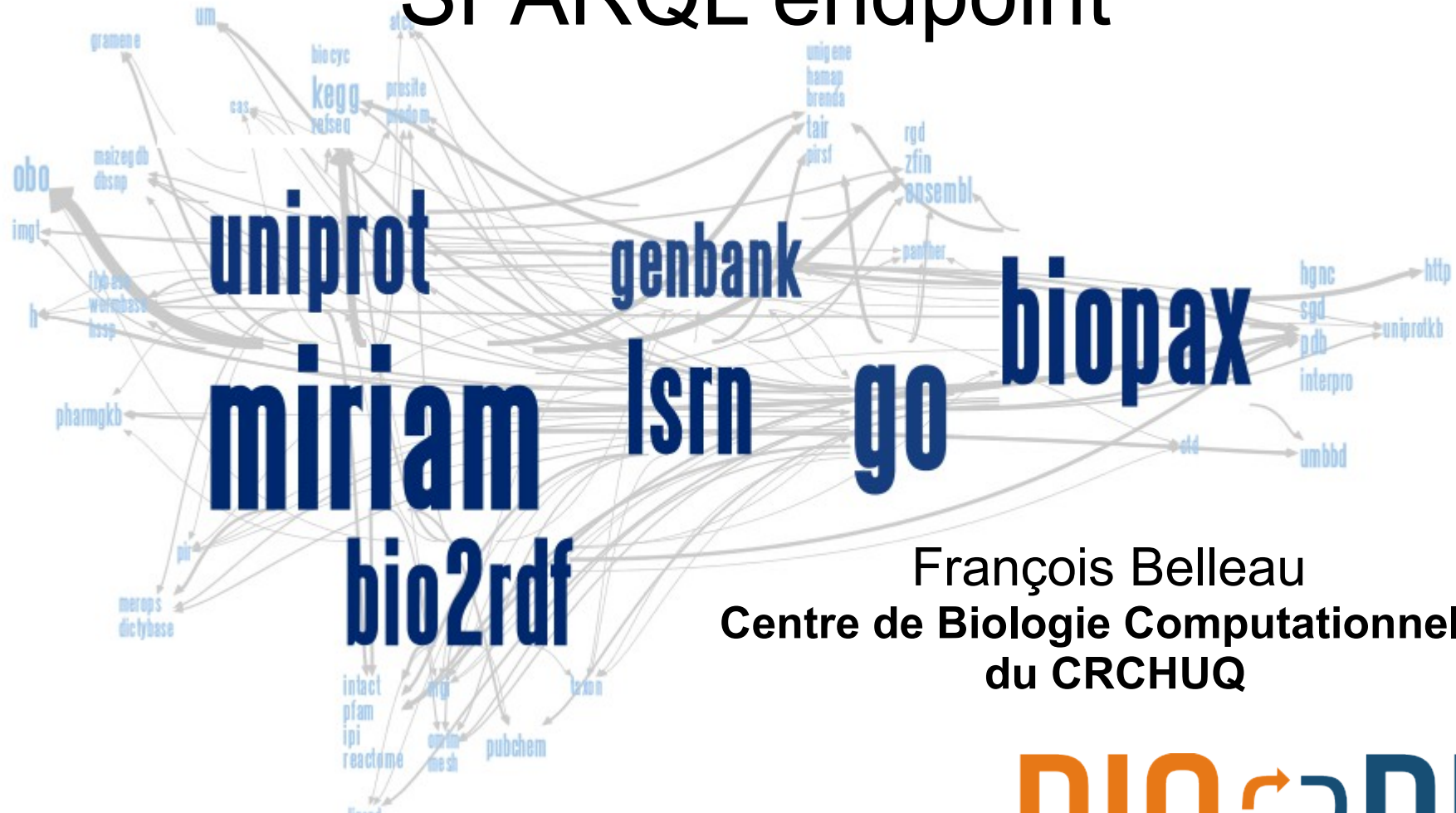


Bio2RDF's namespace SPARQL endpoint



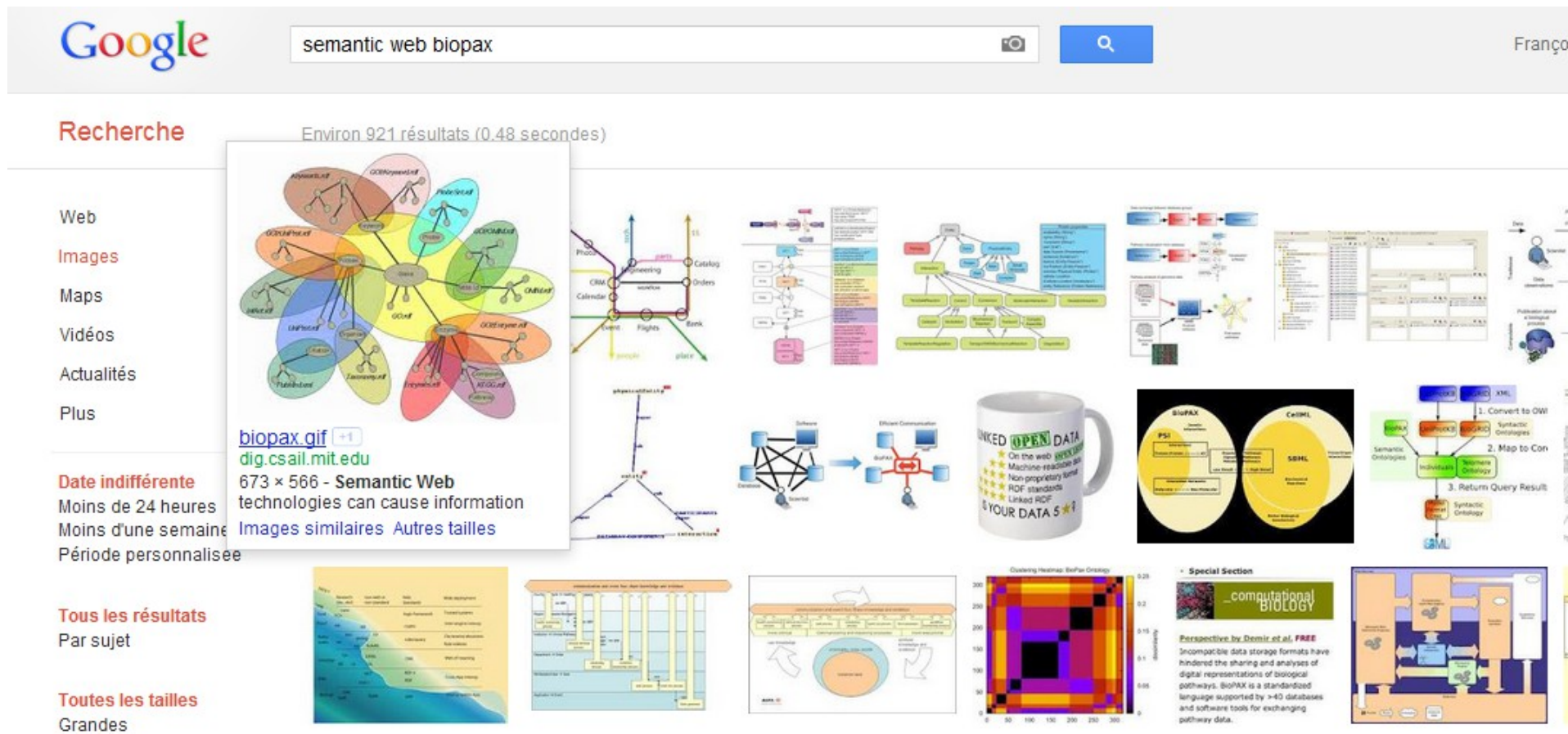
François Belleau
Centre de Biologie Computationnelle
du CRCHUQ

You know them?

