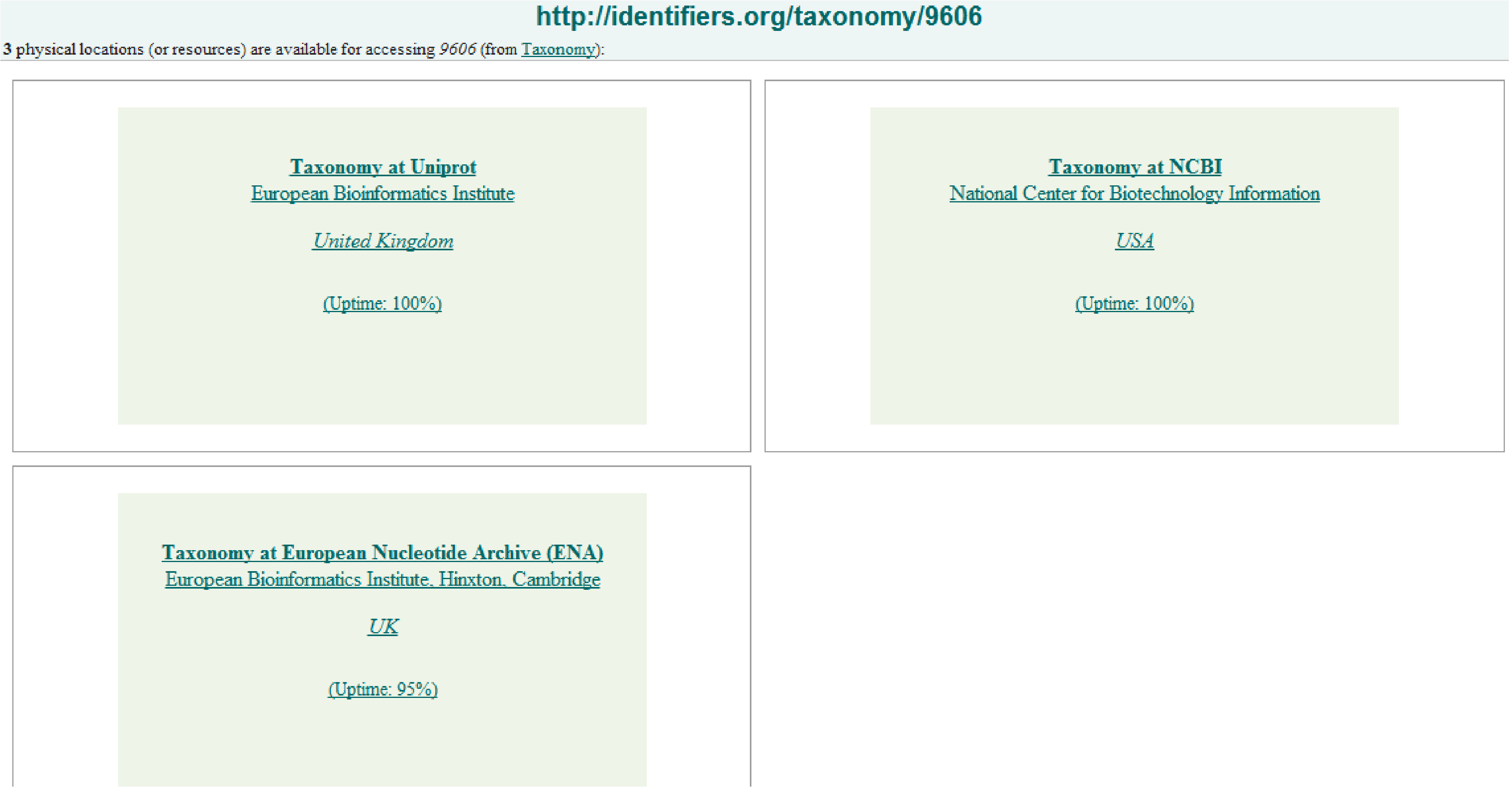
- Michel Dumontier (Carleton University, Ottawa, Canada - Bio2RDF, W3C HCLS)
- Michael Galperin (NCBI, USA - NAR Database issue)
- Pascale Gaudet (Swiss Institute of Bioinformatics, Geneva, Switzerland - BioDBCore)
- Lee Harland (Connected Discovery, UK - Open PHACTS)
- Michael Hucka (California Institute of Technology, Pasadena, USA - SBML)
- Toshiaki Katayama (University of Tokyo, Japan - BioRuby, KEGG)
- Nicolas Le Novère (EMBL-EBI, Hinxton, UK - BioModels Database)
- Philippe Rocca-Serra (Oxford University, Oxford, UK - BioSharing)
- Mark Wilkinson (St. Paul's Hospital/UBC Vancouver, Canada - LSRN, SADI)



Data collection: *Taxonomy*

Identification		
Identifier		MIR:00000006
Name		Taxonomy
Synonyms		NEWT NCBI taxonomy
Information		
Definition		The taxonomy contains the relationships between all living forms for which nucleic acid or protein sequence have been determined.
Identifier pattern		^\d+$
URIs		
Namespace		taxonomy
Root URL	⊞	http://identifiers.org/taxonomy/
Root URN		urn:miriam:taxonomy:
Physical Locations		
Resource MIR:00100019	**Access URL**	http://www.uniprot.org/taxonomy/**$id** [Example: 9606]
	Website	http://www.uniprot.org/taxonomy/
	Description	Taxonomy at Uniprot
	Institution	European Bioinformatics Institute, United Kingdom
Resource MIR:00100007	**Access URL**	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=**$id** [Example: 9606]
	Website	http://www.ncbi.nlm.nih.gov/Taxonomy/
	Description	Taxonomy at NCBI
	Institution	National Center for Biotechnology Information, USA
Resource MIR:00100299	**Access URL**	http://www.ebi.ac.uk/ena/data/view/Taxon:**$id** [Example: 9606]
	Website	http://www.ebi.ac.uk/ena/
	Description	Taxonomy at European Nucleotide Archive (ENA)
	Institution	European Bioinformatics Institute, Hinxton, Cambridge, UK
References		
No reference		

http://identifiers.org/taxonomy/$id



http://identifiers.org/taxonomy/9606

3 physical locations (or resources) are available for accessing *9606* (from Taxonomy):

Taxonomy at Uniprot
European Bioinformatics Institute

United Kingdom

(Uptime: 100%)

Taxonomy at NCBI
National Center for Biotechnology Information

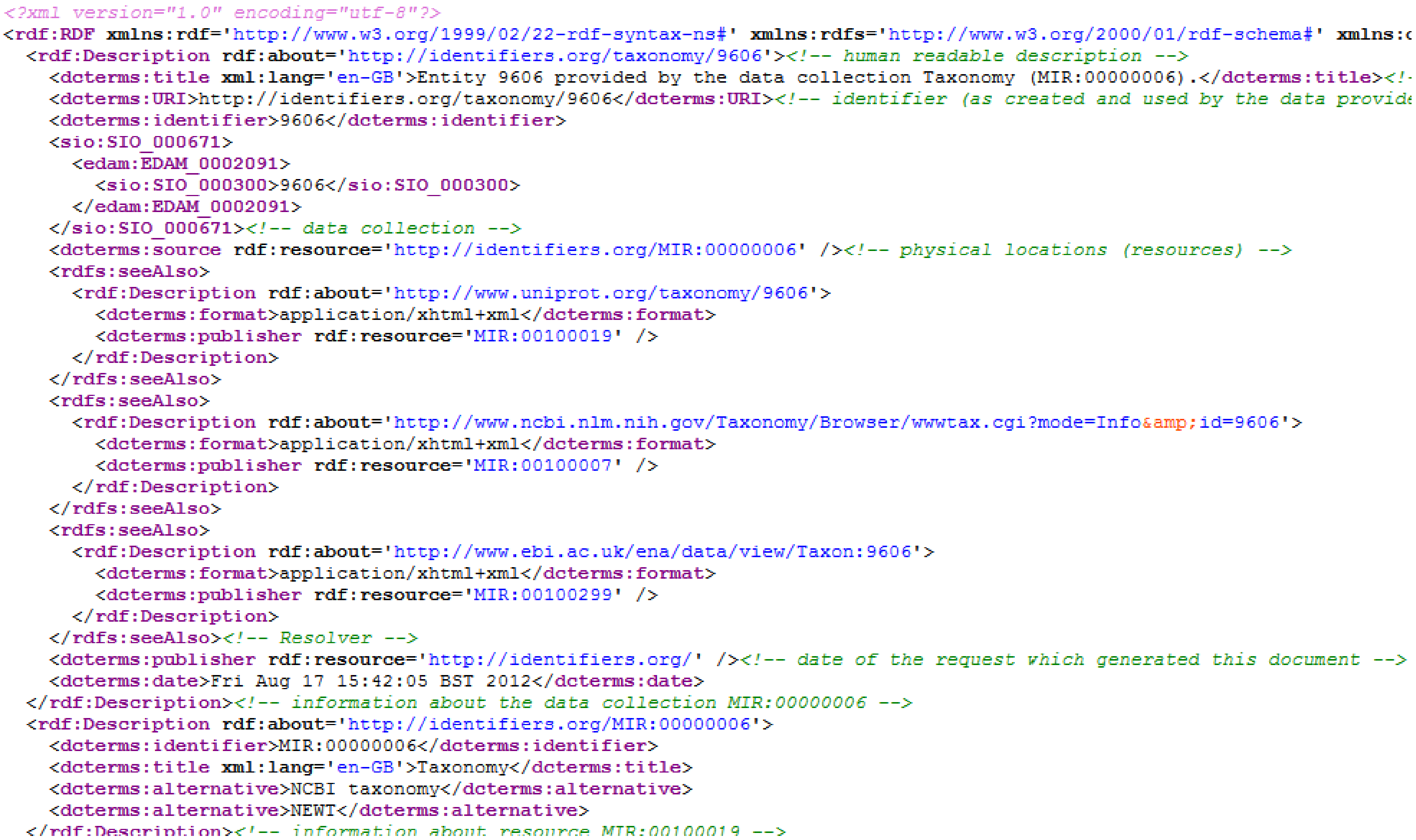
USA

(Uptime: 100%)

