关联数据系统开发实例与平台详解

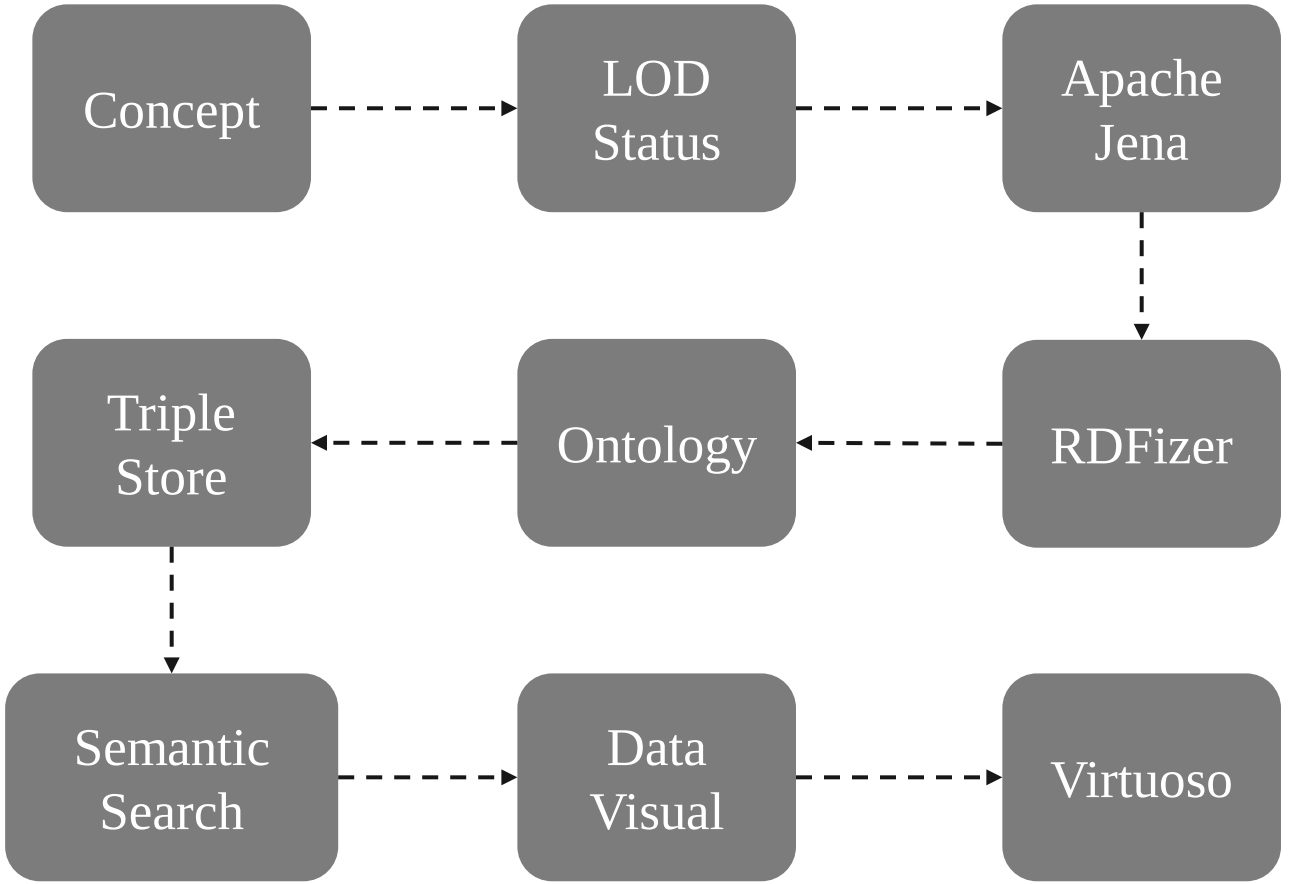
陈 涛

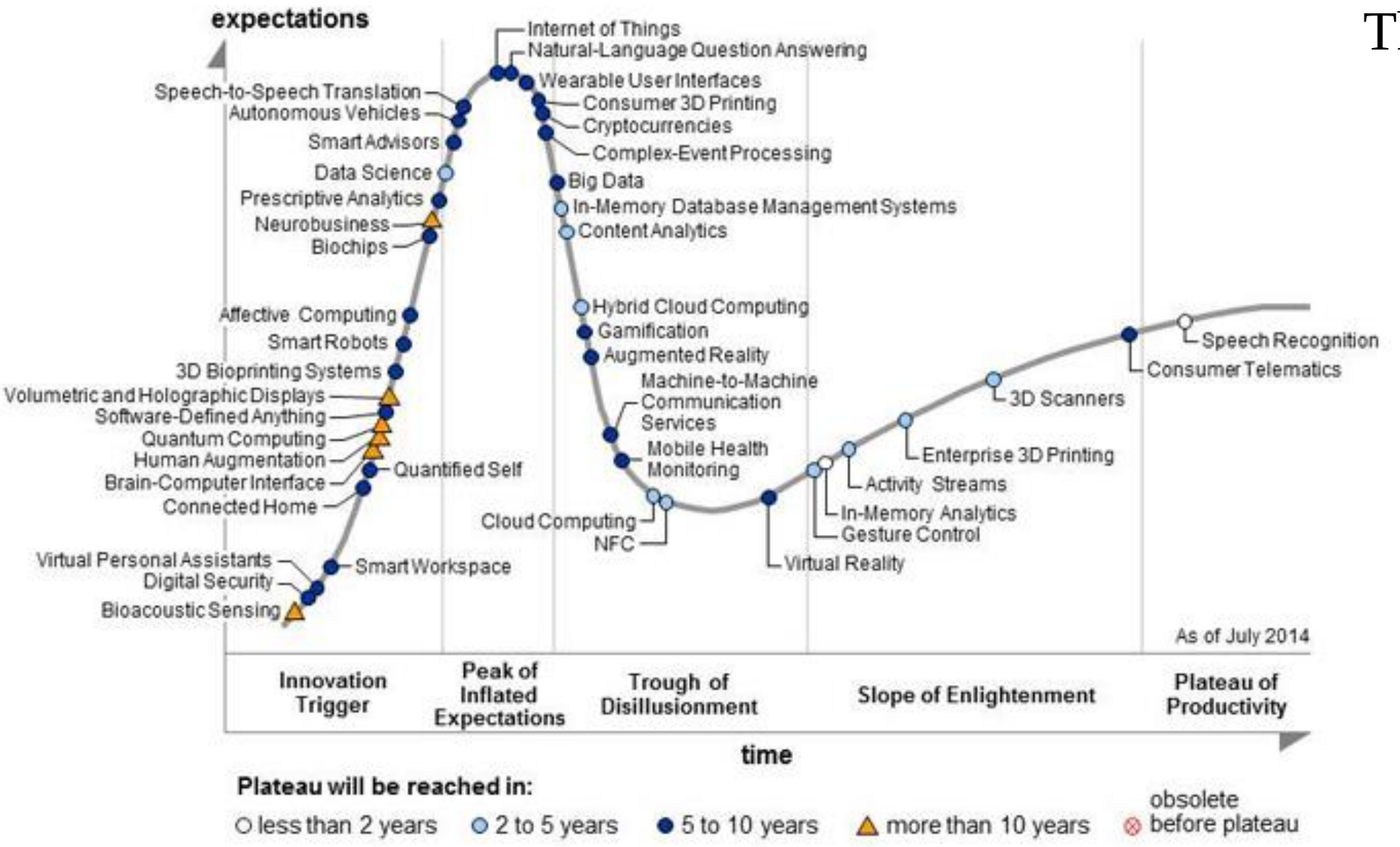
中国科学院上海生命科学信息中心

2015-05-05

QQ群：三人行（语义有你） 150461365

Agenda


Concept
LOD Status
Apache Jena
RDFizer
Ontology
Triple Store
Semantic Search
Data Visual
Virtuoso



The Internet of

~~Things~~

Everything

Gartner

Valuation of IoE

It’s estimated that by the year 2020 more than fifty billion “things” will be connected to the internet.

50,000,000,000

How much is the Internet of everything worth? Cisco says $19 trillion.

19,000,000,000,000

manufacturing (15%) ,healthcare (15%) ,insurance (11%)

Semantic Web

- Semantic web is a **Web of data**.
- The term “Semantic Web” refers to W3C’s vision of the Web of linked data.
- It is about common formats for **integration** and **combination** of data drawn from diverse sources.
- It is also about language for recording how the data **relates** to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.