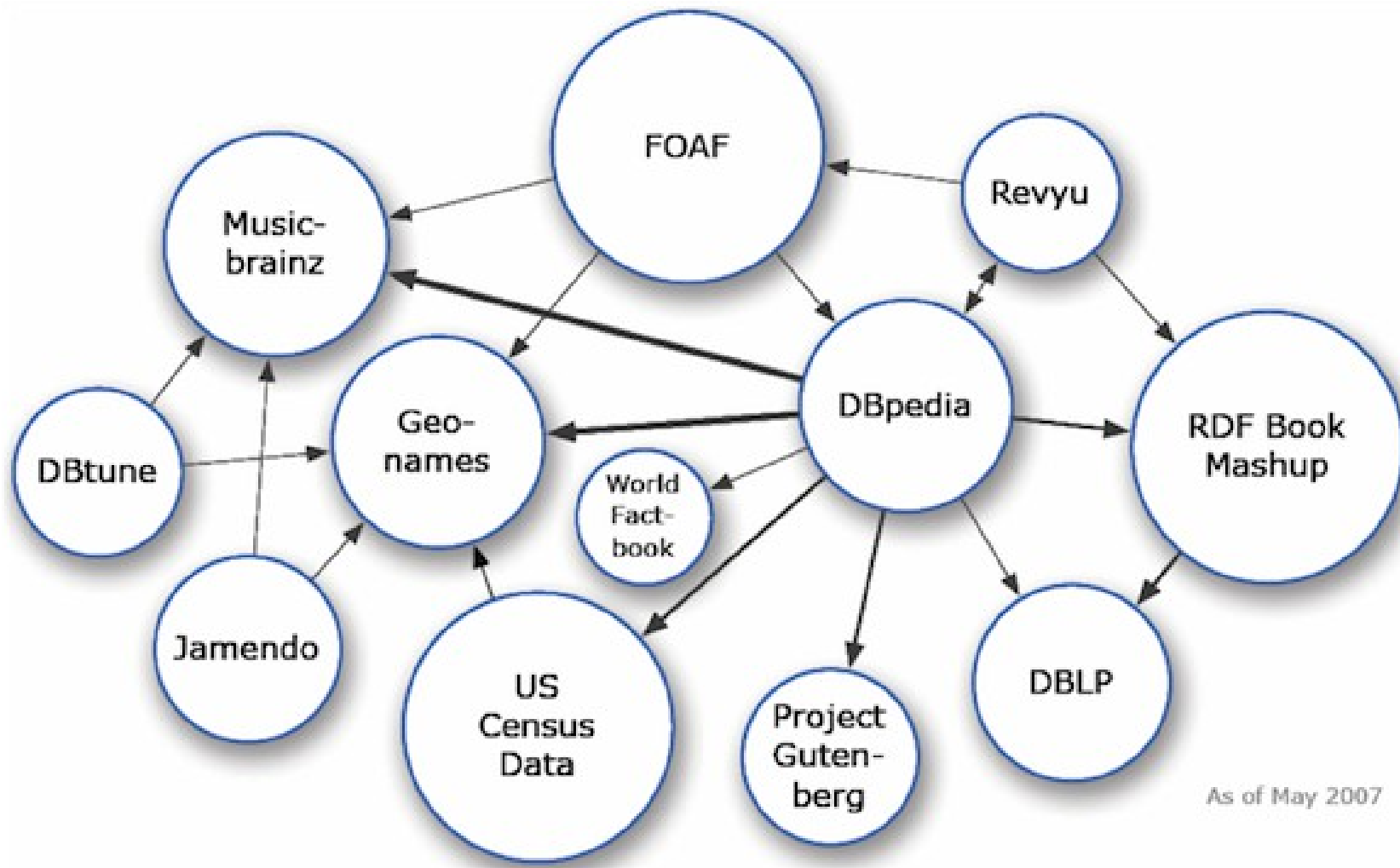
How can we help to navigate in the huge Bioinformatics databases cloud ?



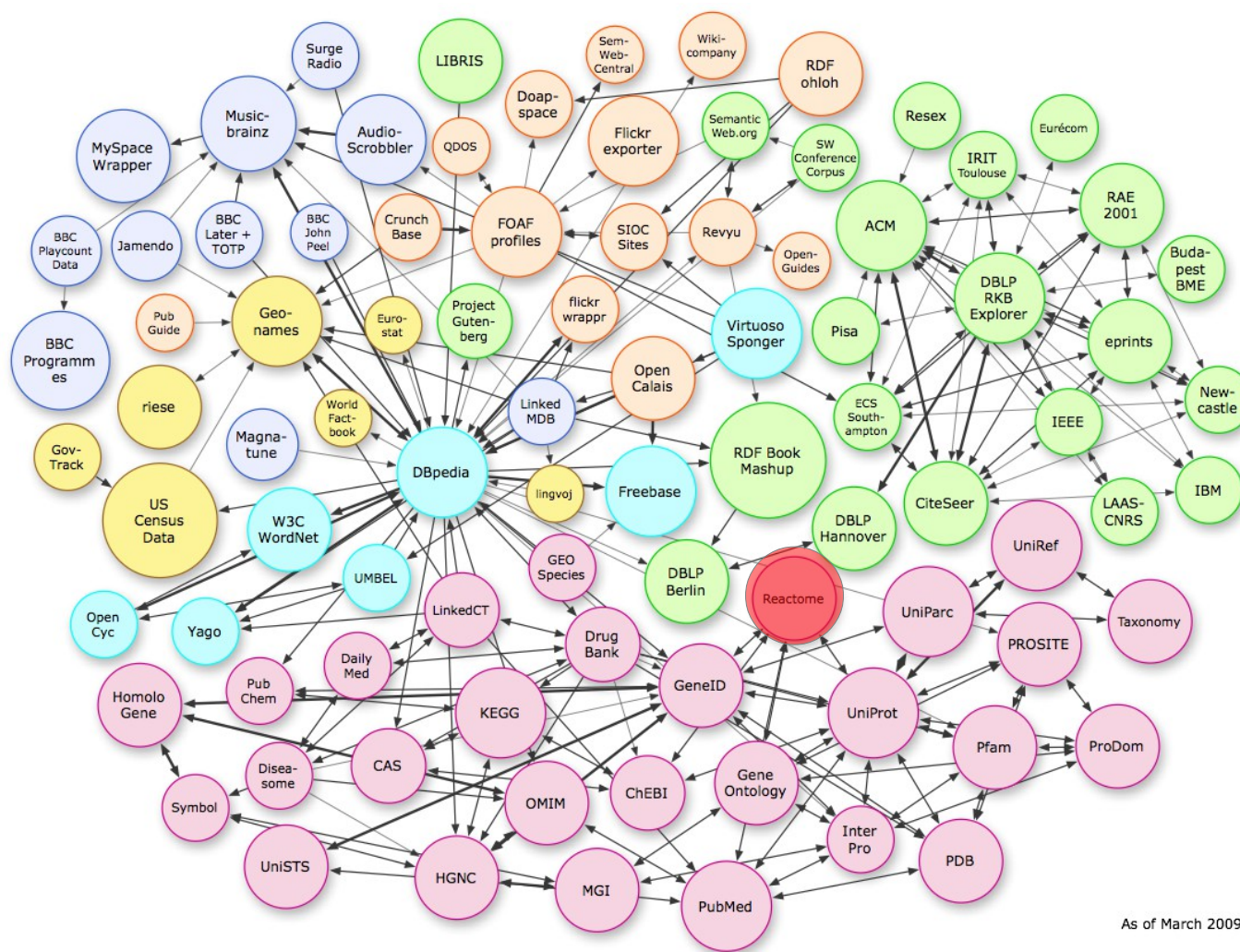
2005 BioPAX.gif next to Semantic Web image vision of Tim Berner Lee