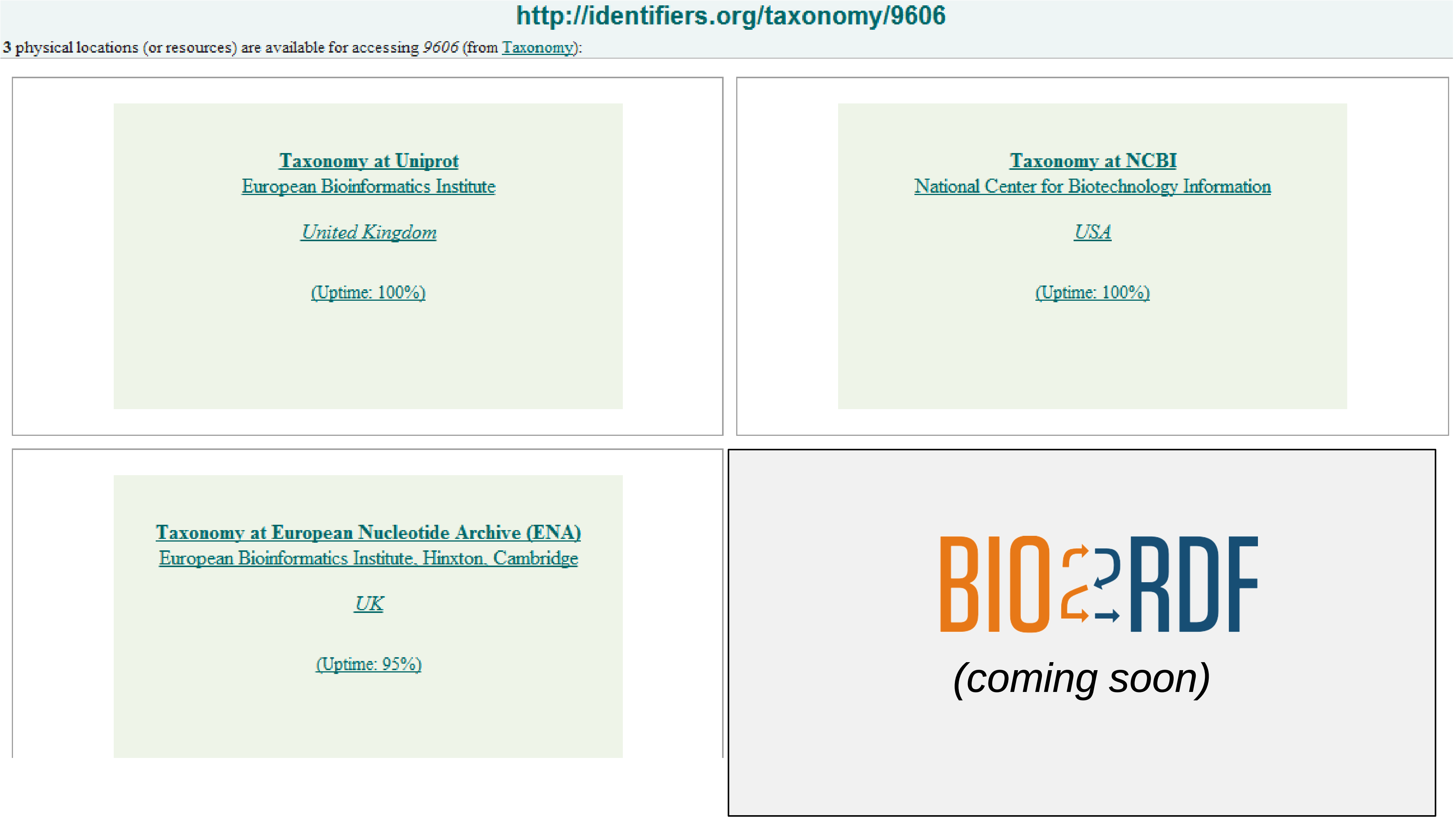
Taxonomy at European Nucleotide Archive (ENA)
European Bioinformatics Institute, Hinxton, Cambridge

UK

(Uptime: 95%)



```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#' xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#' xmlns:c
  <rdf:Description rdf:about='http://identifiers.org/taxonomy/9606'><!-- human readable description -->
    <dcterms:title xml:lang='en-GB'>Entity 9606 provided by the data collection Taxonomy (MIR:00000006).</dcterms:title><!-
    <dcterms:URI>http://identifiers.org/taxonomy/9606</dcterms:URI><!-- identifier (as created and used by the data provide
    <dcterms:identifier>9606</dcterms:identifier>
    <sio:SIO_000671>
      <edam:EDAM_0002091>
        <sio:SIO_000300>9606</sio:SIO_000300>
      </edam:EDAM_0002091>
    </sio:SIO_000671><!-- data collection -->
    <dcterms:source rdf:resource='http://identifiers.org/MIR:00000006' /><!-- physical locations (resources) -->
    <rdfs:seeAlso>
      <rdf:Description rdf:about='http://www.uniprot.org/taxonomy/9606'>
        <dcterms:format>application/xhtml+xml</dcterms:format>
        <dcterms:publisher rdf:resource='MIR:00100019' />
      </rdf:Description>
    </rdfs:seeAlso>
    <rdfs:seeAlso>
      <rdf:Description rdf:about='http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&amp;id=9606'>
        <dcterms:format>application/xhtml+xml</dcterms:format>
        <dcterms:publisher rdf:resource='MIR:00100007' />
      </rdf:Description>
    </rdfs:seeAlso>
    <rdfs:seeAlso>
      <rdf:Description rdf:about='http://www.ebi.ac.uk/ena/data/view/Taxon:9606'>
        <dcterms:format>application/xhtml+xml</dcterms:format>
        <dcterms:publisher rdf:resource='MIR:00100299' />
      </rdf:Description>
    </rdfs:seeAlso><!-- Resolver -->
    <dcterms:publisher rdf:resource='http://identifiers.org/' /><!-- date of the request which generated this document -->
    <dcterms:date>Fri Aug 17 15:42:05 BST 2012</dcterms:date>
  </rdf:Description><!-- information about the data collection MIR:00000006 -->
  <rdf:Description rdf:about='http://identifiers.org/MIR:00000006'>
    <dcterms:identifier>MIR:00000006</dcterms:identifier>
    <dcterms:title xml:lang='en-GB'>Taxonomy</dcterms:title>
    <dcterms:alternative>NCBI taxonomy</dcterms:alternative>
    <dcterms:alternative>NEWT</dcterms:alternative>
  </rdf:Description><!-- information about resource MIR:00100019 -->
```



http://identifiers.org/taxonomy/9606

3 physical locations (or resources) are available for accessing *9606* (from Taxonomy):

Taxonomy at Uniprot
European Bioinformatics Institute

United Kingdom

(Uptime: 100%)

Taxonomy at NCBI
National Center for Biotechnology Information

USA

(Uptime: 100%)