W3C Semantic Web

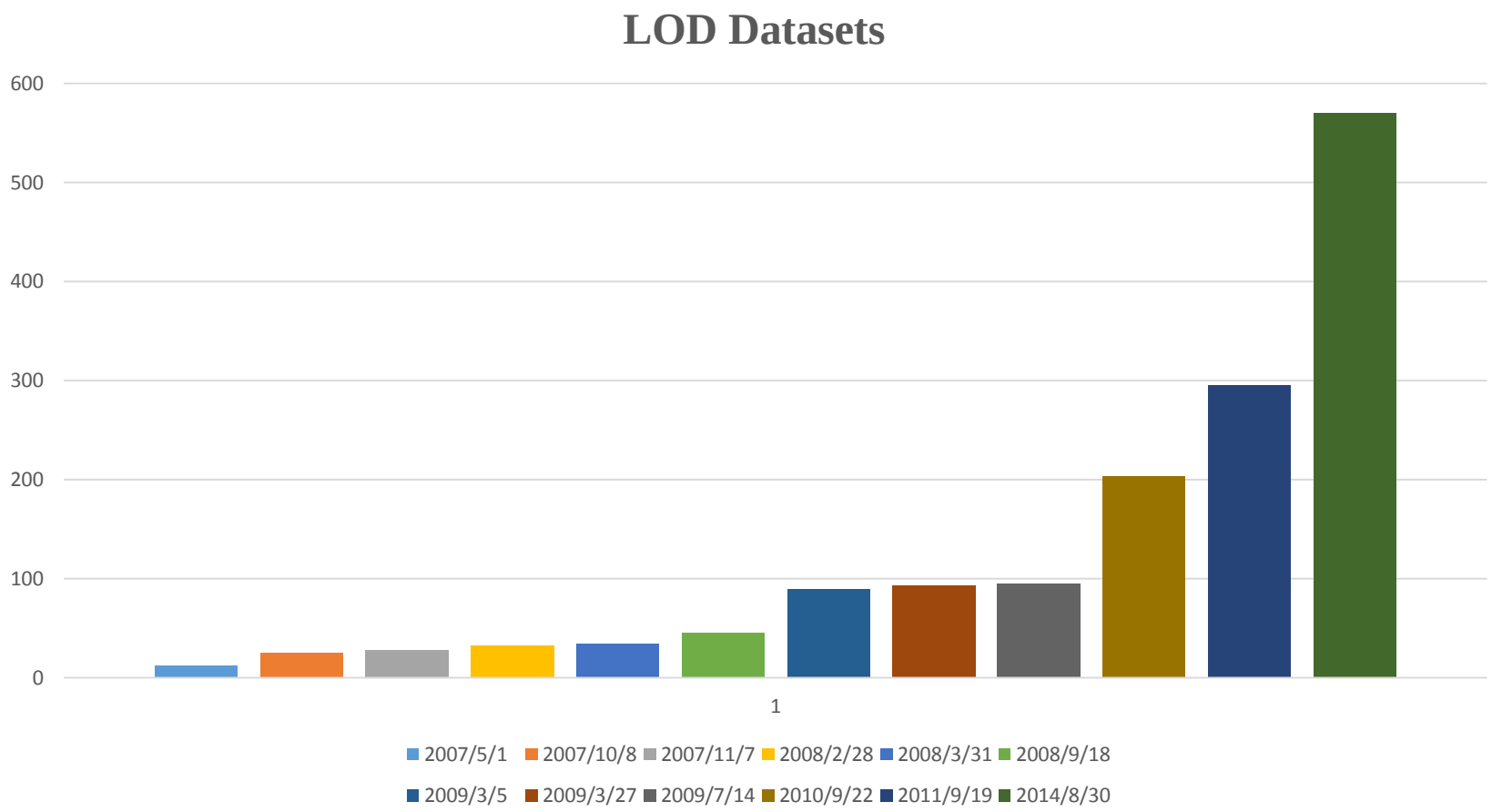
Linked Data

WIKIPEDIA
The Free Encyclopedia

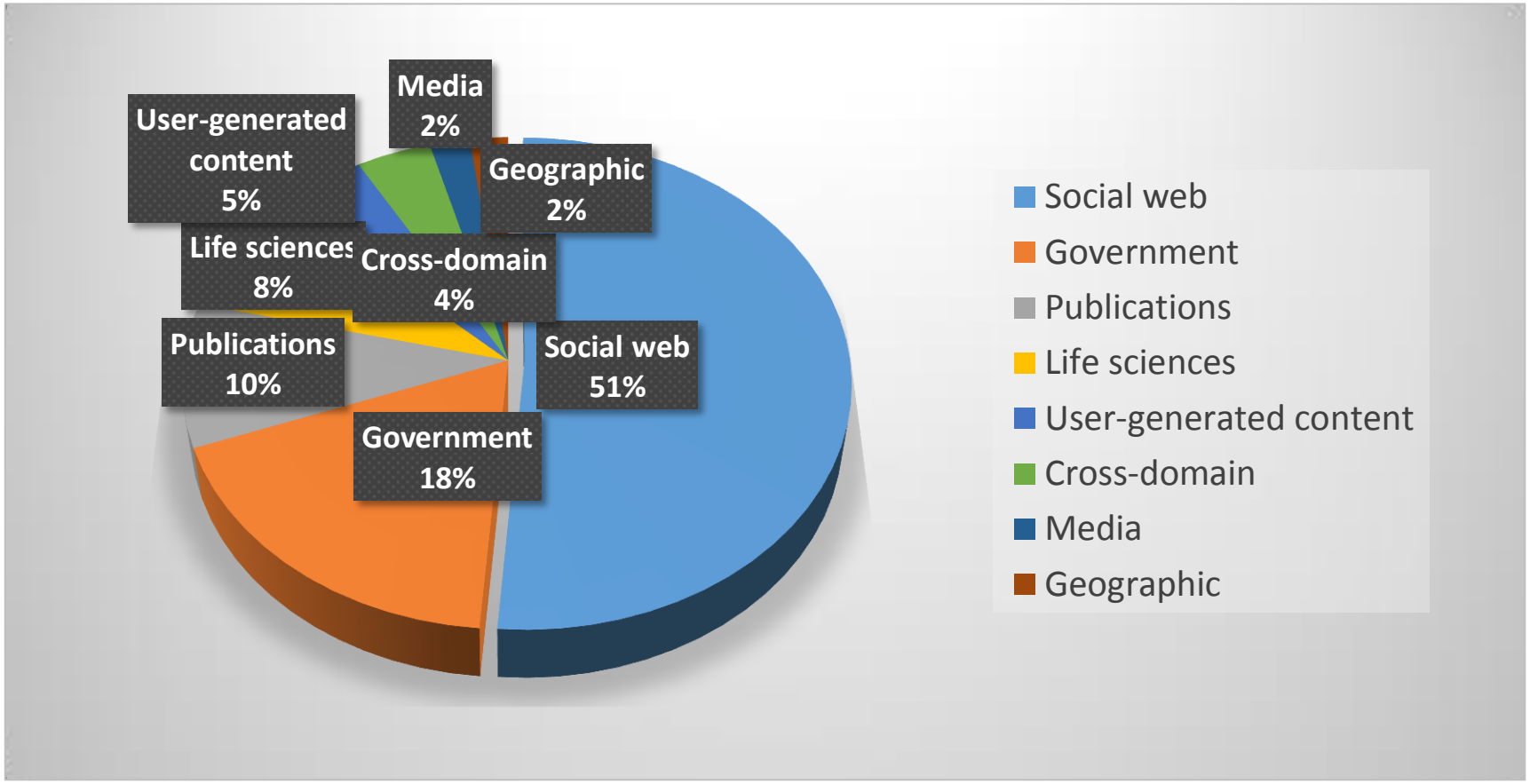
A term used to describe a recommended best practice for exposing, sharing, and connecting pieces of ***data***, ***information***, and ***knowledge*** on the Semantic Web using URIs and RDF.