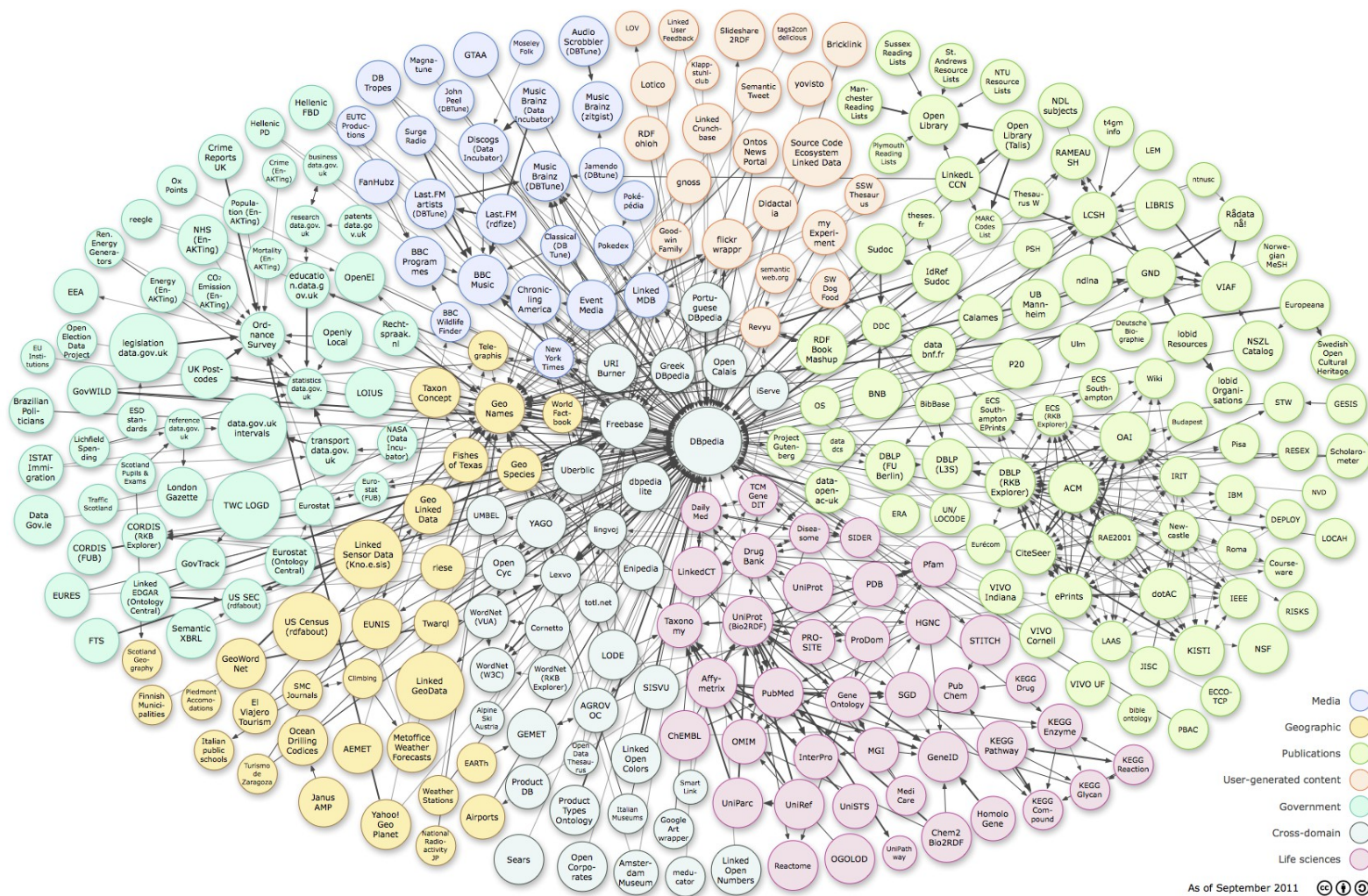
Linked Data in 2007



Bio2RDF contribution in 2009



2011 Linked Data cloud



Databases of databases names

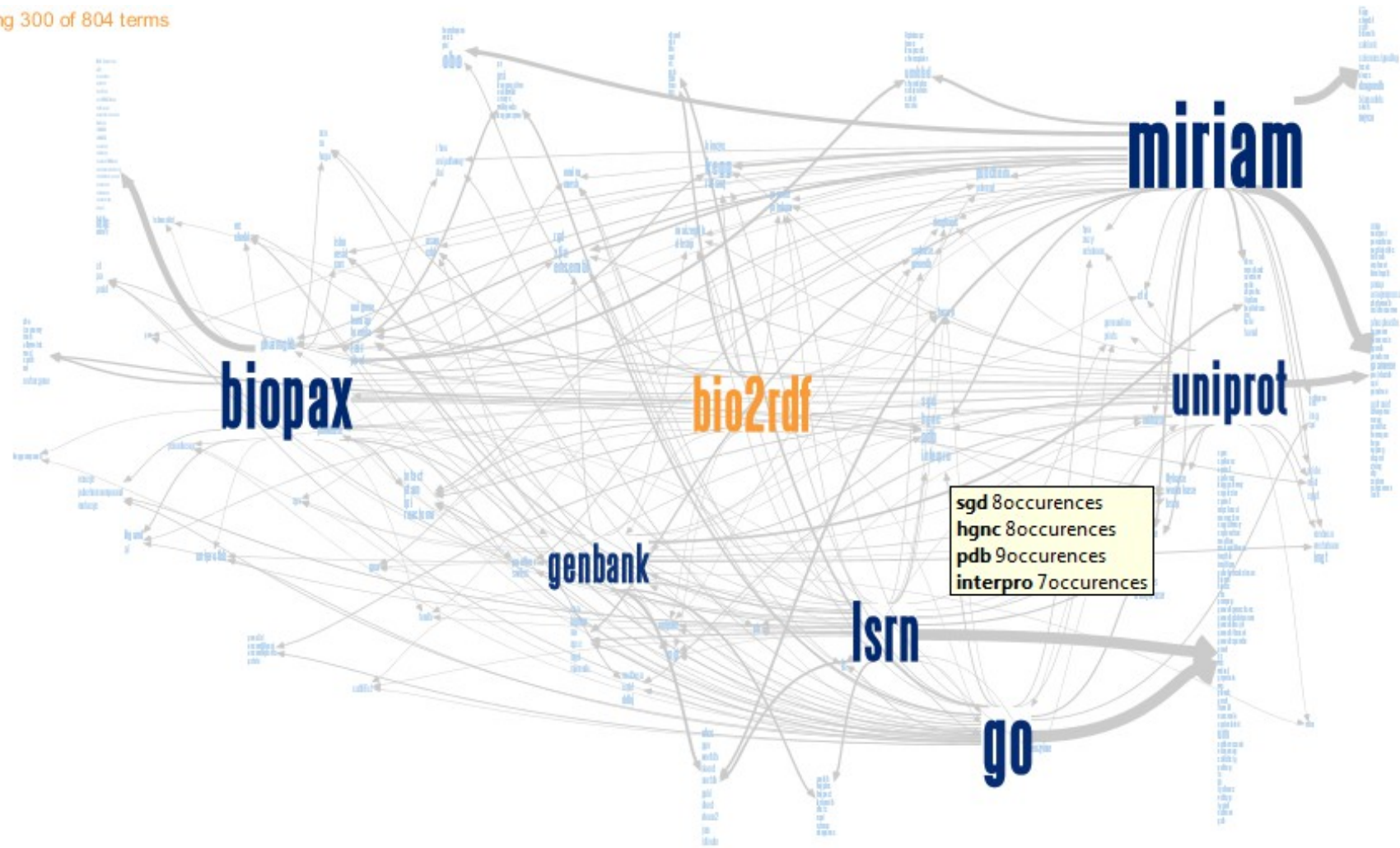
- PathGuide
- Bioinformatics.ca Links Directory
- Annual NAR Database issue
- Go, Uniprot, Genbank cross-reference list
- LSRN initiative
- MIRIAM EBI project