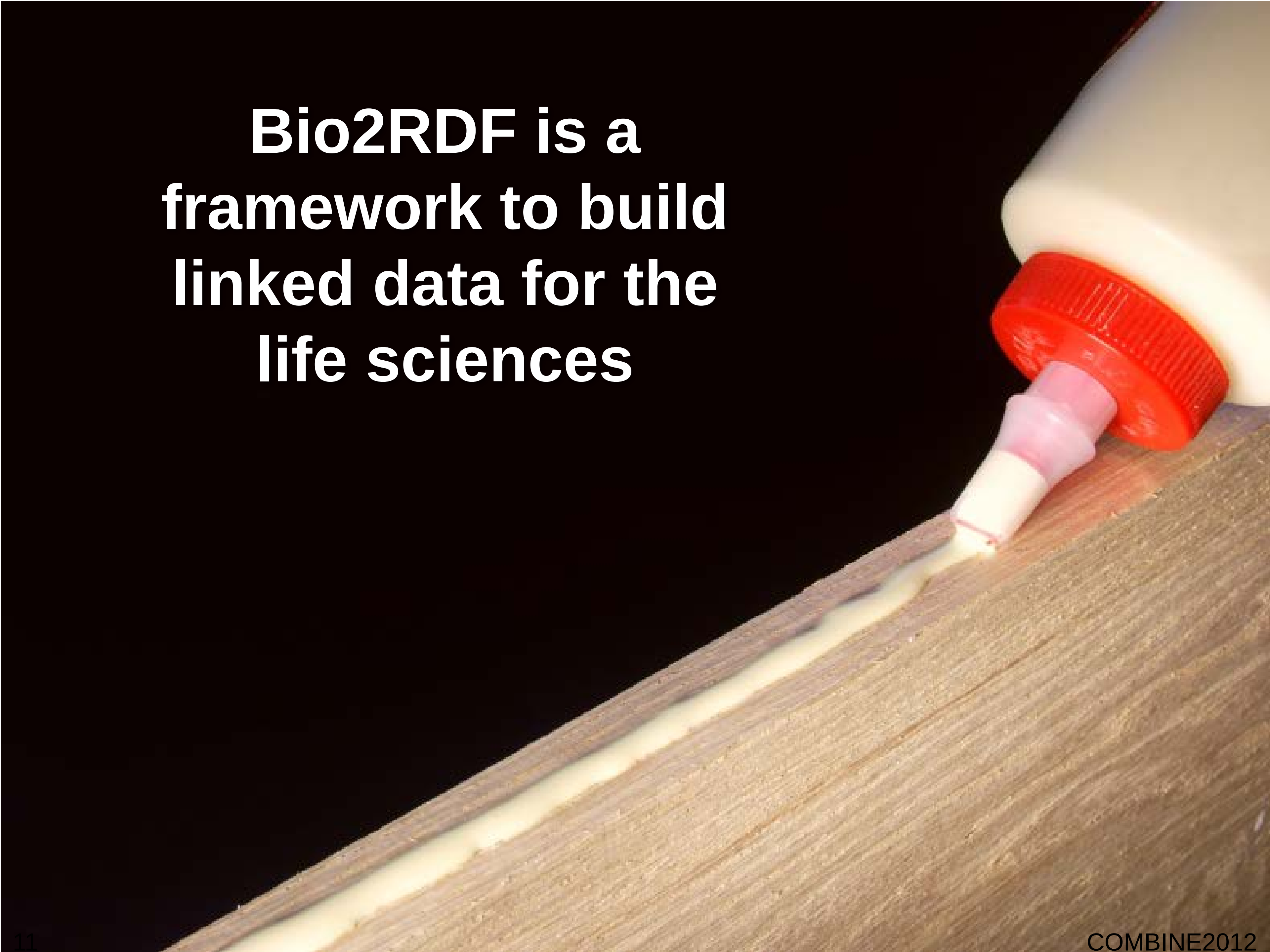
Taxonomy at European Nucleotide Archive (ENA)
European Bioinformatics Institute, Hinxton, Cambridge

UK

(Uptime: 95%)

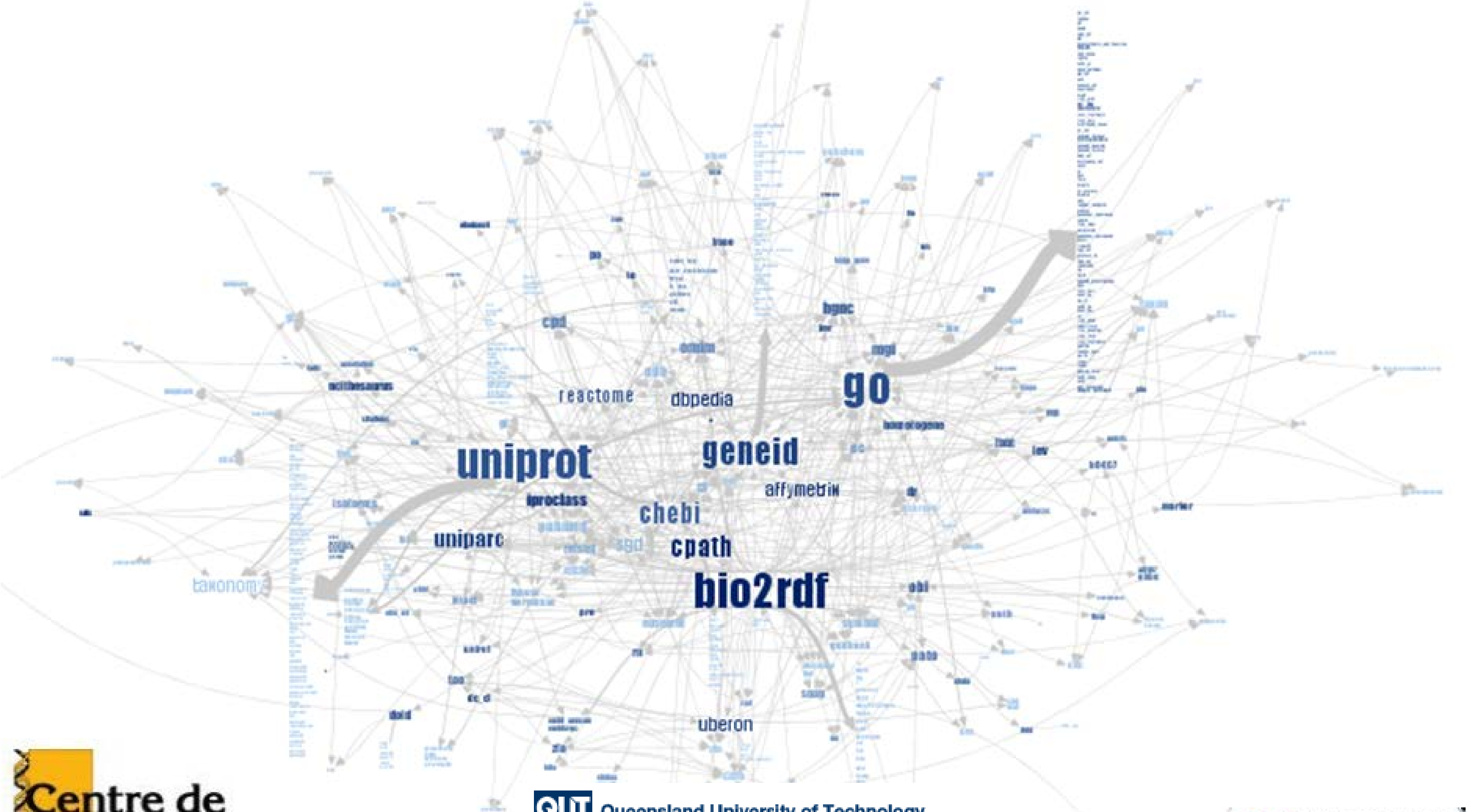
BIO2RDF

(coming soon)