Connect Distributed Data across the Web

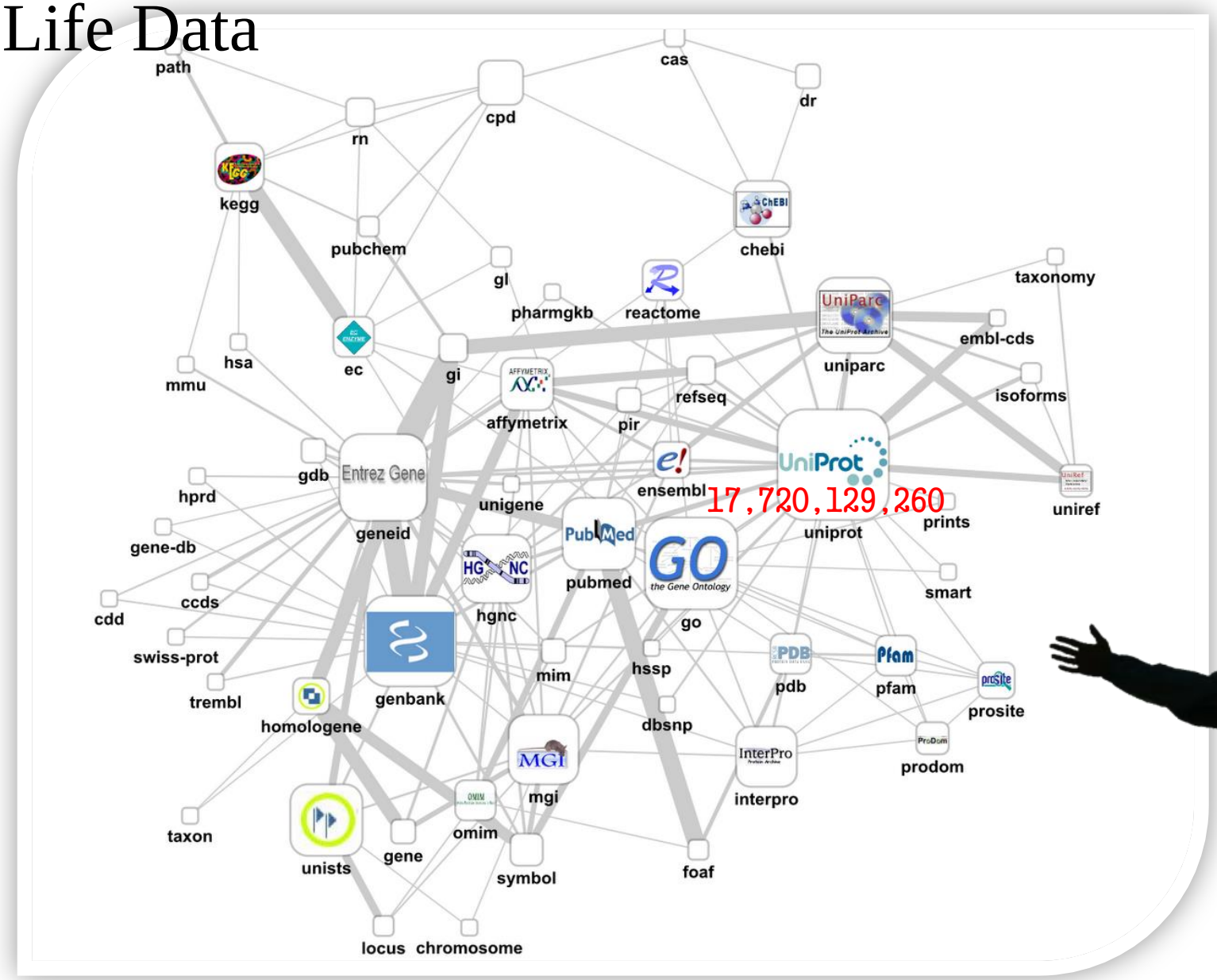
Linked Open Data Cloud



LOD Cloud Status



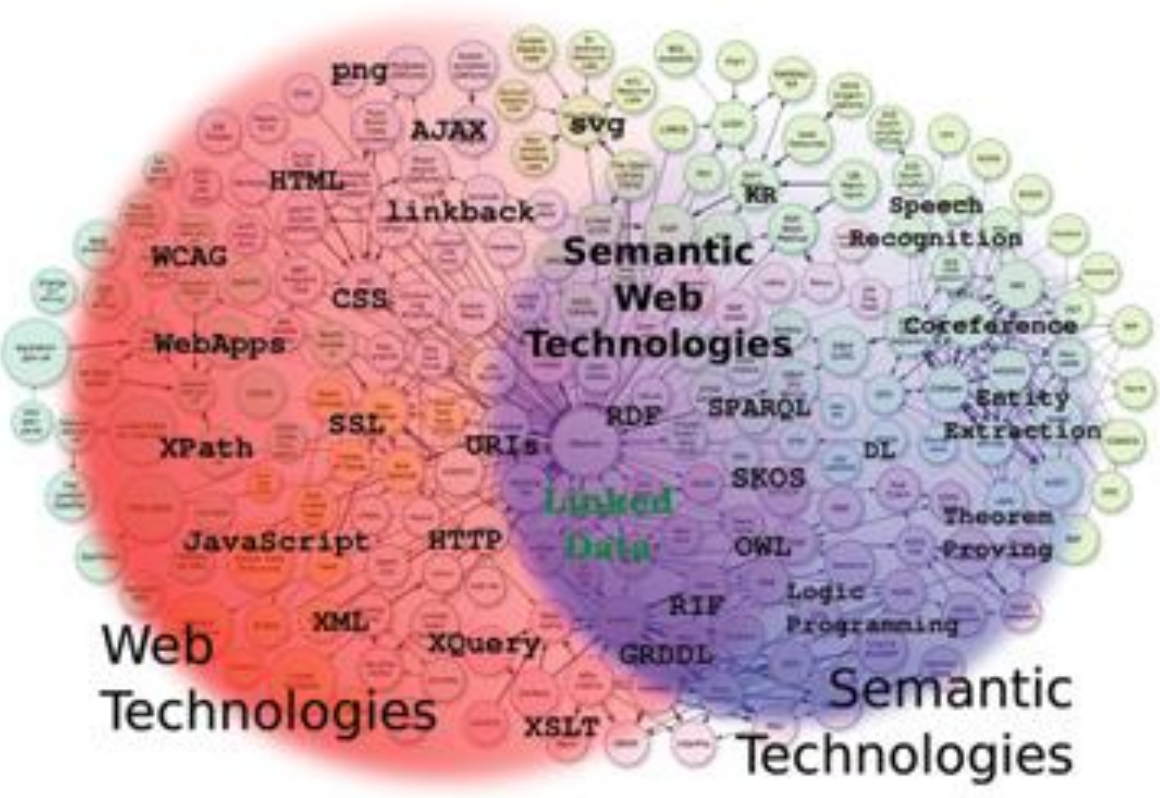
Linked Life Data



Use Cases for SW & LD

Food and Agriculture Organization of the United Nations

Semantic Web Technologies



- RDF
- RDFs
- SPARQL
- SKOS
- OWL
- RIF
- Entity Extraction

… …

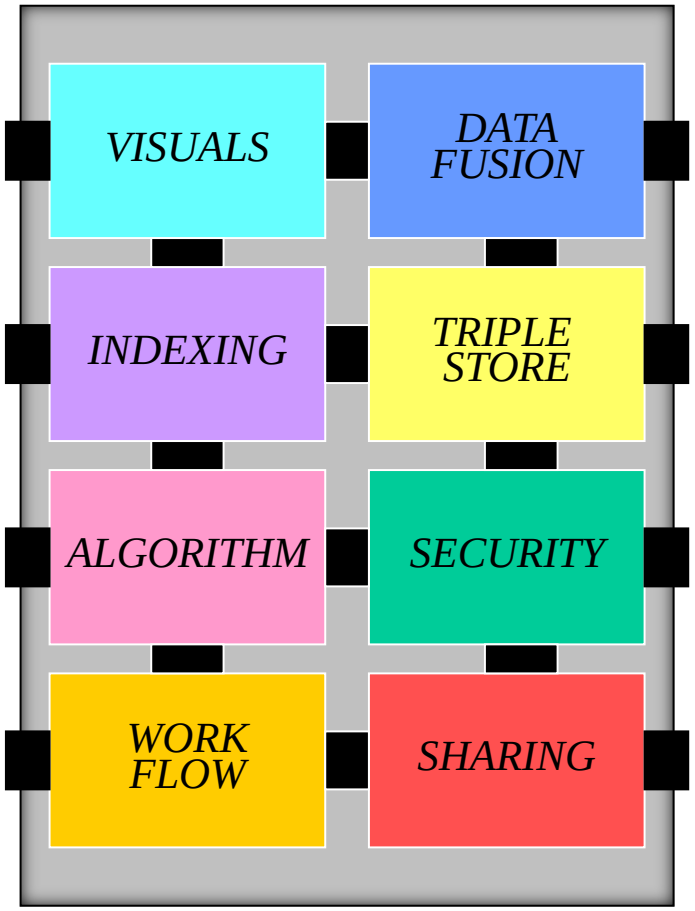
Core Components

Rich perspective on displaying ontology and data

Rapid on free text, complex phrases, semantic search

Apply filter, extractors, transformation algorithms and inference rules

Business or services logic control and work flow



Query and merge data from many different sources, both structured and unstructured

RDF is the core data model; stores triples: Subject, Predicate, Object

The auditing and adjudication of data as it is accessed and transformed

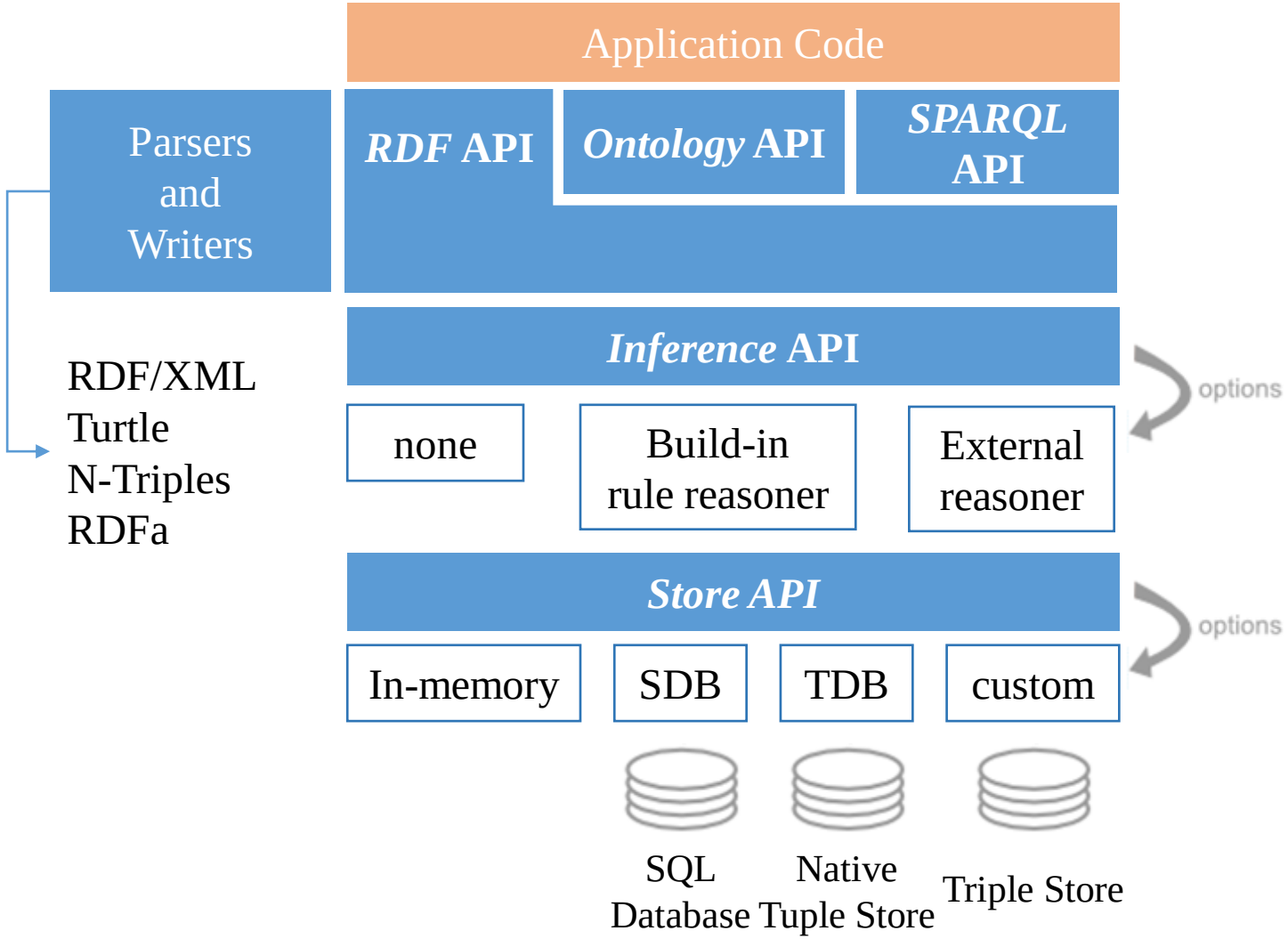
Discovery of web-services, and user workspaces

Framework for *SW* & *LD*

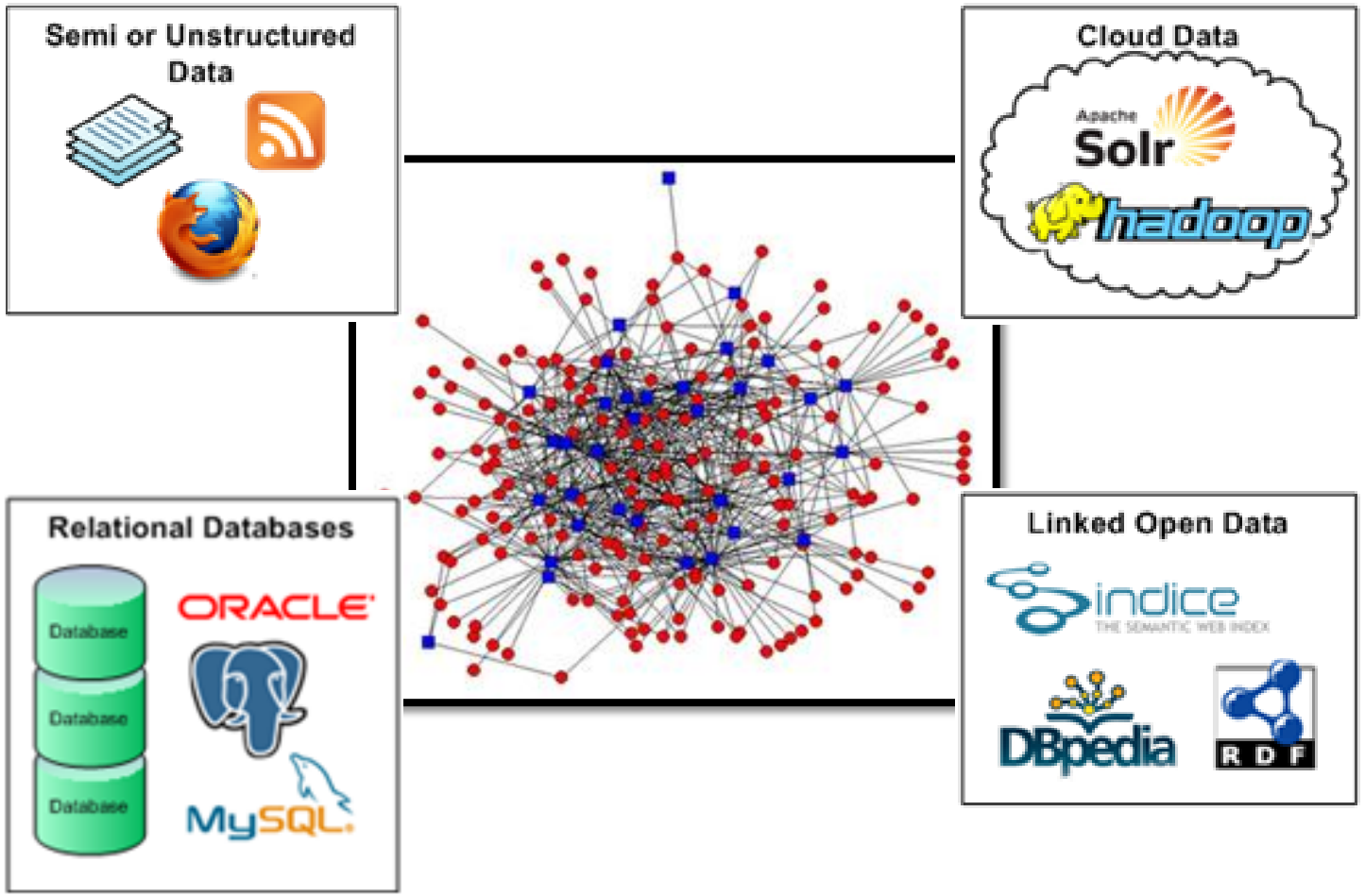
Apache Jena (or Jena in short) is a free and open source Java framework for building semantic web and Linked Data applications.

OpenRDF Sesame is a powerful Java framework for processing and handling RDF data. This includes creating, parsing, storing, inferencing and querying over such data.