- BioPAX dataprovider community
- Bio2RDF Linked Data space

Two interesting questions

- Which namespace are the most popular for identifying database ?
- How far is the BioPAX community to adopt MIRIAM new namespace standard ?

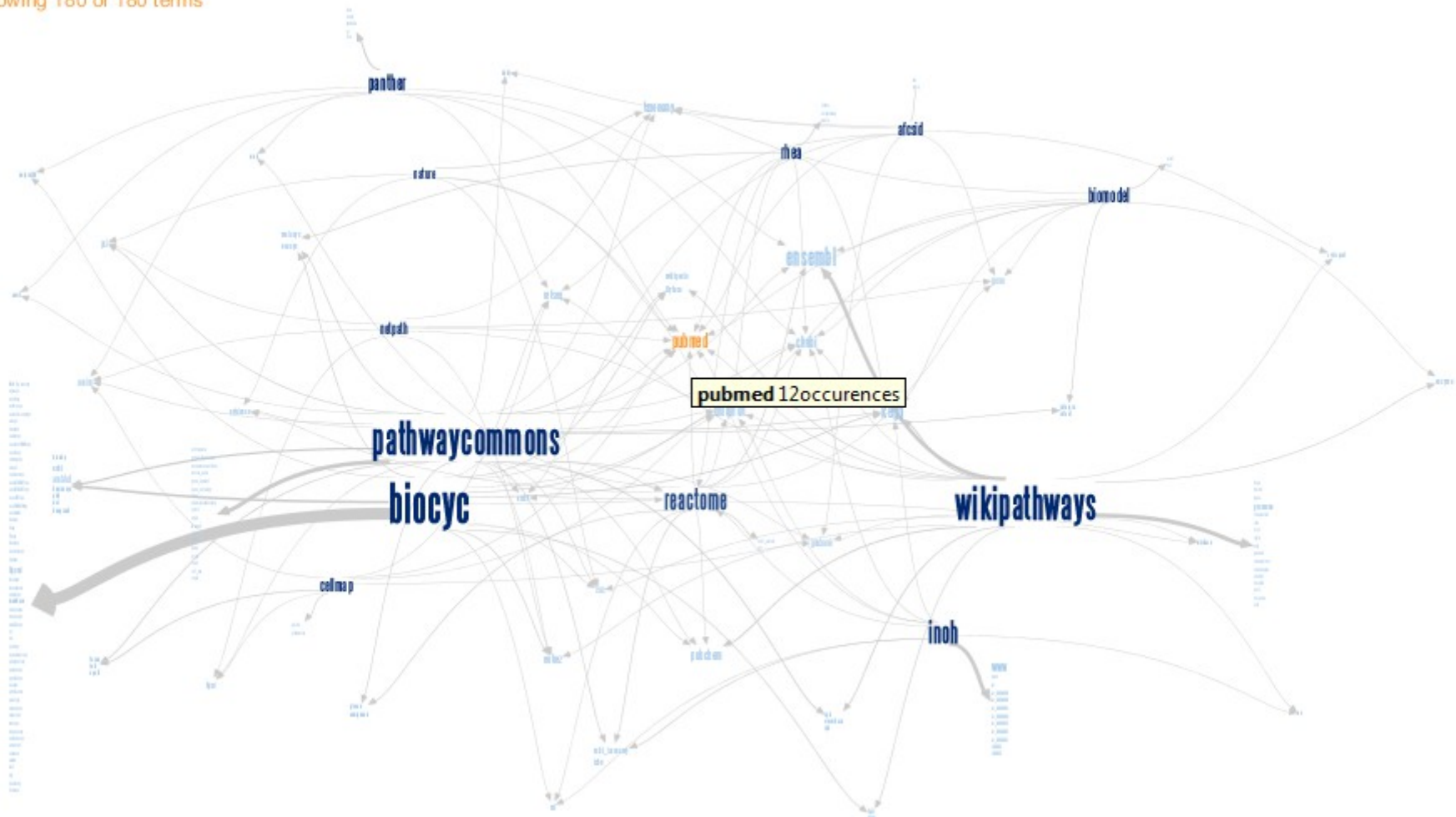
Which namespaces are the most popular to identify a database ?