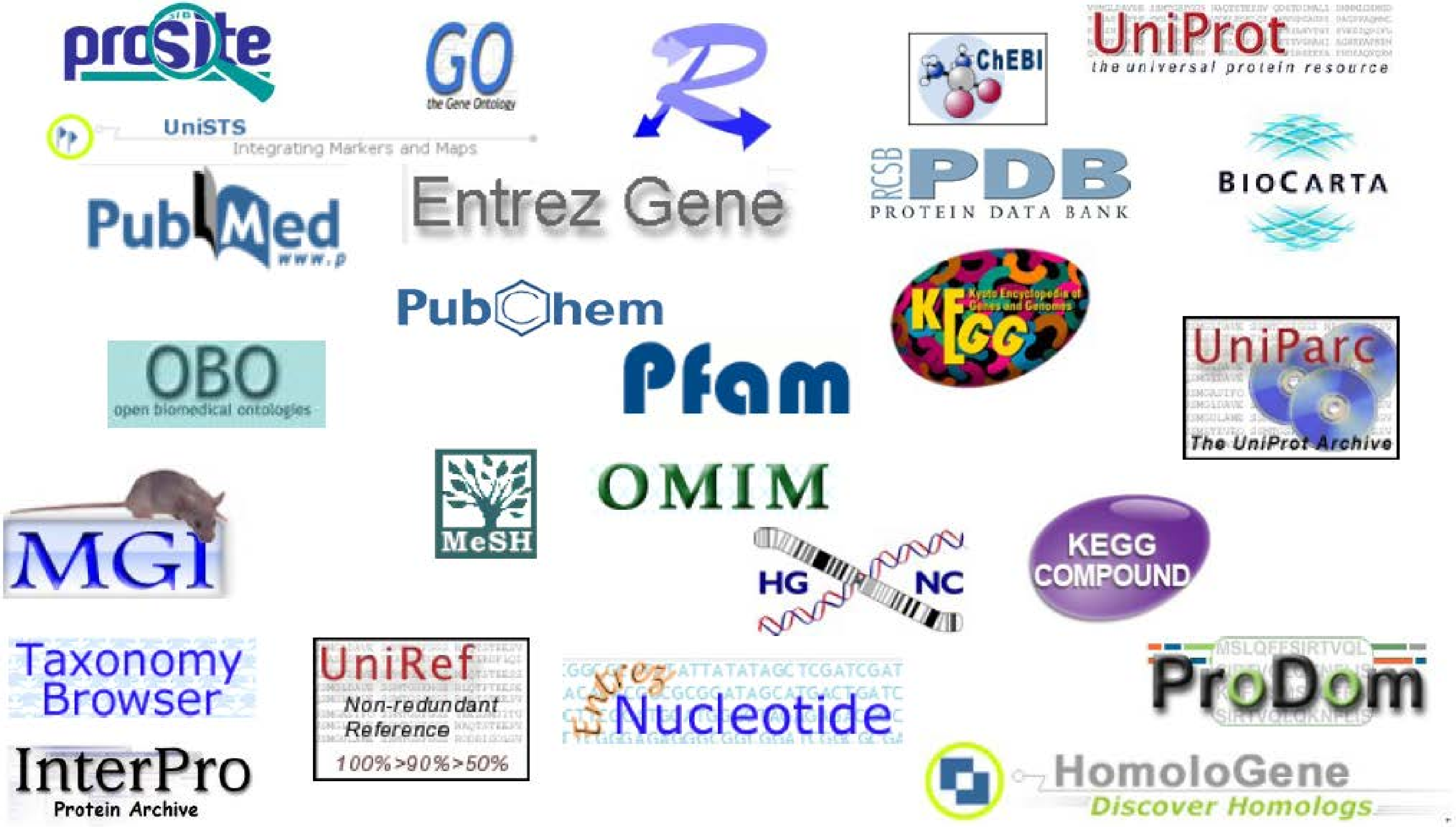
Bio2RDF is a framework to build linked data for the life sciences

Bio2RDF
provides billions of interconnections



Queensland University of Technology
Microsoft QUT eResearch Centre

Bio2RDF covers the major biological databases

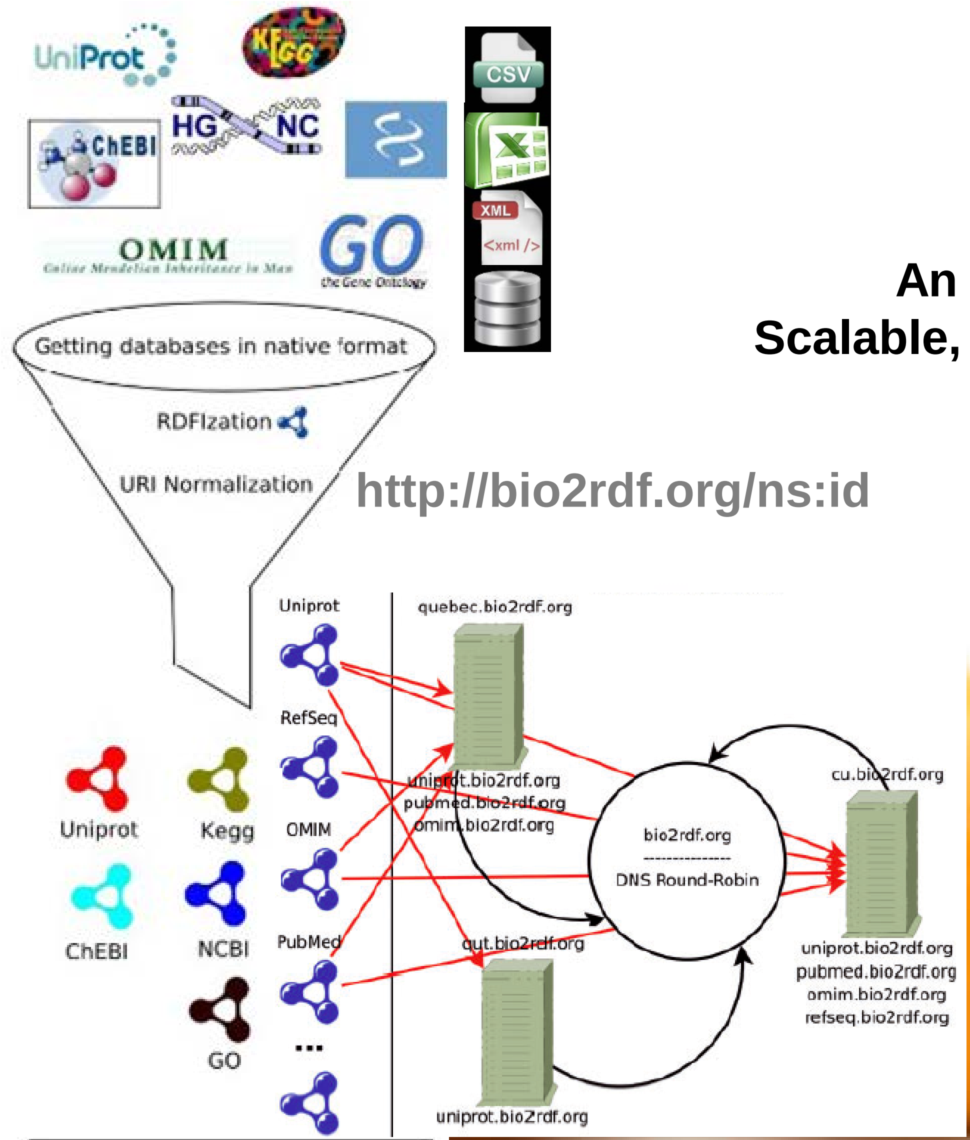


BIO2RDF

linked data for the life sciences

An Open Source Project for the Provision of Scalable, Decentralized Data with Global Mirroring and Customizable Query Resolution