Apache Jena

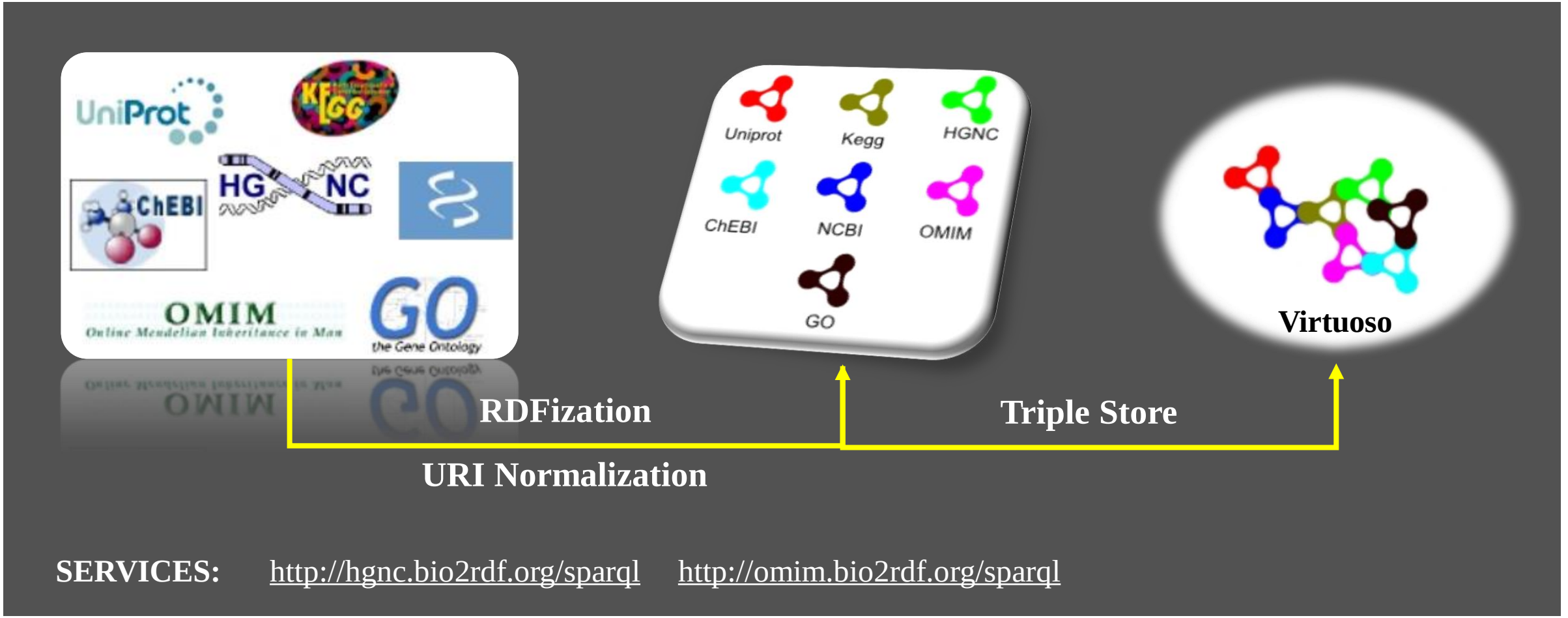

Application Code
Parsers and Writers
RDF API
Ontology API
SPARQL API
RDF/XML
Turtle
N-Triples
RDFa
Inference API
options
none
Build-in rule reasoner
External reasoner
Store API
options
In-memory
SDB
TDB
custom
SQL Database
Native Tuple Store
Triple Store

Data Fusion / Integration Points



Bio2RDF Process

- Dataset: *35* Triples: *11,900,533,153* Entities: *1,108,204,952*

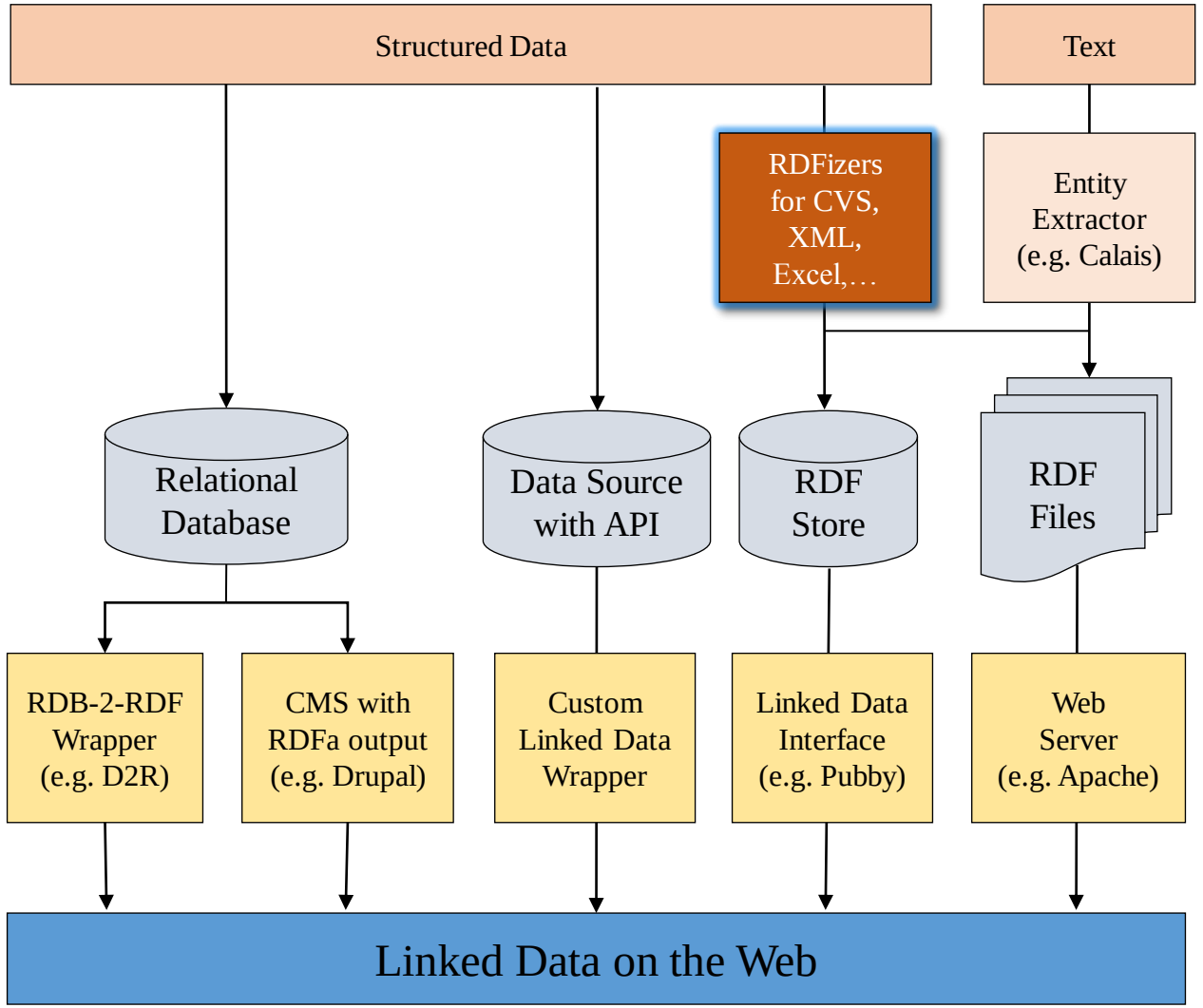


SERVICES: http://hgnc.bio2rdf.org/sparql http://omim.bio2rdf.org/sparql

Fusion Process

1. Download the original data
2. Transform the data into RDF
3. Keeping a link to the original data in the new RDF document
4. Store the converted database in a triple store
5. Give services for these using a RESTful interface