Showing 300 of 804 terms



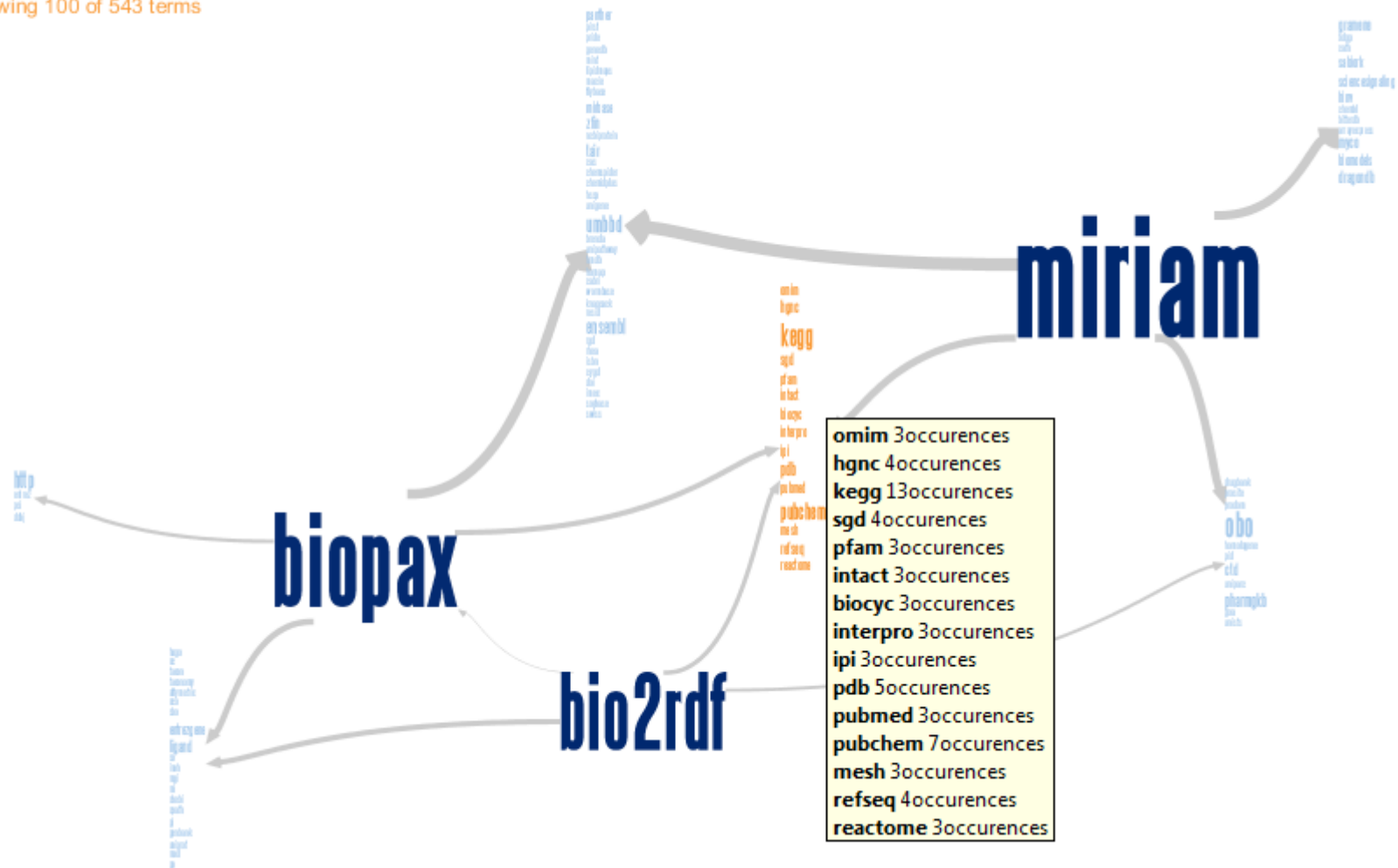
Namespaces collection used by the BioPAX data provider community.

Showing 180 of 180 terms



How far is the BioPAX community to adopt MIRIAM new namespace standard ?

Showing 100 of 543 terms



How we did this ?

To answer a complex question
we first need to build the
database that will potentially
answer it: a semantic mashup.

A ***mashup***, in web development, is a web page, or web application, that uses and combines data, presentation or functionality from two or more sources to create new services. The term implies easy, fast integration, frequently using open Application programming interfaces (API) and data sources to produce enriched results that were not necessarily the original reason for producing the raw source data.

Building a mashup is a lot easier when using Semantic Web technologies like RDF and SPARQL design for data interoperability.

A three steps method

- Get the data from the data provider and transform it into RDF, we use Talend open source Eclipse base ETL software.
- Load the data in a triplestore many software are available (Virtuoso, Sesame, Jena, store, Mulgara, etc.) to load your mashup
- Explore the new dataset using specialised user interface (RelFinder, Virtuoso facet browser)
- Design your SPARQL query and get the answer

Data provider xref resource

- <http://www.ebi.ac.uk/miriam/main/> (XML format)
- Bio2RDF DNS zone description file (text)
- <http://lsrn.org/> (RDF/XML)
- http://www.geneontology.org/doc/GO.xrf_abbs
(key/value format)
- <http://www.uniprot.org/docs/dbxref> (key/value format)
- http://www.ncbi.nlm.nih.gov/genbank/collab/db_xref/
(HTML)
- 12 BioPAX providers (Reactome, Biomodels, Biocyc, Panther, INOH, etc)

Lesson #1

Produce RDF triples with a professional ETL tool.

Talend ETL opensource free software

www.talend.com/index.php

talend*
open integration solutions

downloads | store | blog | company | contacts

Products | Solutions | Services | Resources | Partners | Community

Search...

Big Data for the Masses

Introducing Talend Open Studio for Big Data
Intuitive, open source tools that simplify big data technologies

Learn More | Download Now

Generic | Load | Pig Load and Process 1

tHDFSGet_1 | row2 (Pig) | tPigLoad_1 | tPigAggregate_1 | row3 (Pig) | tPigStoreResult_1

Extract BIG DATA and move to DB2

tHDFSGet_1 | row4 (Main) | tDB2Output_1

tHDFSConnect | tHDFSGet | tHDFSInput | tHDFSOutput | tHDFSPut | Hive | Pig

Big Data | Talend v5 | Cloud | Download

Download Now!

- Data Integration
- Data Quality
- MDM
- Big Data
- ESB
- BPM

White paper

Open Source ESB: Top 10 reasons

Read also:
[Legally Avoid the Data Tax](#)
[Leveraging Open Source Data Quality](#)

Webinar

Talend Open Studio for Dummies

Watch also:
[Rapid MDM Deployment](#)
[Big Data with Talend](#)

talend* forge

Join our community!

NEWS

Talend Partners with Google for BigQuery
Talend Press Release

Talend Visionary in Two New Gartner Magic Quadrants
Talend Press Release

<http://www.talend.com/index.php>

Talend workflows to convert HTML Genbank page to triples and MIRAM XML

The image displays three screenshots of the Talend Open Studio for Data Integration (5.1.0.r82787) interface, showing workflows for converting HTML Genbank pages to triples and MIRAM XML.