Laval University, Carleton University,
Queensland University of Technology

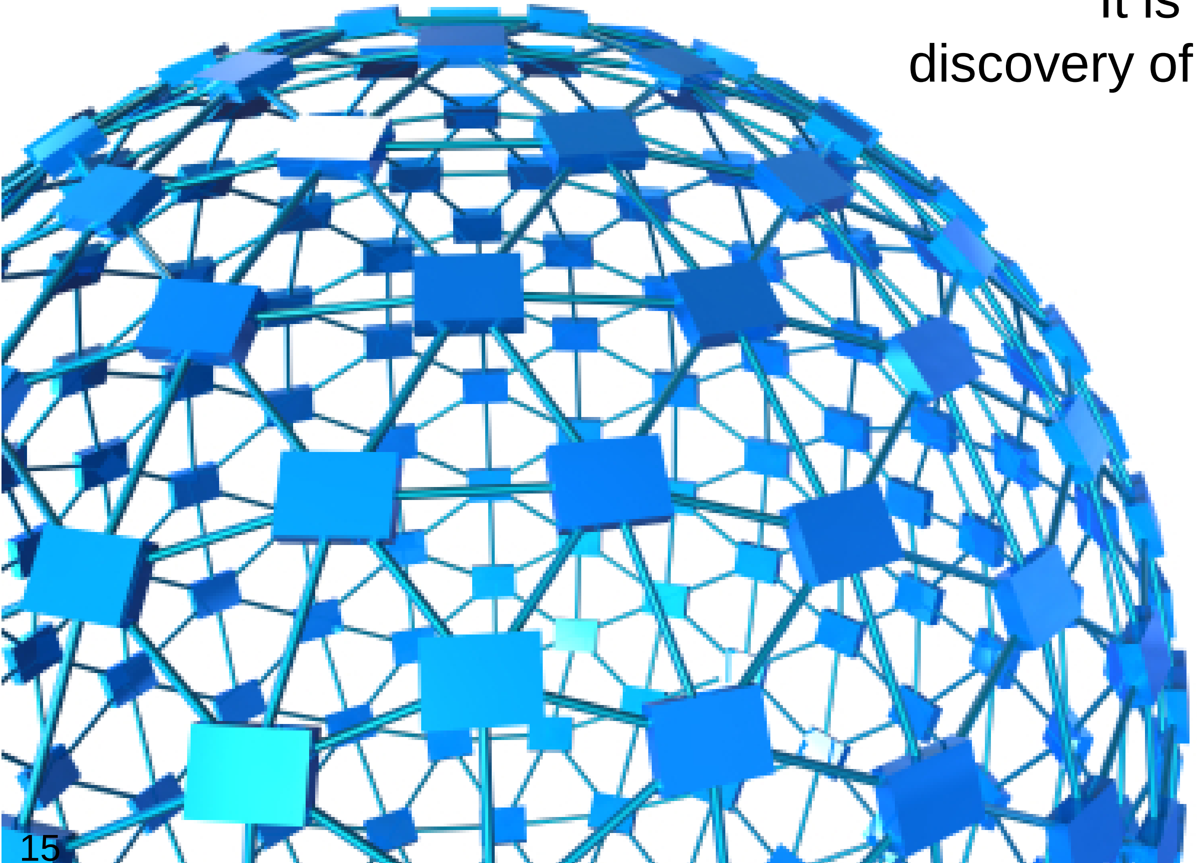


COMBINE2012

The Semantic Web is the new global **web of knowledge**

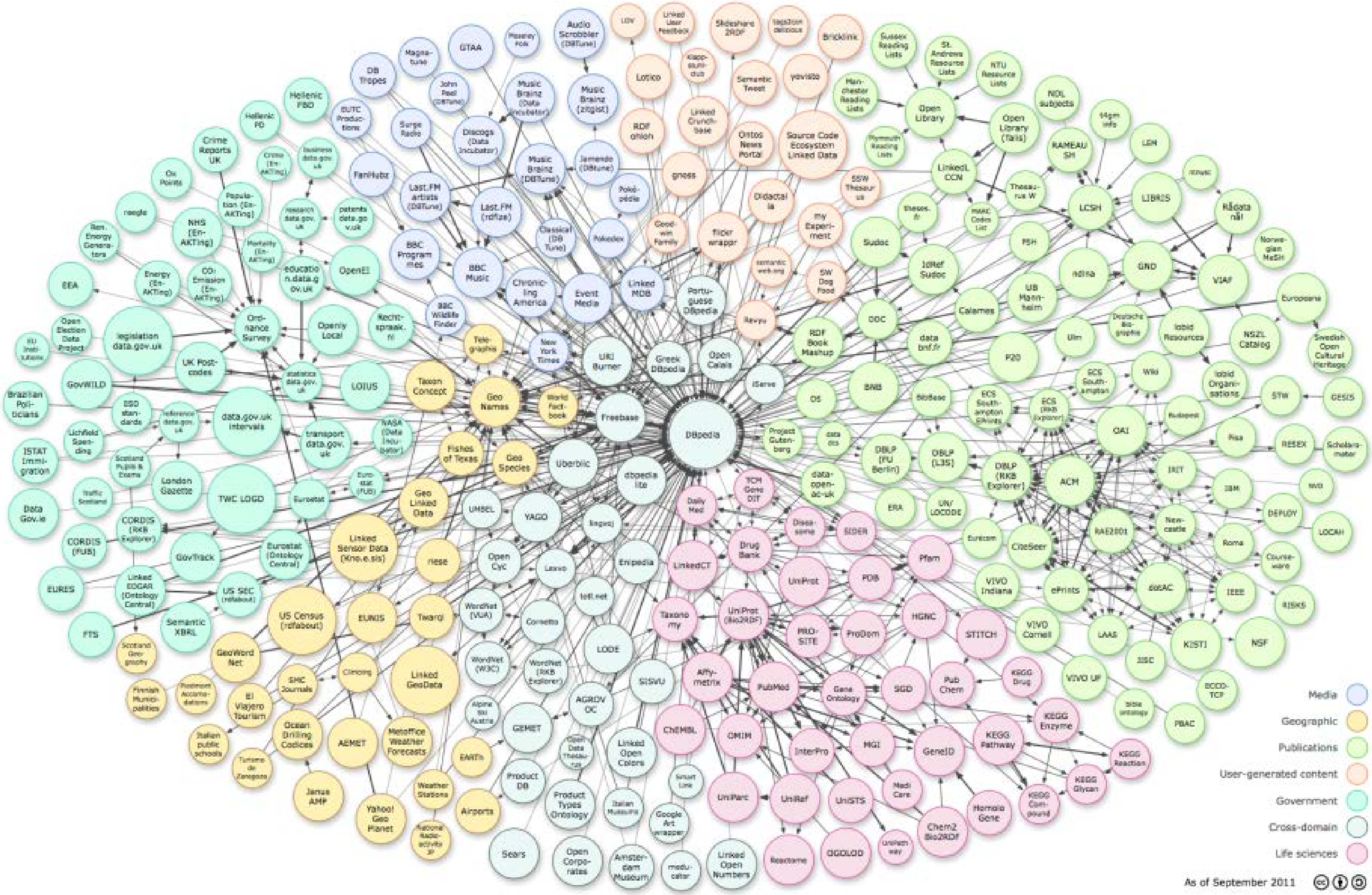
It involves standards for publishing, sharing and querying **facts, expert knowledge and services**

It is a scalable approach to the discovery of *independently formulated* and *distributed* knowledge



W3C® Semantic Web

A continuously growing web of linked data



“Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/”

But BioPAX data isn’t ready for the semantic web

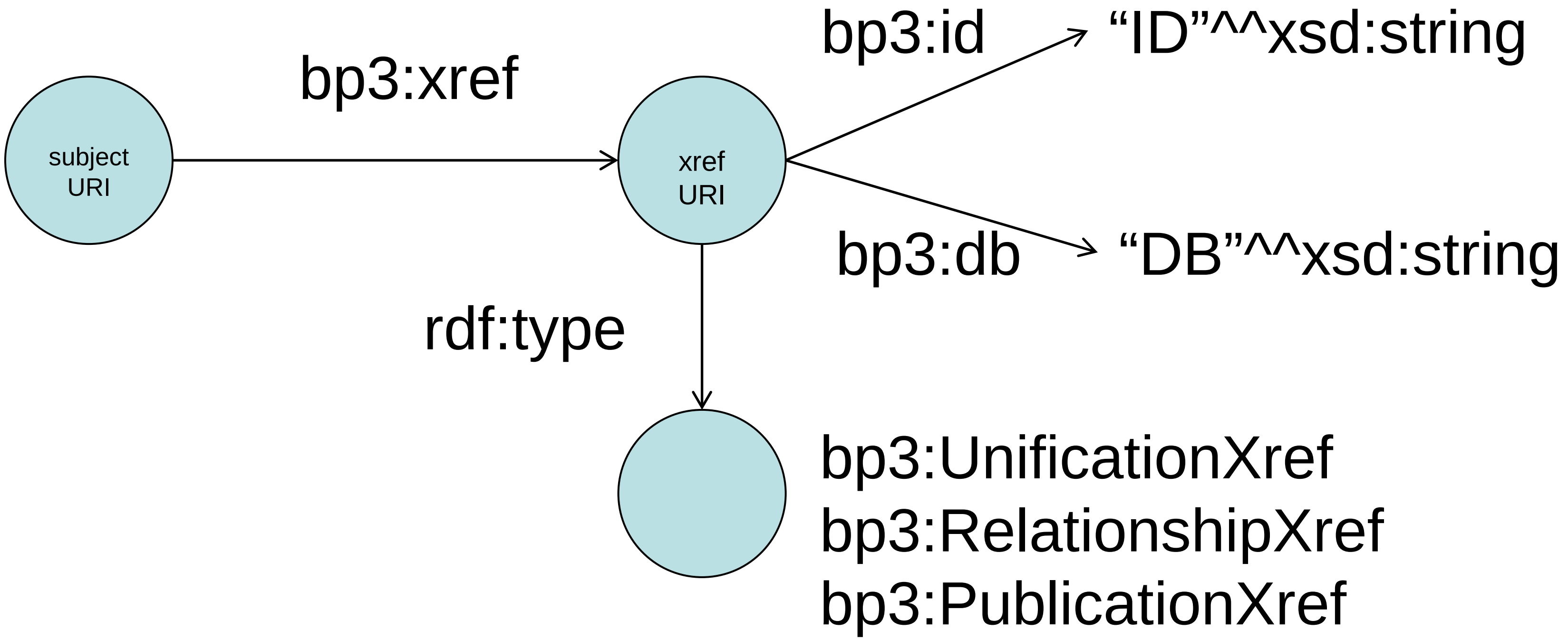
Biomodels

```
<bp:Protein rdf:about="DL">
   <bp:xref rdf:resource="http://identifiers.org/interpro/IPR002394" />
   <bp:xref rdf:resource="http://identifiers.org/obo.go/GO:0005892" />
</bp:Protein>
```