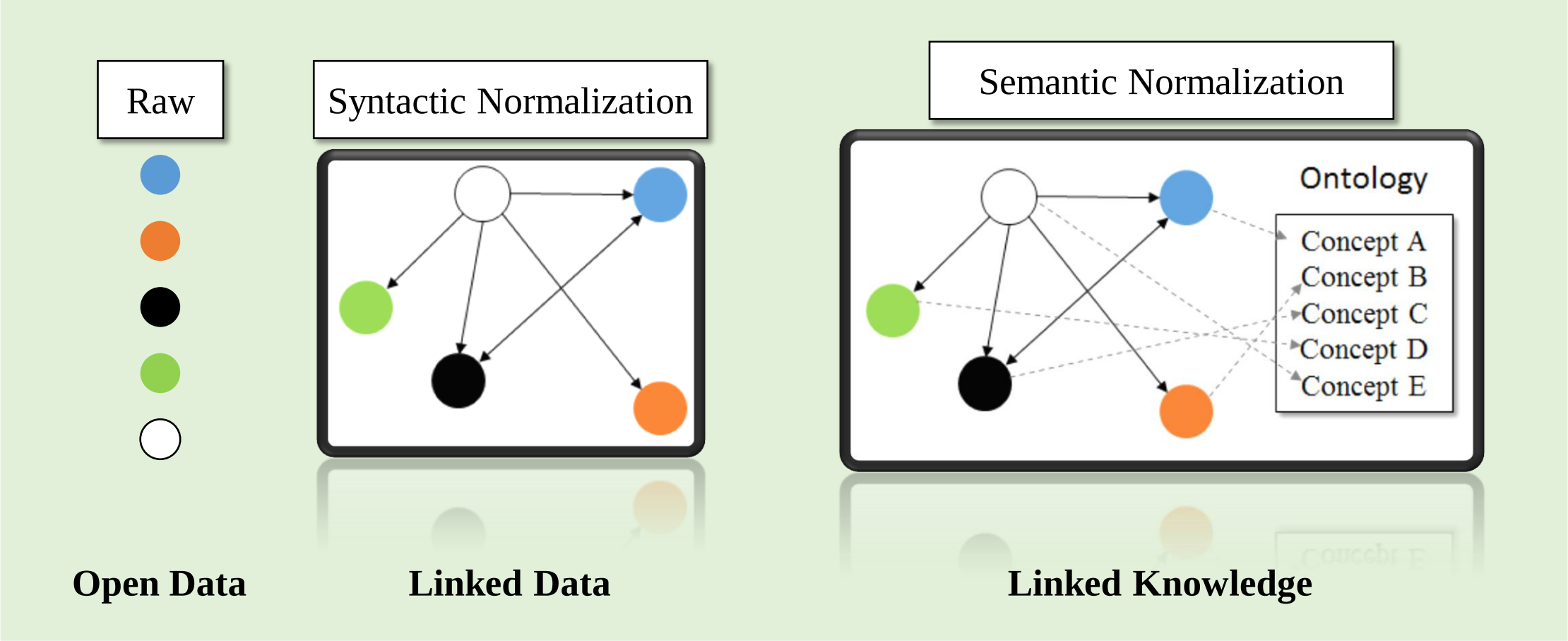
Transformation / RDFization



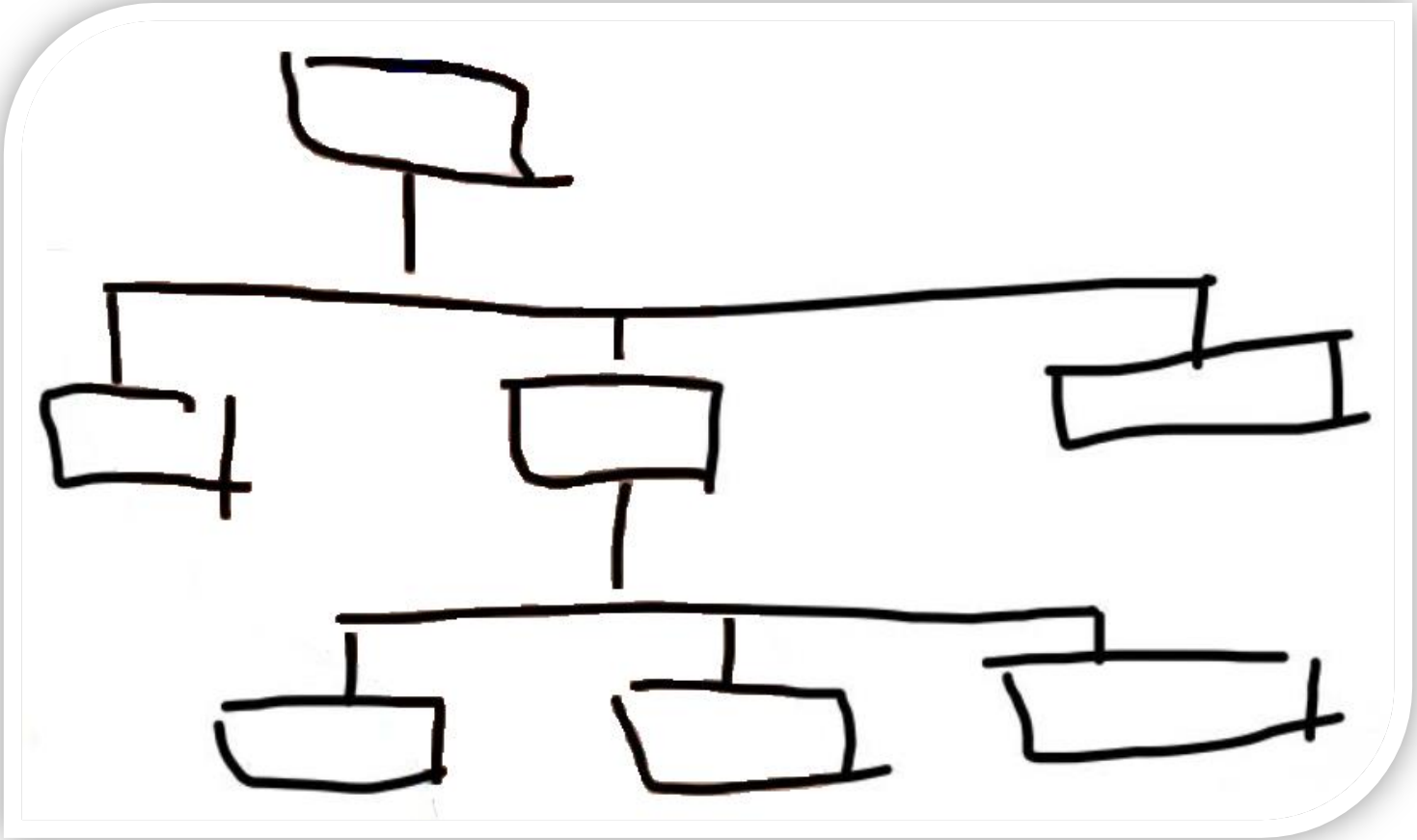
Convert to RDF Tools

Format	Tools
BibTex	• BibBase • bibtex2rdf • Java BibTex-To-RDF Converter
SQL	• D2RQ • OpenLink Virtuoso • R2RML
Excel	• XLWrap • Excel2rdf • Sheet2RDF
MARC	• easyM2R • marcmods2rdf
XML	• GRDDL • SPARQL2XQuery
Microformats	• Any23