Top Screenshot: Shows the main workspace with a workflow diagram. The left sidebar lists the repository structure, including folders like `links_directory`, `nar`, `pathguide`, and `PROD`. The right sidebar shows the component palette with categories like `Applications Métier`, `Bas de données`, `Big Data`, `Business Intelligence`, `Cloud`, `Code Utilisateur`, `Drivers`, `DotNET`, `ESB`, `Fichier`, and `Internet`.

Bottom Left Screenshot: Shows the `tSplitRow_1` component configuration. The `Mapping des colonnes` tab is active, displaying a table with columns: `subject`, `predicate`, `object`, and `isLiteral`. The table contains mappings for various RDF properties and namespaces.

subject	predicate	object	isLiteral
out5.subject	"bio2rdf#name"	out5.object	true
out5.subject	"rdf:type"	"http://genbank.xr..."	false
out5.subject	"rdfs:label"	out5.object + "(" + "	true
out5.subject	"dc:identifier"	"genbank_xref" + "	true
out5.subject	"bio2rdf#namespac..."	"ns:genbankxref"	true
out5.subject	"bio2rdf#urn"	"urn:bm:" + names...	false
out5.subject	"bmm:" + namespa..."	"http://ns.bio2rdf.o..."	false
out5.subject	"rdfs:label"	namespace:IdMemo...	true
out5.subject	"dc:identifier"	"ns:" + namespace...	true
out5.subject	"bmm:" + namespa..."	"bio2rdf#namespac..."	true
out5.subject	"bmm:" + namespa..."	"urn:bm:" + names...	false
out5.subject	"bmm:" + namespa..."	"owl:sameAs"	false

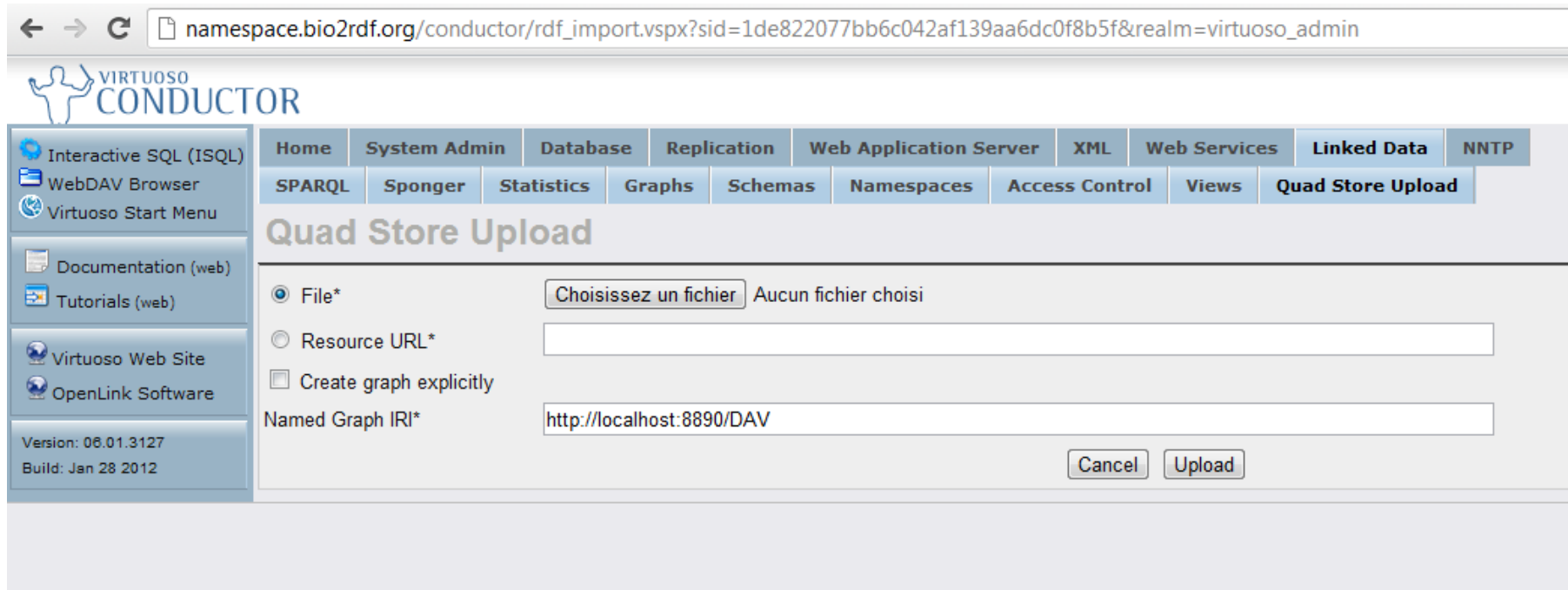
Bottom Right Screenshot: Shows the `tSplitRow_6` component configuration. The `Mapping des colonnes` tab is active, displaying a table with columns: `subject`, `predicate`, `object`, and `isLiteral`. The table contains mappings for various RDF properties and namespaces.

subject	predicate	object	isLiteral
id_tMemorizeRows...	"dc:identifier"	out2.id.toLowerCa...	true
id_tMemorizeRows...	"rdfs:label"	out2.name + "(" + "	true
id_tMemorizeRows...	"rdf:type"	"miriam#datatype"	false
id_tMemorizeRows...	"rdf:type"	"http://mir.bio2rdf..."	false
id_tMemorizeRows...	"mirid"	out2.id	true
id_tMemorizeRows...	"mirurn"	out2.urn	false

Lesson #2

Publish with a SPARQL endpoint.
(to get a free 5 stars cup)

Load RDF triples into a triplestore (we use Openlink Virtuoso)



The screenshot shows the Virtuoso Conductor web interface in a browser. The address bar displays the URL: `namespace.bio2rdf.org/conductor/rdf_import.vsp?sid=1de822077bb6c042af139aa6dc0f8b5f&realm=virtuoso_admin`. The page features a top navigation bar with tabs: Home, System Admin, Database, Replication, Web Application Server, XML, Web Services, Linked Data, and NNTP. Below this is a secondary navigation bar with tabs: SPARQL, Sponger, Statistics, Graphs, Schemas, Namespaces, Access Control, Views, and Quad Store Upload. The main content area is titled "Quad Store Upload" and contains the following form elements:

- ☒ File*: A button labeled "Choisissez un fichier" and the text "Aucun fichier choisi".
- ☐ Resource URL*: An empty text input field.
- ☐ Create graph explicitly
- Named Graph IRI*: A text input field containing the value `http://localhost:8890/DAV`.
- Buttons: "Cancel" and "Upload".

A left sidebar contains links to "Interactive SQL (ISQL)", "WebDAV Browser", "Virtuoso Start Menu", "Documentation (web)", "Tutorials (web)", "Virtuoso Web Site", and "OpenLink Software". At the bottom of the sidebar, it shows "Version: 06.01.3127" and "Build: Jan 28 2012".