Pathway Commons

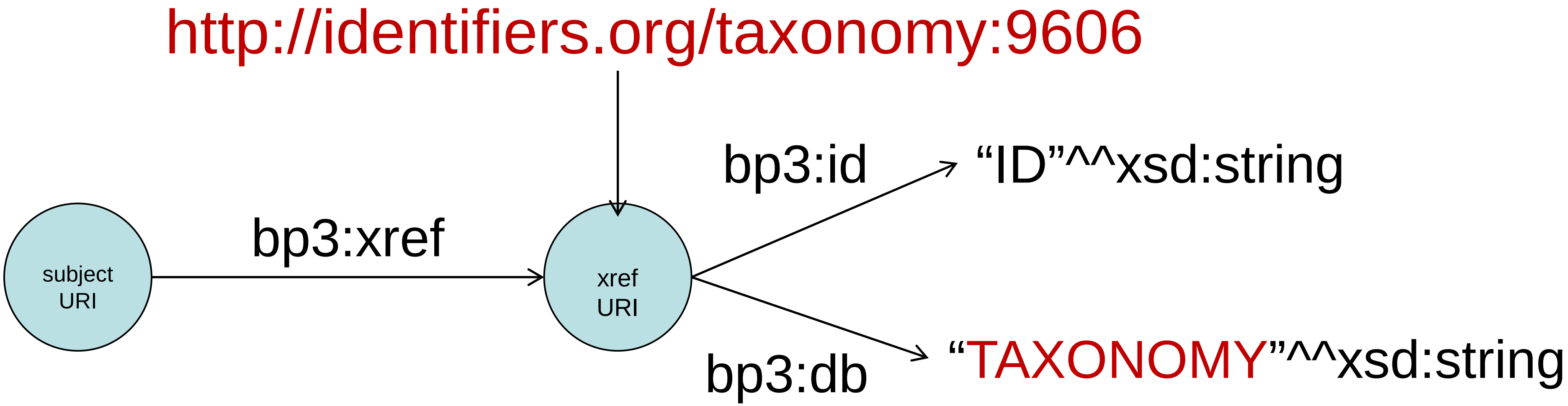
```
 <bp:protein rdf:ID="CPATH-310">
   <bp:ORGANISM>
     <bp:bioSource rdf:ID="CPATH-LOCAL-1446">
       <bp:NAME rdf:datatype="xsd:string">Homo sapiens</bp:NAME>
       <bp:TAXON-XREF>
         <bp:unificationXref rdf:ID="CPATH-LOCAL-1447">
           <bp:DB rdf:datatype="xsd:string">NCBI_TAXONOMY</bp:DB>
           <bp:ID rdf:datatype="xsd:string">9606</bp:ID>
         </bp:unificationXref>
       </bp:TAXON-XREF>
     </bp:bioSource>
   </bp:ORGANISM>
```

BioPAX L3 xrefs



bp3:UnificationXref
bp3:RelationshipXref
bp3:PublicationXref

Use identifiers.org to normalize the syntax of the xref and make it resolvable



http://identifiers.org/taxonomy:9606

subject URI

bp3:xref

xref URI

bp3:id

“ID”^^xsd:string

bp3:db

“TAXONOMY”^^xsd:string

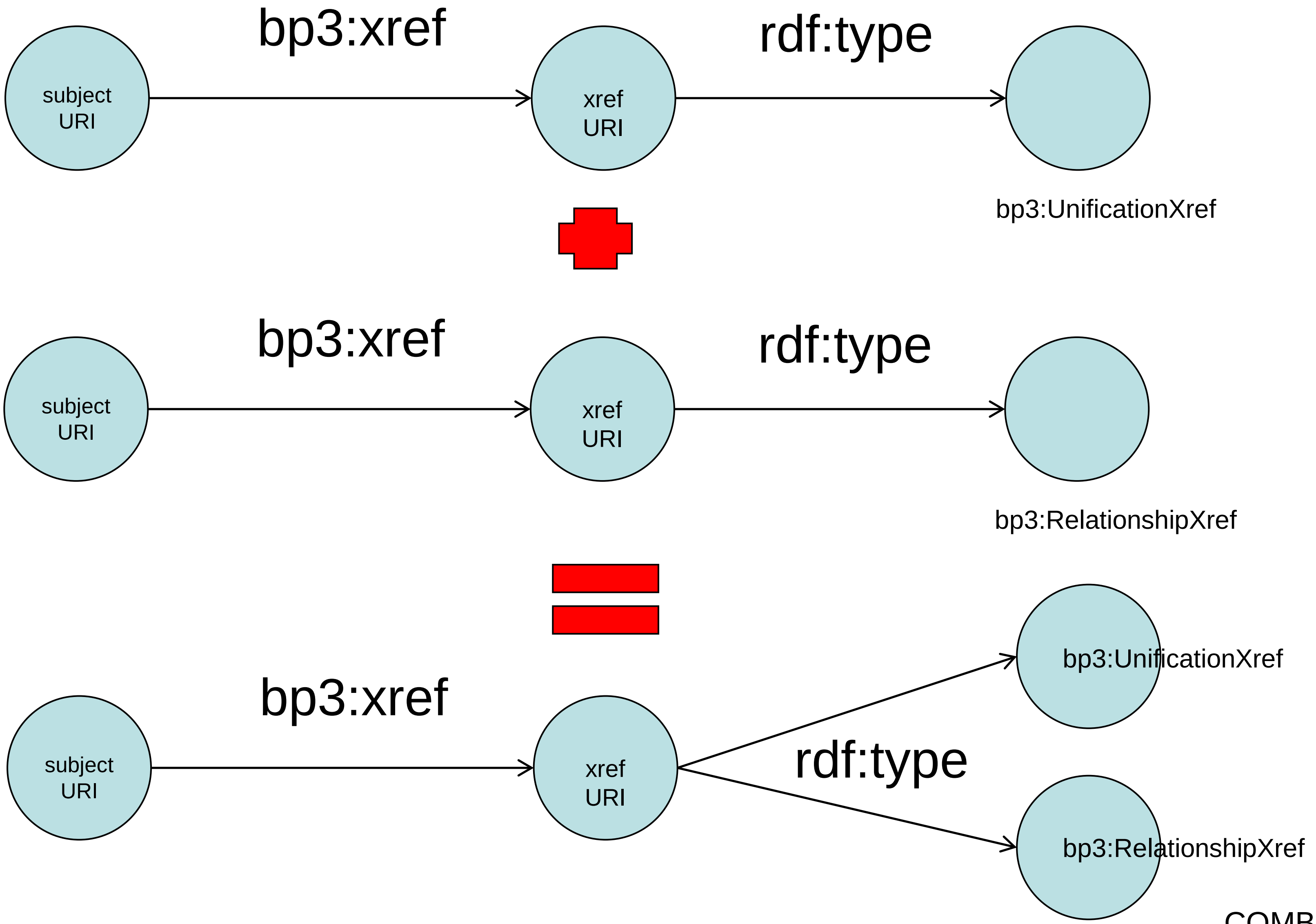
specificity of Xref type gets lost on integration

- Xrefs are typed
  - unificationXref, relationshipXref, publicationXref

```
<bp:UnificationXref rdf:about="http://identifiers.org/obo.go/GO:0004889">
  <bp:id rdf:datatype = "http://www.w3.org/2001/XMLSchema#string">GO:0004889</bp:id>
  <bp:db rdf:datatype = "http://www.w3.org/2001/XMLSchema#string">Gene Ontology</bp:db>
</bp:UnificationXref>
```