RDF is only the First step, not End

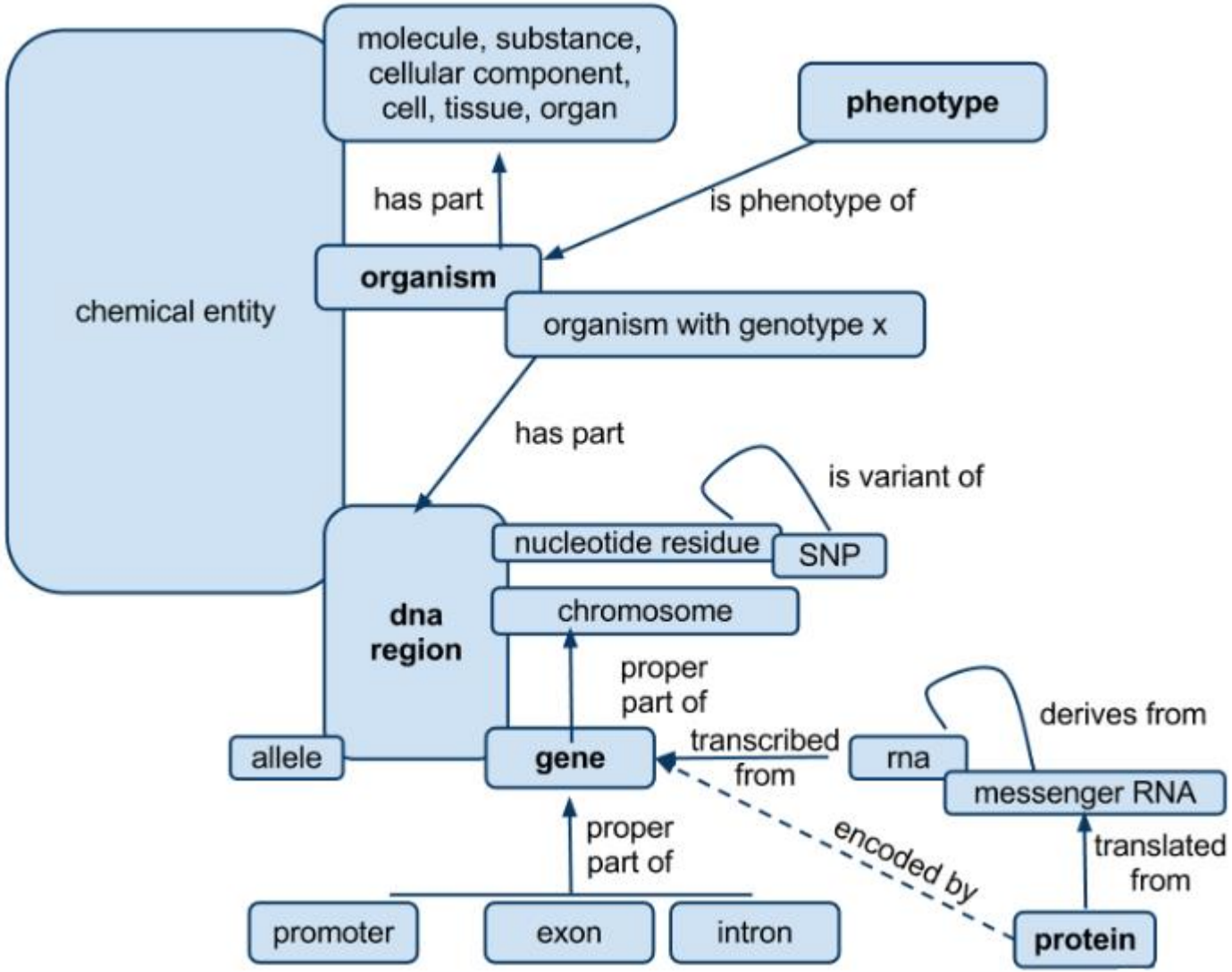


Ontology

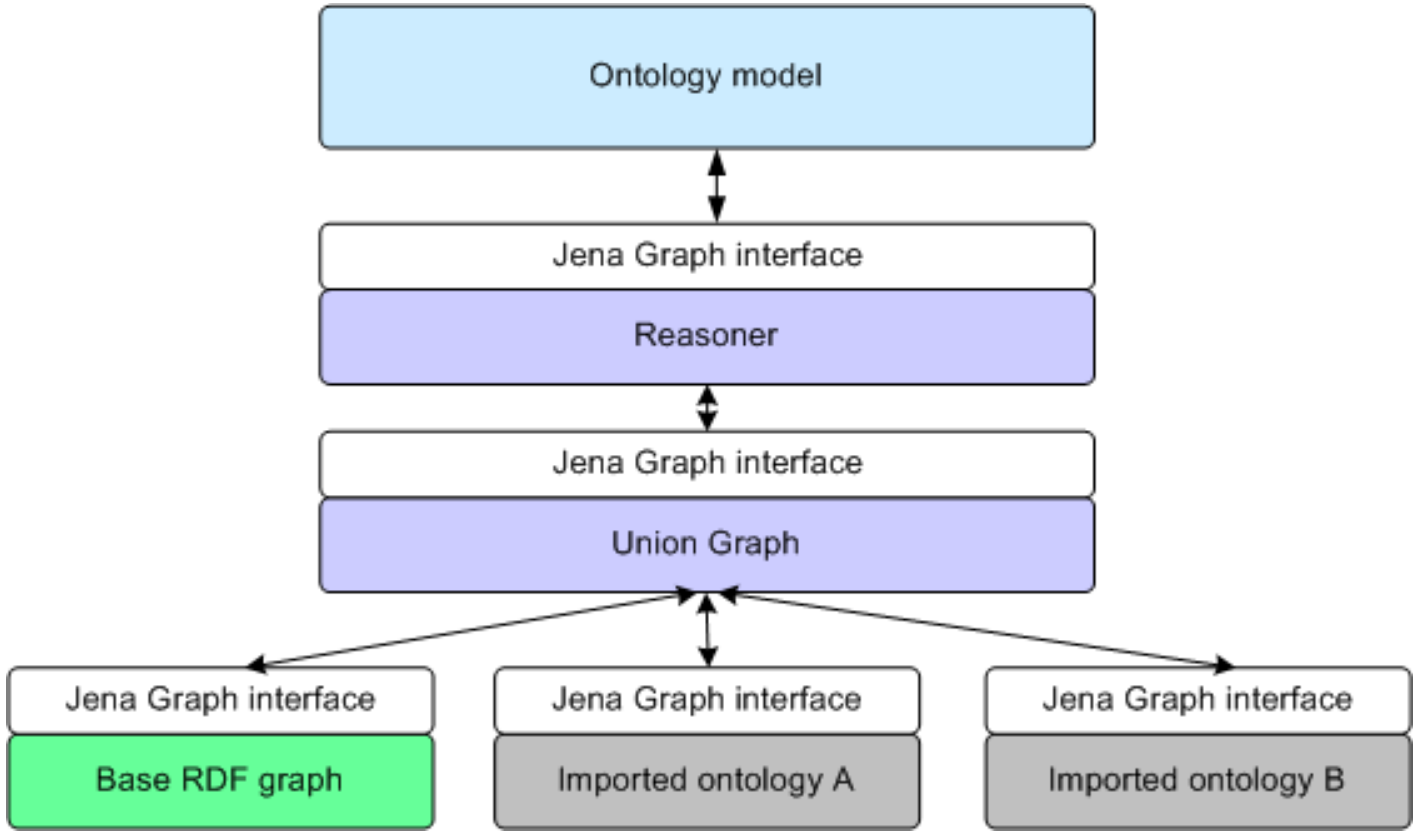


ontology as a strategy to formally represent and integrate knowledge

Ontology for the representations of diverse biomedical knowledge



Jena Ontology API



Ontology model compound document structure for imports

- Jena OntModel
- RDF Graph - SPARQL

Jena Inference API

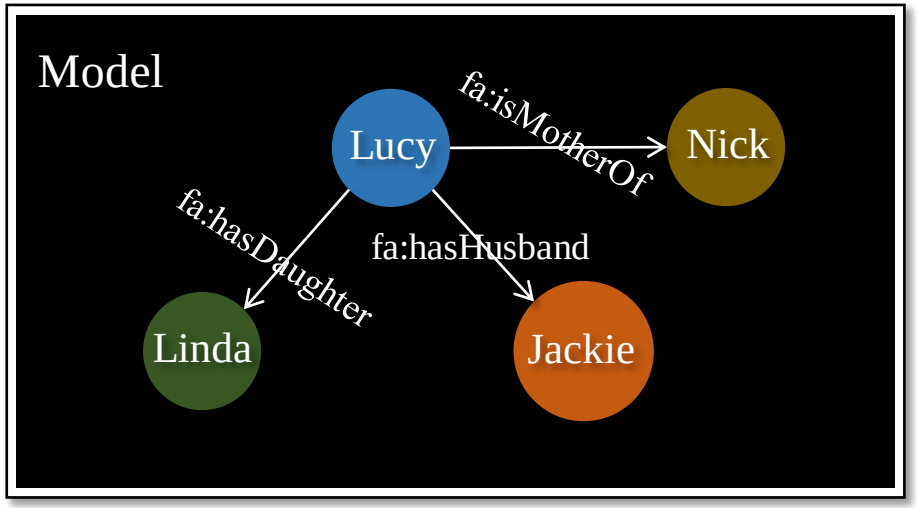
- **Transitive reasoner**: Provides support for storing and traversing class and property lattices. This implements just the *transitive* and *reflexive* properties of *rdfs:subPropertyOf* and *rdfs:subClassOf*.
- **RDFs rule reasoner**: Implements a configurable subset of the RDFs entailments.
- **OWL, OWL Mini, OWL Micro Reasoners**: A set of useful but incomplete implementation of the OWL/Lite subset of the OWL/Full language.
- **Generic rule reasoner**: A rule based reasoner that supports user defined rules. Forward chaining, tabled backward chaining and hybrid execution strategies are supported.

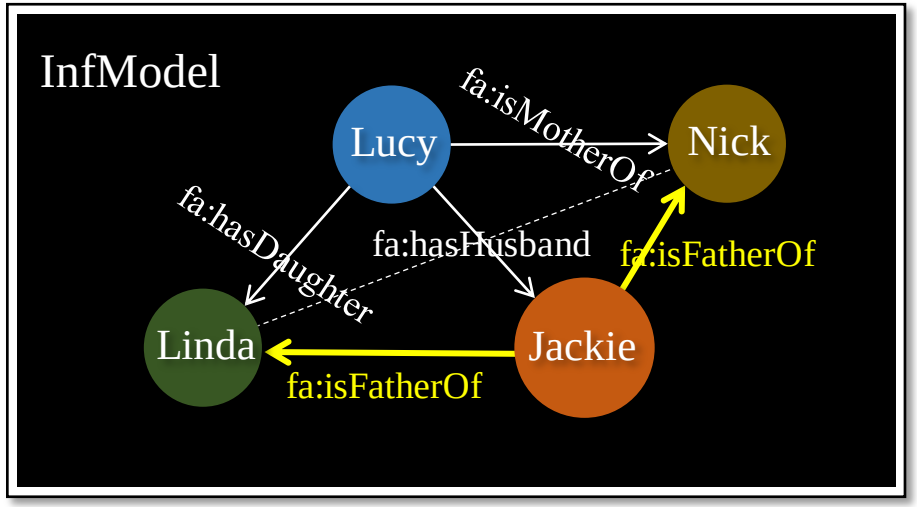
Jena Inference API

```
[rule1: (?a fa:hasHusband ?b)(?a fa:isMotherOf ?c)->(?b fa:isFatherOf ?c)]
[rule2: (?a fa:hasHusband ?b)(?a fa:hasDaughter ?c)->(?b fa:isFartherOf ?c)]
```

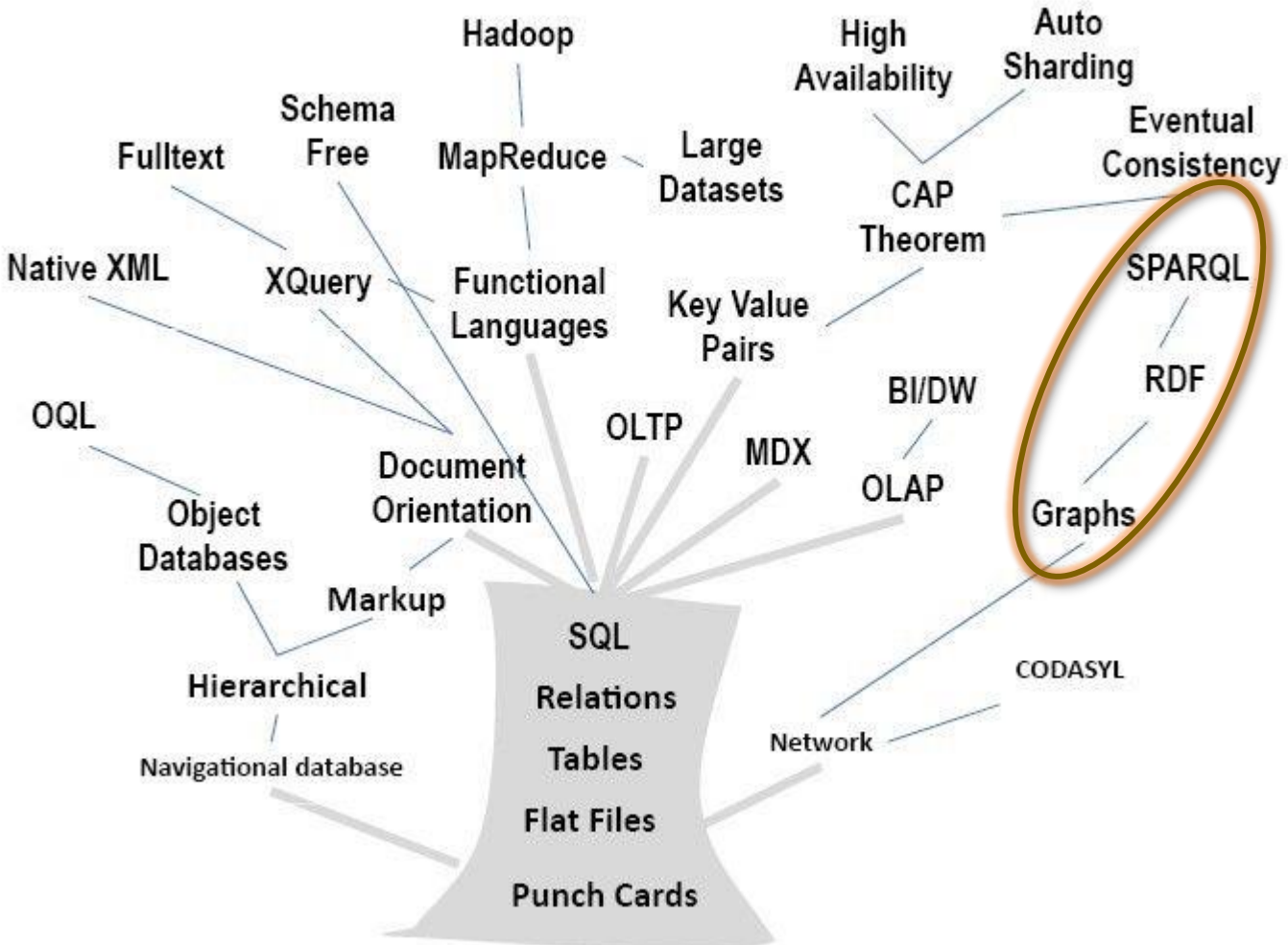
```
Reasoner reasoner = new GenericRuleReasoner(Rule.parseRules(rules));
InfModel inf = ModelFactory.createInfModel(reasoned, rawData);
```

Lucy fa:hasHusband Jackie ; fa:isMotherOf Nick ; fa:hasDaughter Linda.





A NoSQL Concept Tree



RDF Databases / Triple Store

- **Schema flexibility** - it's possible to do the equivalent of a schema change to an RDF store live, and without any downtime, or redesign .
- **More modern** - RDF stores are typically queried over HTTP, it's very easy to fit them into Service Architectures. Also they handle internationalized content better than typical SQL databases.
- **Standardization** - the level of standardization of implementations using RDF and SPARQL is much higher than SQL. Moving data between stores is easy, as they all speak the same language.
- **Expressivity** - it's much easier to model complex data in RDF than in SQL, and the query language makes it easier to do things like LEFT JOINs (called OPTIONAL in SPARQL).
- **Provenance** - SPARQL lets you track where each piece of information came from, and you can store metadata about it, letting you easily do sophisticated queries.

RDF Databases / Triple Store

Databases	Customers
OpenLink Virtuoso	DBpedia, BIO2RDF, LOD2
AllegroGraph	Public Health Agency of Canada, Pfizer, NOVARTIS
Stardog	NASA, BEST BUY, inQuest TECHNOLOGIES
Oracle NoSQL	eDBA, eTouch, opalsoft

Others: Jena TDB, BigData, Sesame, 4store, MarcLogic, YARS2

How to access Linked Data Sets?

- Downloads
- Web interface (lookup + services)
- SPARQL endpoint
- Flint Sparql editor
- Virtuoso Faceted Browser

RDF Serialization

```
model.write(output, "format") ;
```

Jena writer name	RIOT RDFFormat
"TURTLE"	TURTLE
"TTL"	TURTLE
"Turtle"	TURTLE
"N-TRIPLES"	NTRIPLES
"N-TRIPLE"	NTRIPLES
"NT"	NTRIPLES
"RDF/XML-ABBREV"	RDFXML
"RDF/XML"	RDFXML_PLAIN
"N3"	N3
"JSON-LD"	JSONLD
"RDF/JSON"	RDFJSON

Semantic Search

Semantic search provides more meaningful search results by evaluating and understanding the search phrase and finding the most relevant results in a website, database or any other data repository.