<http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/>

Full text search

← → ↻ namespace.bio2rdf.org/fct/ ☆ [icons]

OPENLINK
SOFTWARE

POWERED BY
VIRTUOSO

Text Search Entity Label Lookup Entity URI Lookup [Featured](#) | [Demo Queries](#) | [About](#)

Precision Search & Find

Search Text

Hint: You can [add this engine](#) in search bar

Faceted Search & Find



OpenLink Virtuoso version 06.01.3127, on Linux (x86_64)
Copyright © 2009-2011 OpenLink Software

← → ↻ namespace.bio2rdf.org/fct/facet.vsp?cmd=text&sid=85

OPENLINK
SOFTWARE

POWERED BY
VIRTUOSO

Displaying Ranked Entity Names and Text summaries where:

Entity1 has any Attribute with Value "go" Drop.

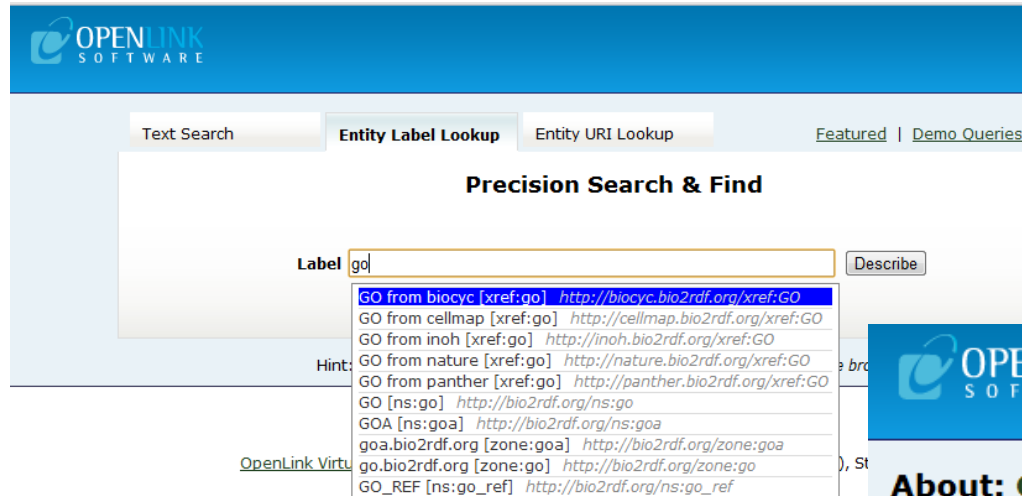
[View query as SPARQL](#) [Facet permalink](#)

Showing 1-15 of 15 total

Isrn:GO_REF	Gene Ontology Database references [Isrn:GO_REF]	
Isrn:TIGR_REF		... org tdb GO REF GO REF.
http://panther.bio2rdf.org/xref:GO	GO from panther [xref:go]	
http://biocyc.bio2rdf.org/xref:GO	GO from biocyc [xref:go]	GO from biocyc xref:go biopax xref:go.
http://nature.bio2rdf.org/xref:GO	GO from nature [xref:go]	
http://cellmap.bio2rdf.org/xref:GO	GO from cellmap [xref:go]	
http://inoh.bio2rdf.org/xref:GO	GO from inoh [xref:go]	
ns:go	GO [ns:go]	ns:gons:gons:gons:gogogoGO ns:go GO ns:go GO ns:go.
ns:go_ref	GO_REF [ns:go_ref]	
uniprot_dbxref:DB-0037	Gene Ontology [uniprot_dbxref:db-0037]	GO.
uniprot_dbxref:DB-0071	PeptideAtlas [uniprot_dbxref:db-0071]	... net sbeams cgi PeptideAtlas Search action GO search key...
Isrn:INSD	International Nucleotide Sequence Database Collaboration (GenBank, EMBL, DDBJ) [Isrn:INSD]	... uk cgi bin emblfetch style html Submit Go id.

<http://namespace.bio2rdf.org/fct/>

Discover entity name and browse the triplestore



OPENLINK SOFTWARE

Text Search Entity Label Lookup Entity URI Lookup Featured | Demo Queries

Precision Search & Find

Label Describe

Hint:

- GO from biocyc [xref:go] <http://biocyc.bio2rdf.org/xref:GO>
- GO from cellmap [xref:go] <http://cellmap.bio2rdf.org/xref:GO>
- GO from inoh [xref:go] <http://inoh.bio2rdf.org/xref:GO>
- GO from nature [xref:go] <http://nature.bio2rdf.org/xref:GO>
- GO from panther [xref:go] <http://panther.bio2rdf.org/xref:GO>
- GO [ns:go] <http://bio2rdf.org/ns:go>
- GOA [ns:goa] <http://bio2rdf.org/ns:goa>
- goa.bio2rdf.org [zone:goa] <http://bio2rdf.org/zone:goa>
- go.bio2rdf.org [zone:go] <http://bio2rdf.org/zone:go>
- GO_REF [ns:go_ref] http://bio2rdf.org/ns:go_ref

OpenLink Virtuoso



OPENLINK SOFTWARE

Facets Description

About: GO [ns:go]

An Entity of Type : ns.bio2rdf.org, within Data Space : namespace.bio2rdf.org
[Constrain facet on this type](#)

Attributes	Values
rdf:type	http://ns.bio2rdf.org
rdfs:label	GO [ns:go]
sameAs	Gene Ontology Database [Isrn:GO] Gene Ontology Database identifier [qenbank_xref:go] go.bio2rdf.org [zone:go] Gene Ontology [uniprot_dbxref:db-0037]
dc:identifier	ns:go
bio2rdf_ontology:namespace	go nsns
bio2rdf_ontology:ns	nsns
bio2rdf_ontology:urn	urn:bm:go

Lesson #3

Consume as you like.

HTTP GET to obtain RDF from URI,
SPARQL endpoint,
SOAP services returning RDF,
semantic web new software...