- But integration of data would lose the *nature* of relationship

BioPAX L3 xrefs

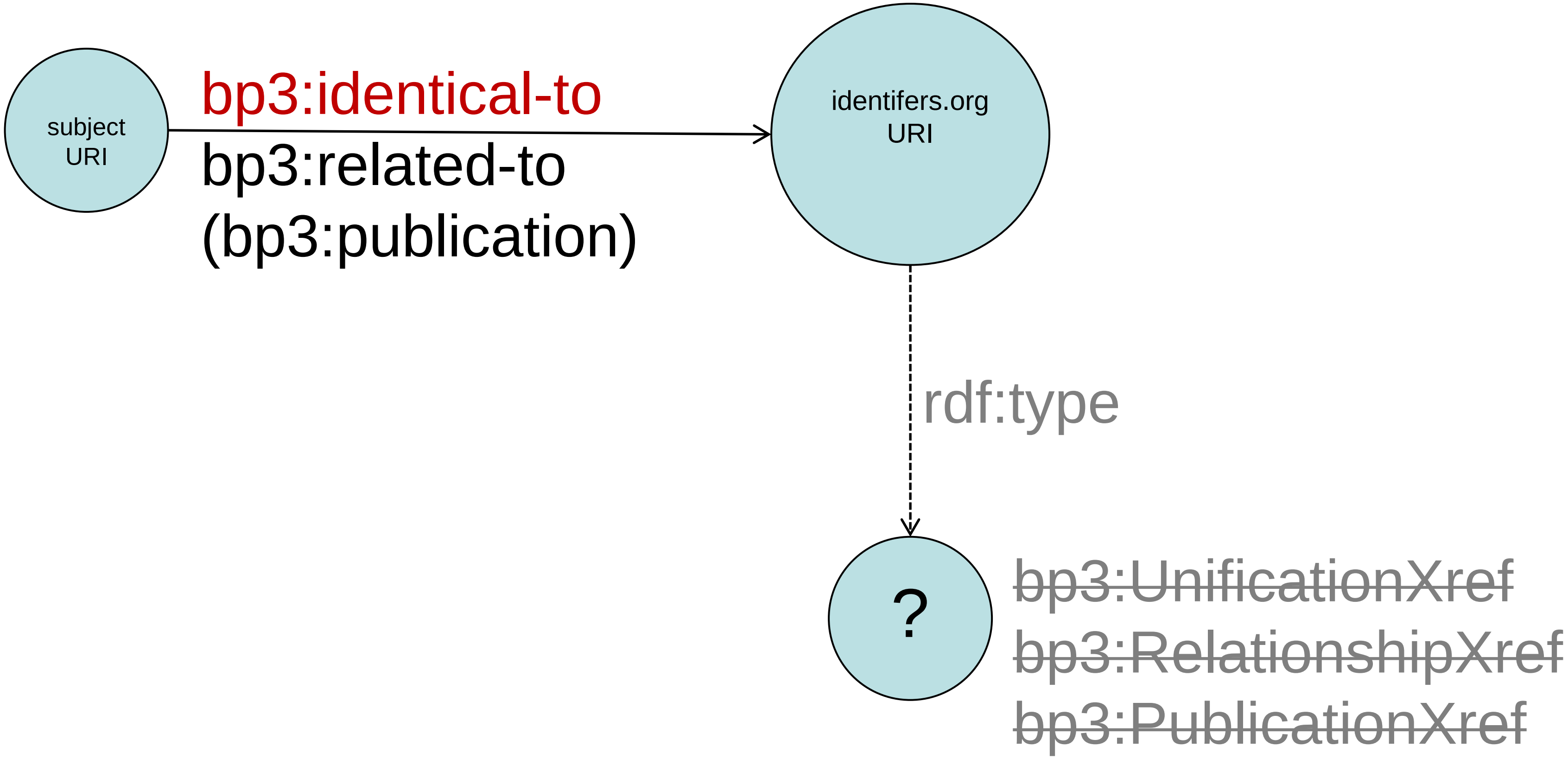


specificity of Xref type gets lost on integration

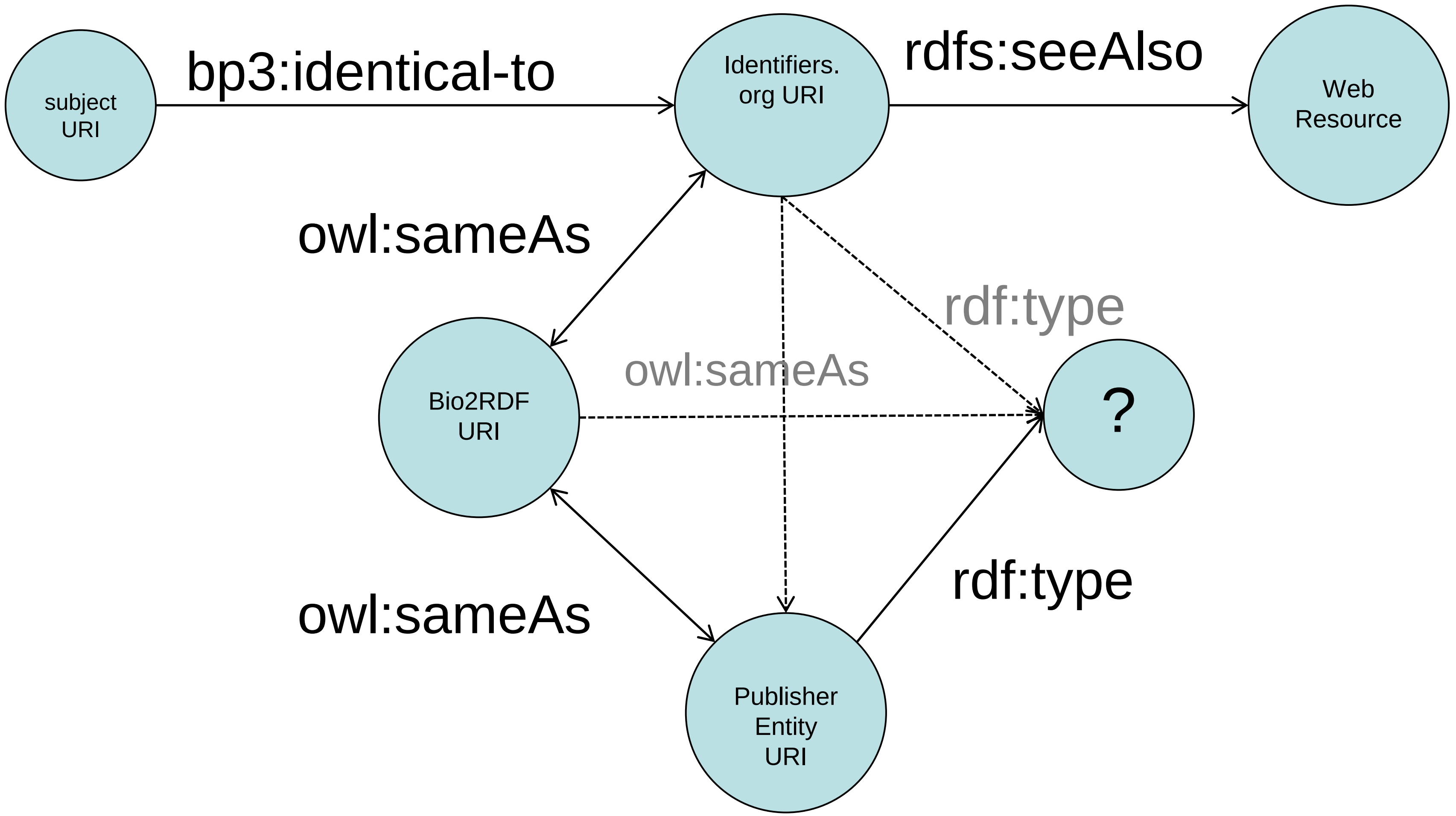
necessary to *reify* the relation

- complex: role-based representation (OBI, SIO)
- simple: use more specific predicates (SIO)

Xrefs – specify the role in the predicate so as to maintain the relationship



Xrefs – Bio2RDF driven integration



Bio2RDF coverage

PathwayCommons

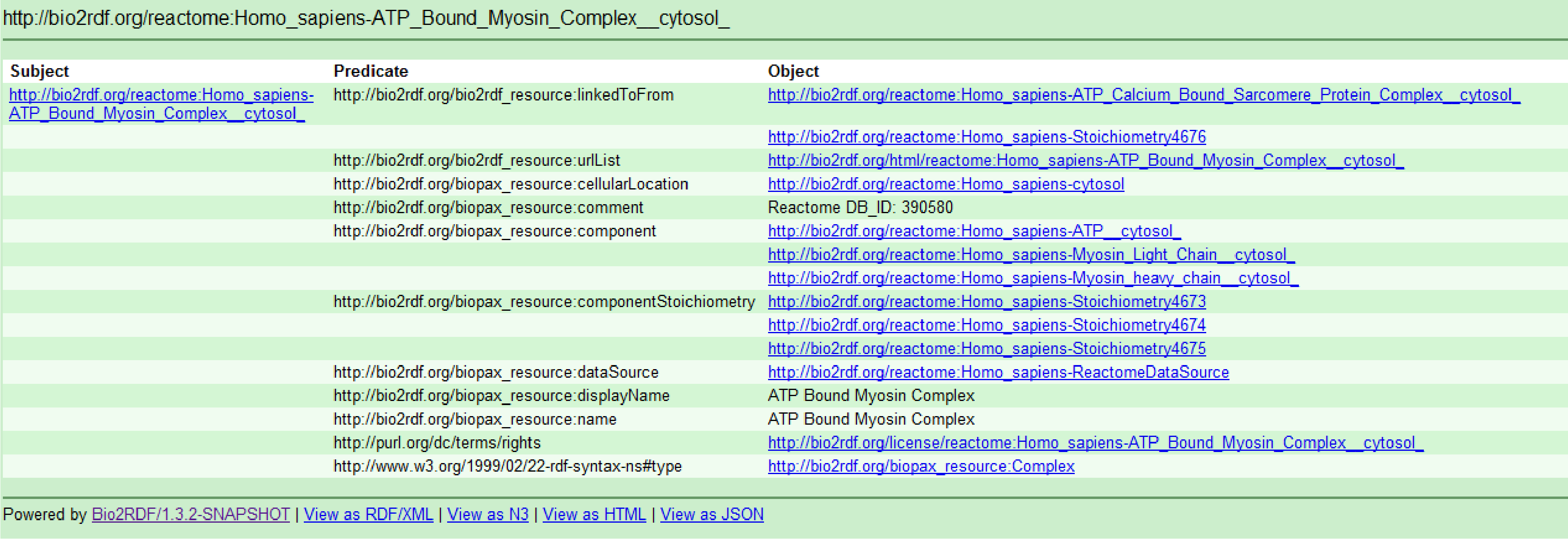
AraCYC, ECOCYC, METACYC, **HUMANCYC,** BRENDA, CABRI, CAS, **ChEBI**, **CPATH**, CYGD, **DDBJ/EMBL/GENBANK**, ENSEMBL, ENSEMBL GENOMES, **NCBI GENE, Enzyme Nomenclature**, **Evidence Code Ontology, Gene Ontology, HGNC Gene Symbol, BioGRID, HPRD, INTACT**, **KEGG, IPI**, **INTERPRO**, IOB, KNAPSACK, **MINT**, **NCBI TAXONOMY,** NCI, NEWT, **PDB**, PRIDE, **PSI-MI, PSI-MOD**, **PUBCHEM**, **REACTOME, RefSeq**, **RESID**, **SGD**, UMBBD-COMPOUNDS, **UNIPARC**, **UNIPROT**, WORMBASE, **WIKIPEDIA**