One of the primary goals of semantic web:

- Not only keyword full-text search
- Query includes relations between resources
- Connecting data: mash-up from different sources

Get the phenotypes of knock-out mouse models for the targets of a selected drug

data sources: DrugBank, HGNC, MGI, BioPortal

Drug: Imatinib

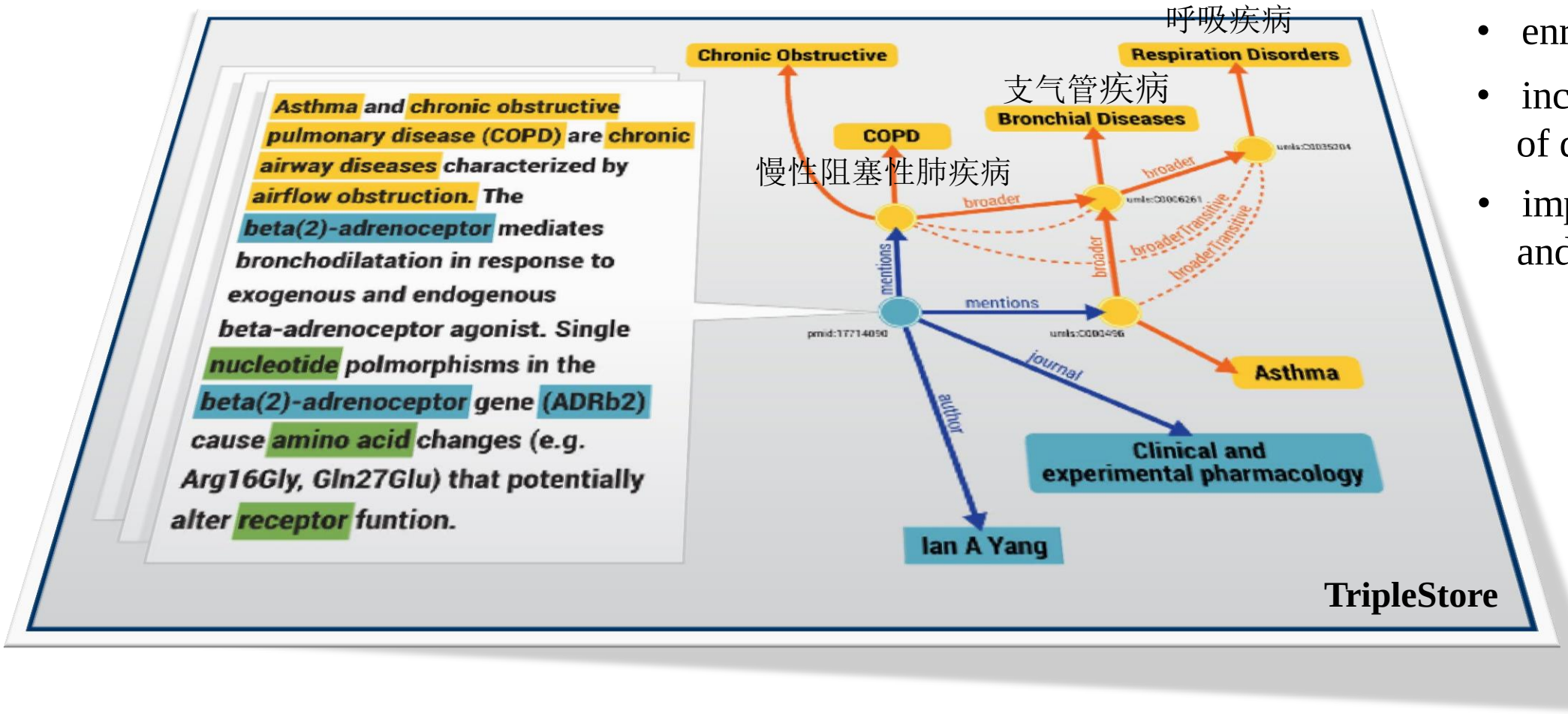
```
PREFIX dct: <http://purl.org/dc/terms/>
SELECT DISTINCT ?phenotype_label
WHERE {
 SERVICE <http://drugbank.bio2rdf.org/sparql> {
  ?drug <http://bio2rdf.org/drugbank_vocabulary:target> ?target .
  FILTER(?drug = <http://bio2rdf.org/drugbank:DB00619>)
  ?target <http://bio2rdf.org/drugbank_vocabulary:x-hgnc> ?hgnc .
 }
 SERVICE <http://hgnc.bio2rdf.org/sparql> {
  ?hgnc <http://bio2rdf.org/hgnc_vocabulary:x-mgi> ?marker .
 }
 SERVICE <http://mgi.bio2rdf.org/sparql> {
  ?model <http://bio2rdf.org/mgi_vocabulary:marker> ?marker .
  ?model <http://bio2rdf.org/mgi_vocabulary:allele> ?all .
  ?all <http://bio2rdf.org/mgi_vocabulary:allele-attribute> ?allele_type .
  ?model <http://bio2rdf.org/mgi_vocabulary:phenotype> ?phenotypes .
  FILTER (str(?allele_type) = "Null/knockout")
 }
 SERVICE <http://bioportal.bio2rdf.org/sparql> {
  ?phenotypes rdfs:label ?phenotype_label .
 }
}
```

Results:

"hemorrhage [mp:0001914]"@en
"intracranial hemorrhage [mp:0001915]"@en
"perinatal lethality [mp:0002081]"@en
"hydrops fetalis [mp:0002192]"@en
"kidney hemorrhage [mp:0000533]"@en
"skin hemorrhage [mp:0011514]"@en
"anemia [mp:0001577]"@en

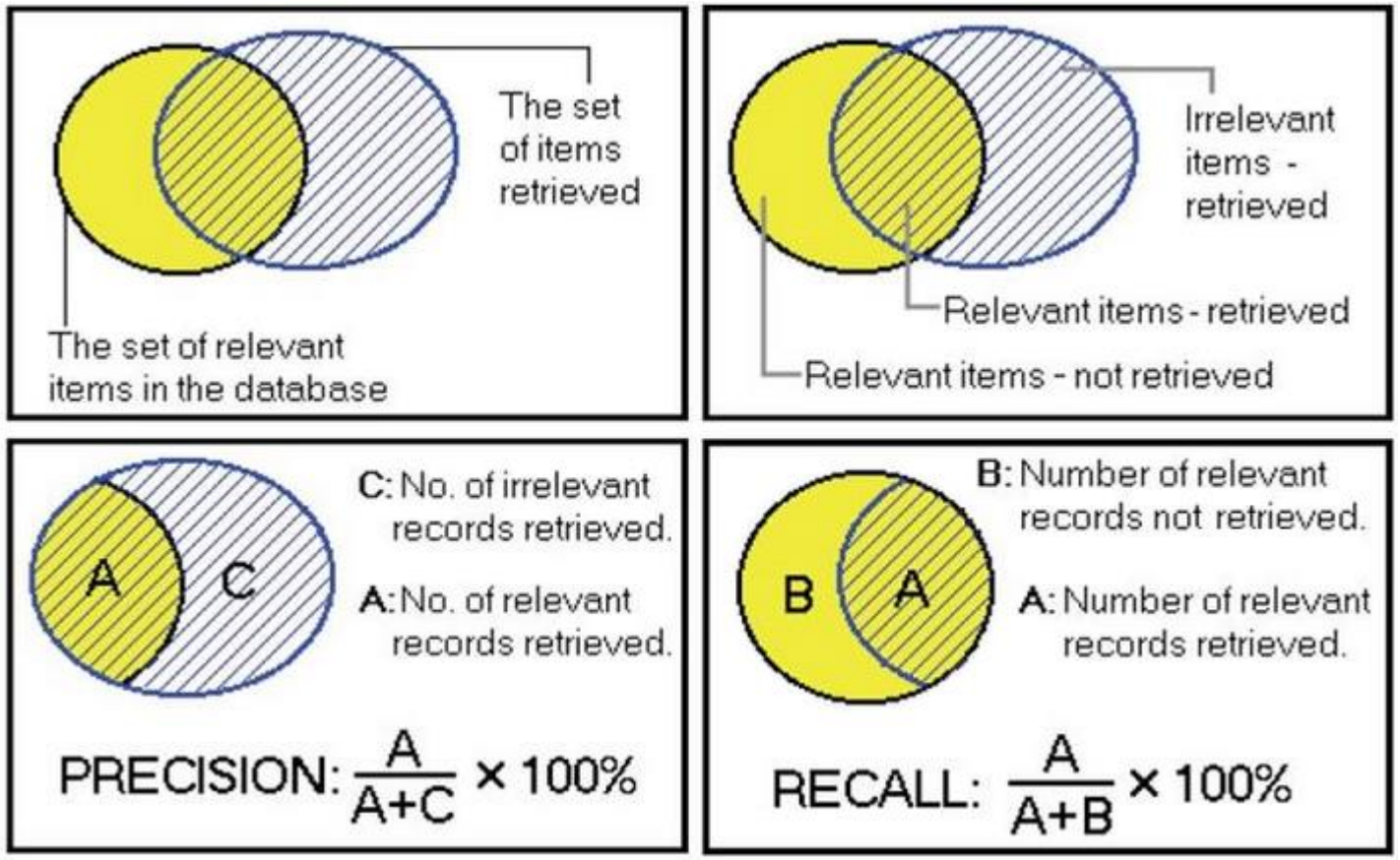
"出血"
"颅内出血"
"围产期的杀伤力"
"胎儿水肿"
"肾出血"
"皮肤出血"
"贫血"