The needed SPARQL query to draw the previous graph using ManyEyes service

← → ↻ namespace.bio2rdf.org/sparql

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

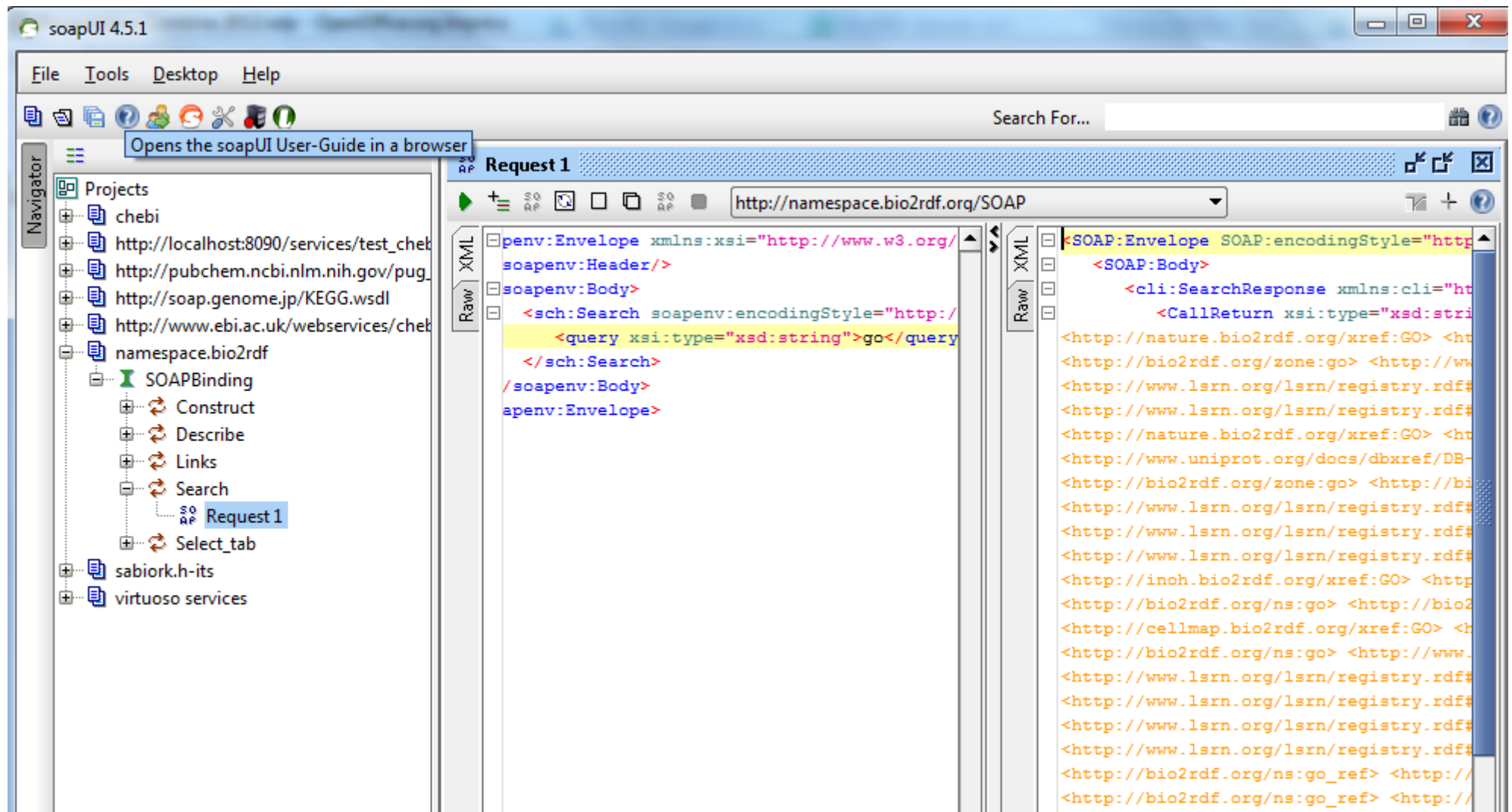
```
select distinct ?type ?urn count(*)
where {
  ?subject <http://bio2rdf.org/bio2rdf_ontology#urn> ?urn .
  ?subject a ?type .
}
order by 1 2
```

'sparql?default-graph-uri=&query=select+distinct+%3Ftype+%3Furn+coi

	urn	callret-2
	urn:bm:	1
	urn:bm:10403s_rastcyc	1
	urn:bm:affy	1
	urn:bm:affymetrix	1
	urn:bm:agrocyc	1
	urn:bm:ano2cyc	1
	urn:bm:anthracyc	1
http://biopax.bio2rdf.org/xref	urn:bm:aracyc	2
http://biopax.bio2rdf.org/xref	urn:bm:aurantimonascyc	1
http://biopax.bio2rdf.org/xref	urn:bm:biocyc	1
http://biopax.bio2rdf.org/xref	urn:bm:brenda	2
http://biopax.bio2rdf.org/xref	urn:bm:bsubcyc	1
http://biopax.bio2rdf.org/xref	urn:bm:c00002	1
http://biopax.bio2rdf.org/xref	urn:bm:c00035	1
http://biopax.bio2rdf.org/xref	urn:bm:cabri	1

<http://namespace.bio2rdf.org/sparql>

Use a SOAP service



<http://namespace.bio2rdf.org/bio2rdf/services.wsdl>

Discover your relations graphically with RelFinder

RelFinder interface showing a query and its results.

Query:

- (1) Gene Ontology Database [Isrn:GO]
- (2) Gene Ontology Database identifier [gen]
- (3) go.bio2rdf.org [zone:go]

Filter by: relations: (6/6)

Results:

number of objects	num	vi
1	6/6	

GO [ns:go]

More Infos: bio2rdf.org

Graph:

The graph displays relationships between various entities. Key nodes include:

- Gene Ontology Database [Isrn:GO]
- Gene Ontology Database identifier [gen]
- go.bio2rdf.org [zone:go]
- GO [ns:go]
- urn:bm:go
- urn

Relationships are indicated by arrows, with labels such as "sameAs" and "urn".

Edit Settings:

Endpoint URI: <http://namespace.bio2rdf.org>

Method: POST

Max Relation Length: 2

Autocomplete URIs: <http://www.w3.org/2000/01/rdf-schema#label>

Autocomplete Language: en

<http://www.visualdataweb.org/refinder.php>

Conclusion

- Building a mashup is easy with the actual software, we still need the RDF data.
- One SPARQL query in the proper triplestore (Bio2RDF's namespace mashup) could answer our two initial questions.
- Why not consider publish your own SPARQL endpoint to make semantic hacker's life easier ?

Acknowledgements

- Bio2RDF is a community project available at <http://bio2rdf.org>
- The community can be joined at
<https://groups.google.com/forum/?fromgroups#!forum/bio2rdf>
- This work was done under the supervision of Dr Arnaud Droit, assistant professor and director of the **Centre de Biologie Computationnelle du CRCHUQ at Laval University**, where a mirror of Bio2RDF is hosted.
- Michel Dumontier, from the Dumontier Lab at Carleton University, is also hosting Bio2RDF server and actually leads the project
- Thanks to all the people member of the Bio2RDF community, and especially Marc-Alexandre Nolin and Peter Ansell, initial developers.

Come in Montreal July 2013 with your SPARQL endpoint and get a FREE cup!

Semantic Trilogy '13

Montreal, Qc, Canada

HOME	Semantic Trilogy Montreal, Quebec, Canada July 7th - 13th 2013
ICBO 2013	1) July 7th - Workshops
CSWC 2013	2) 4th International Conference on Biomedical Ontology - ICBO July 8th ICBO Day 1 July 9th ICBO Day 2
DILS 2013	3) 4th Canadian Semantic Web Conference - CSWC July 10th CSWC
SEMANTIC TRILOGY WORKSHOPS	4) 9th Data Integration in Life Science - DILS July 11th DILS Day 1 July 12th DILS Day 2
SEMANTIC TRILOGY HACKATHON	5) Life Science Hackathon - ST:LSH July 13th
ORGANIZERS	



<http://www.unbsj.ca/sase/csas/data/semantic-trilogy-2013/>