Biomodels

BioModels Database, Brenda Tissue Ontology, Cell Cycle Ontology, Cell Type Ontology, ChEBI, DOI**,** Ensembl**, Enzyme Nomenclature, FMA, Gene Ontology, Human Disease Ontology,** ICD**, IntAct, InterPro, KEGG Compound, KEGG Pathway, KEGG Reaction,** NARCIS**, OMIM, PATO, PIRSF, Protein Modification Ontology, PubMed, Reactome, Taxonomy, UniProt**

something you can lookup or search for with rich descriptions

Linked Open Data

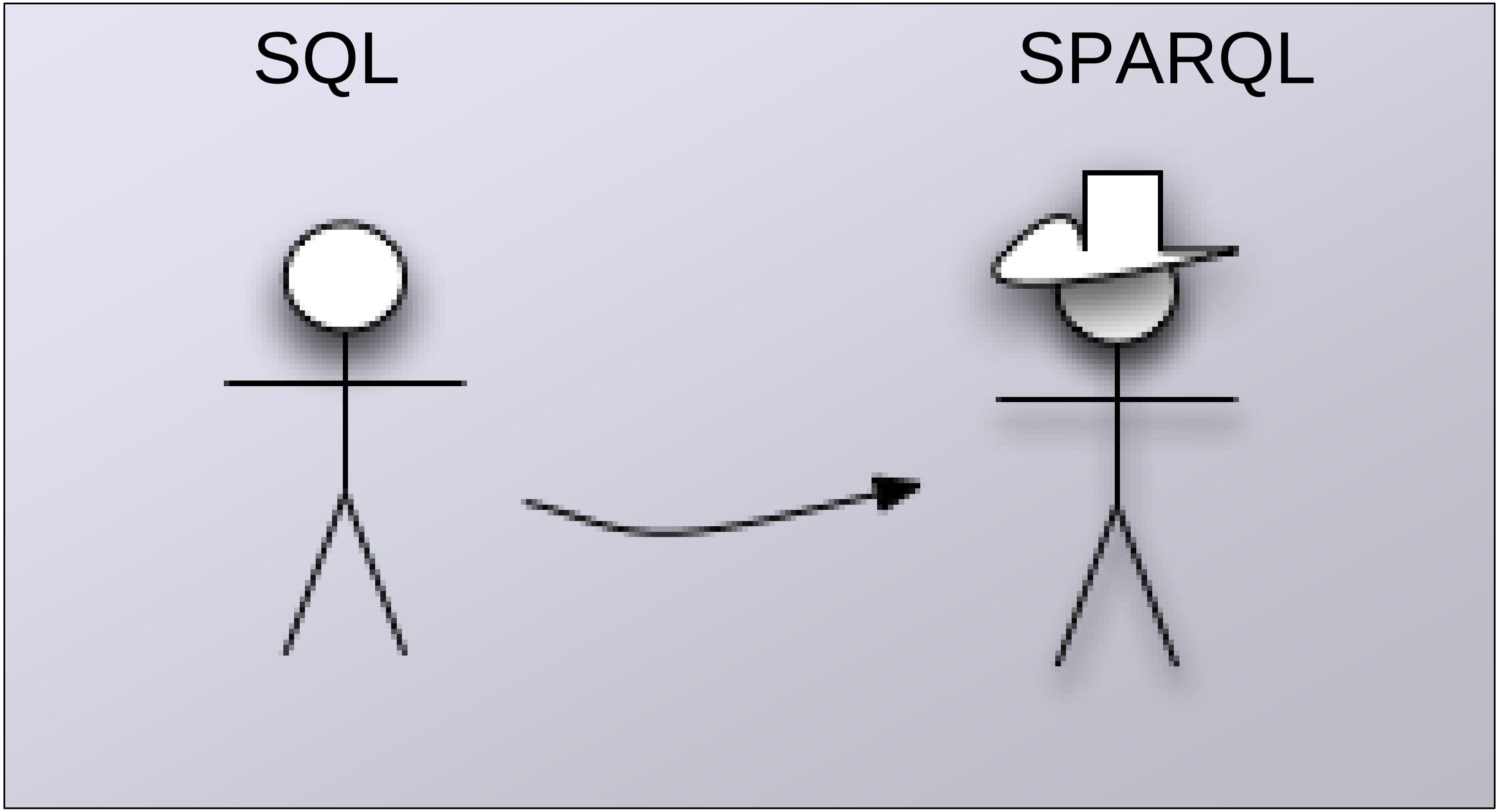


http://bio2rdf.org/reactome:Homo_sapiens-ATP_Bound_Myosin_Complex__cytosol_

Subject	Predicate	Object
http://bio2rdf.org/reactome:Homo_sapiens-ATP_Bound_Myosin_Complex__cytosol_	http://bio2rdf.org/bio2rdf_resource:linkedToFrom	http://bio2rdf.org/reactome:Homo_sapiens-ATP_Calcium_Bound_Sarcomere_Protein_Complex__cytosol_
		http://bio2rdf.org/reactome:Homo_sapiens-Stoichiometry4676
	http://bio2rdf.org/bio2rdf_resource:urlList	http://bio2rdf.org/html/reactome:Homo_sapiens-ATP_Bound_Myosin_Complex__cytosol_
	http://bio2rdf.org/biopax_resource:cellularLocation	http://bio2rdf.org/reactome:Homo_sapiens-cytosol
	http://bio2rdf.org/biopax_resource:comment	Reactome DB_ID: 390580
	http://bio2rdf.org/biopax_resource:component	http://bio2rdf.org/reactome:Homo_sapiens-ATP__cytosol_
		http://bio2rdf.org/reactome:Homo_sapiens-Myosin_Light_Chain__cytosol_
		http://bio2rdf.org/reactome:Homo_sapiens-Myosin_heavy_chain__cytosol_
	http://bio2rdf.org/biopax_resource:componentStoichiometry	http://bio2rdf.org/reactome:Homo_sapiens-Stoichiometry4673
		http://bio2rdf.org/reactome:Homo_sapiens-Stoichiometry4674
		http://bio2rdf.org/reactome:Homo_sapiens-Stoichiometry4675
	http://bio2rdf.org/biopax_resource:dataSource	http://bio2rdf.org/reactome:Homo_sapiens-ReactomeDataSource
	http://bio2rdf.org/biopax_resource:displayName	ATP Bound Myosin Complex
	http://bio2rdf.org/biopax_resource:name	ATP Bound Myosin Complex
	http://purl.org/dc/terms/rights	http://bio2rdf.org/license/reactome:Homo_sapiens-ATP_Bound_Myosin_Complex__cytosol_
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/biopax_resource:Complex

Powered by Bio2RDF/1.3.2-SNAPSHOT | View as RDF/XML | View as N3 | View as HTML | View as JSON

SPARQL is the **new** *cool* kid on the query block



BioPAX SPARQL Endpoints

- Current temporary endpoint:
  - http://bio2rdf.semanticscience.org:8010/sparql
  - http://bio2rdf.semanticscience.org:8010/fct
  - pathwaycommons (l2) + biomodels (l3)

- Francois has collected ~15 BioPAX datasets, we'll load and process (add rdfs:labels)
- I'll be using the pc2 webservices for l3 pathwaycommons
- We're going to work with data providers to generate valid (identifier.org) URIs
- Official Bio2RDF BioPAX endpoint (to be updated)
  - http://biopax.bio2rdf.org/sparql

Summary

- Use identifers.org as a source of external references – minimally for DB field and for the xref URL – optimally for ALL URIs

- Define a more specific predicate to specify “identity” and “related” by some community-drafted guiding criteria

- Bio2RDF can provide integration with external resources that are part of the Bio2RDF network
  - we would like to host BioPAX SPARQL endpoint

special thanks to Bio2RDF team

Francois Belleau (CHUQ)
Marc-Alexandre Nolin (Laval)
Peter Ansell (Queensland)

Alison Callahan (Carleton)
Jose Cruz-Toledo (Carleton)
Dana Klassen (DERI)

Gary Bader (Toronto)

dumontierlab.com

michel_dumontier@carleton.ca

Website: [*http://dumontierlab.com*](http://dumontierlab.com)
Presentations: http://slideshare.com/micheldumontier

MITACS
NSERC CRSNG
Canada Foundation for Innovation
Fondation canadienne pour l'innovation
canarie
Carleton UNIVERSITY
Health Canada
Ontario