Semantic Search



- enriched content
- increased wealth of data
- improved precision and recall

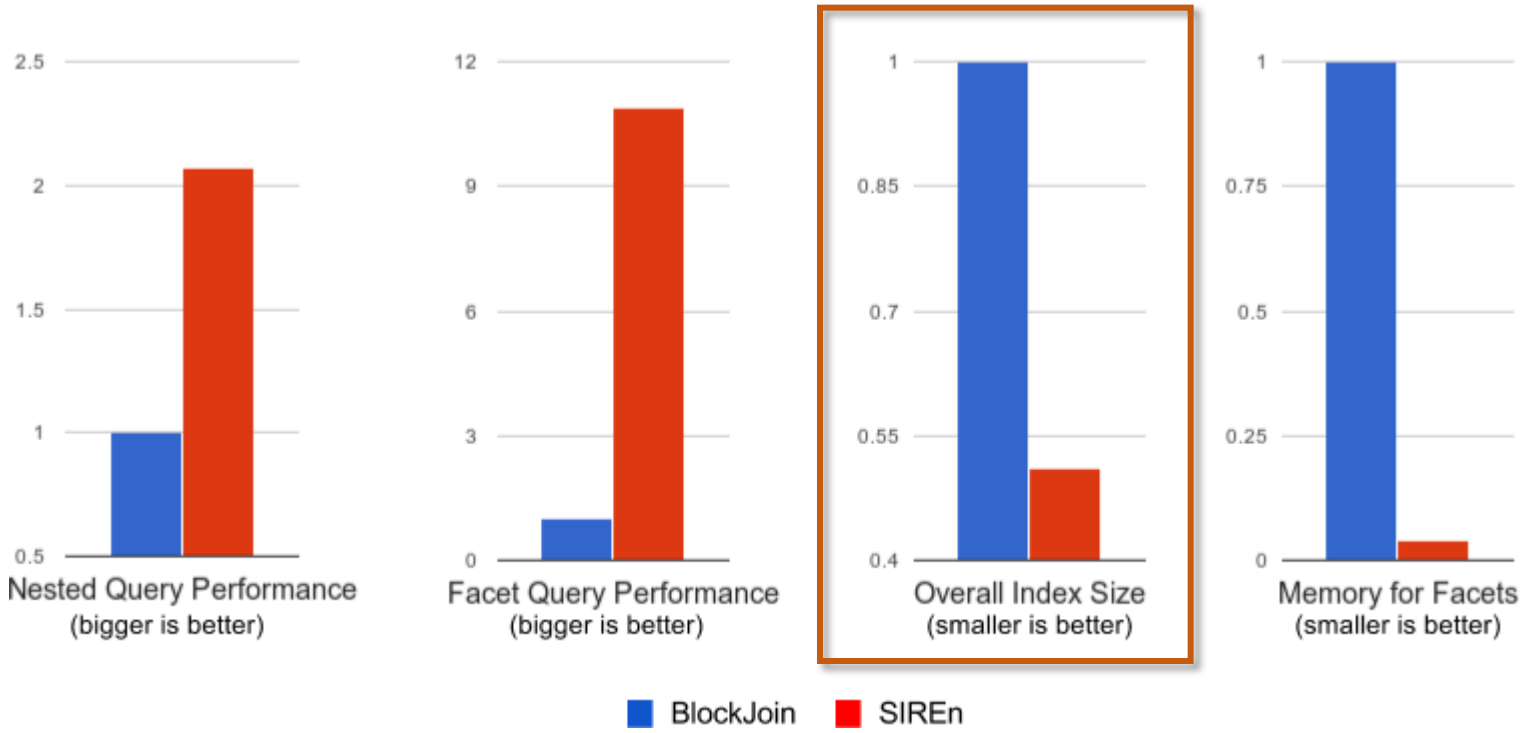
Precision & Recall



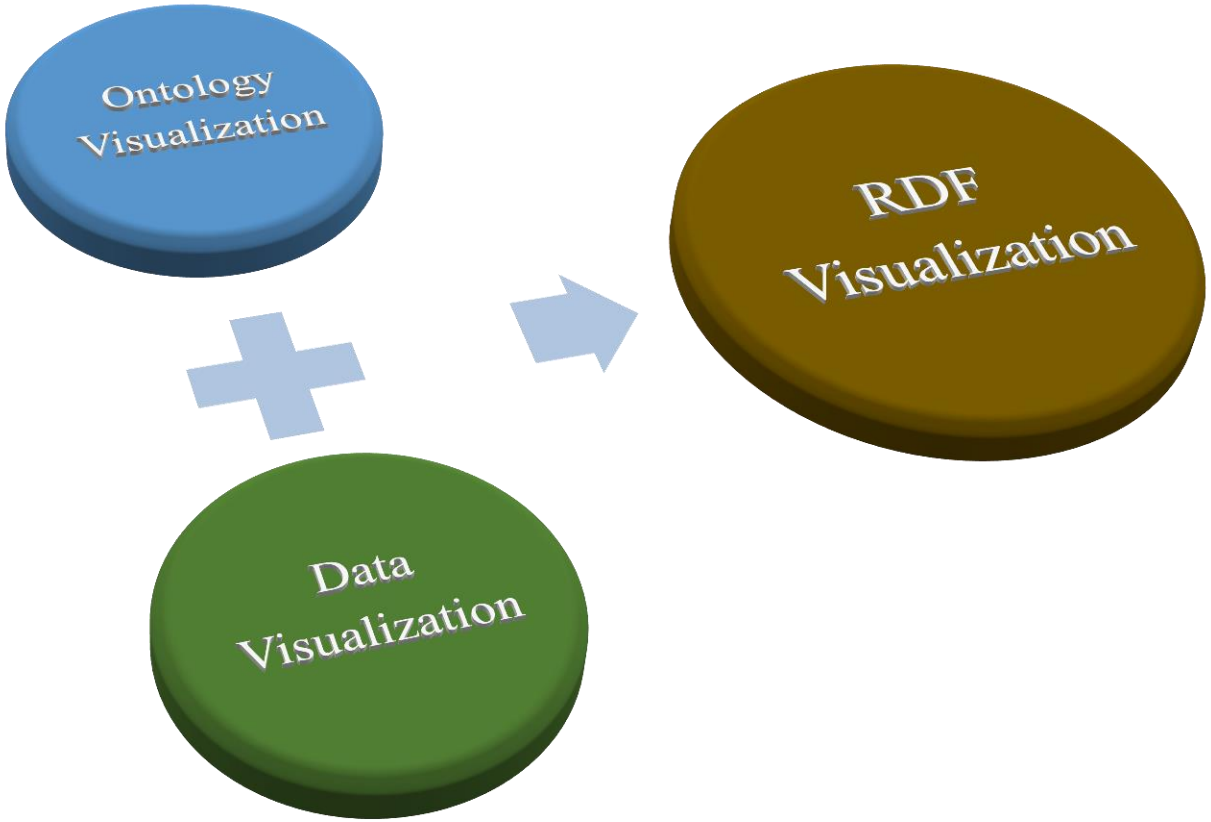
SIREn {Semantic Information Retrieval Engine}

SIREn is a highly scalable open-source full-text search engine especially suited for **nested** and **schemaless** data.

Lucence's BlockJoin: 100*100 = 10,000 docs
SIREn: 100 = 100 docs



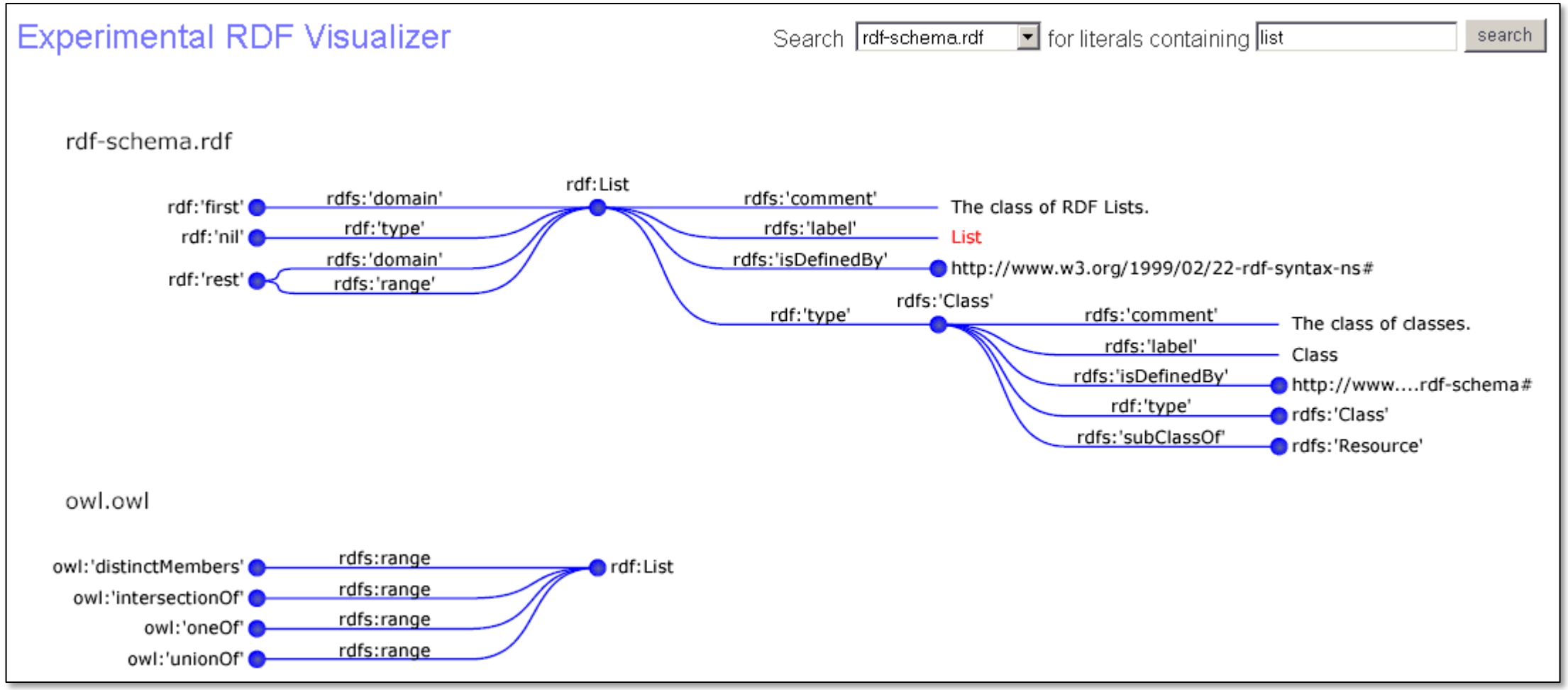
Visual Tools – RDF Graph



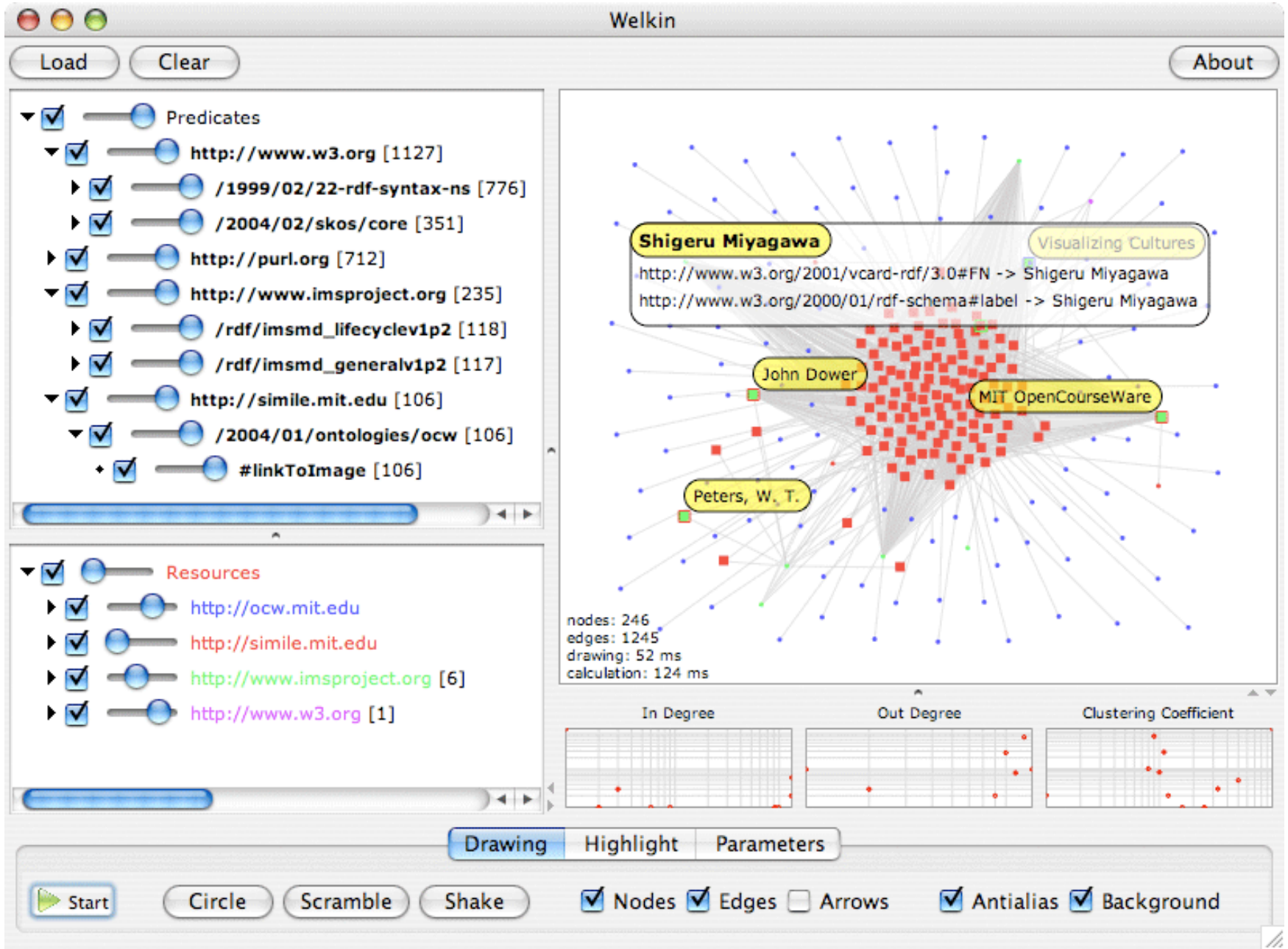
RDF/OWL Ontology Visualization

- **HP RDF Graph Visualization** describes an experimental node-centric approach to RDF
- **RDF Gravity** is a tool for visualizing RDF/OWL Graphs/ontologies
- **Welkin** is a graph-based RDF visualizer
- **RelFinder** extracts and visualizes relationships between given objects in RDF data and makes these relationships interactively explorable

HP RDF Graph Visualization

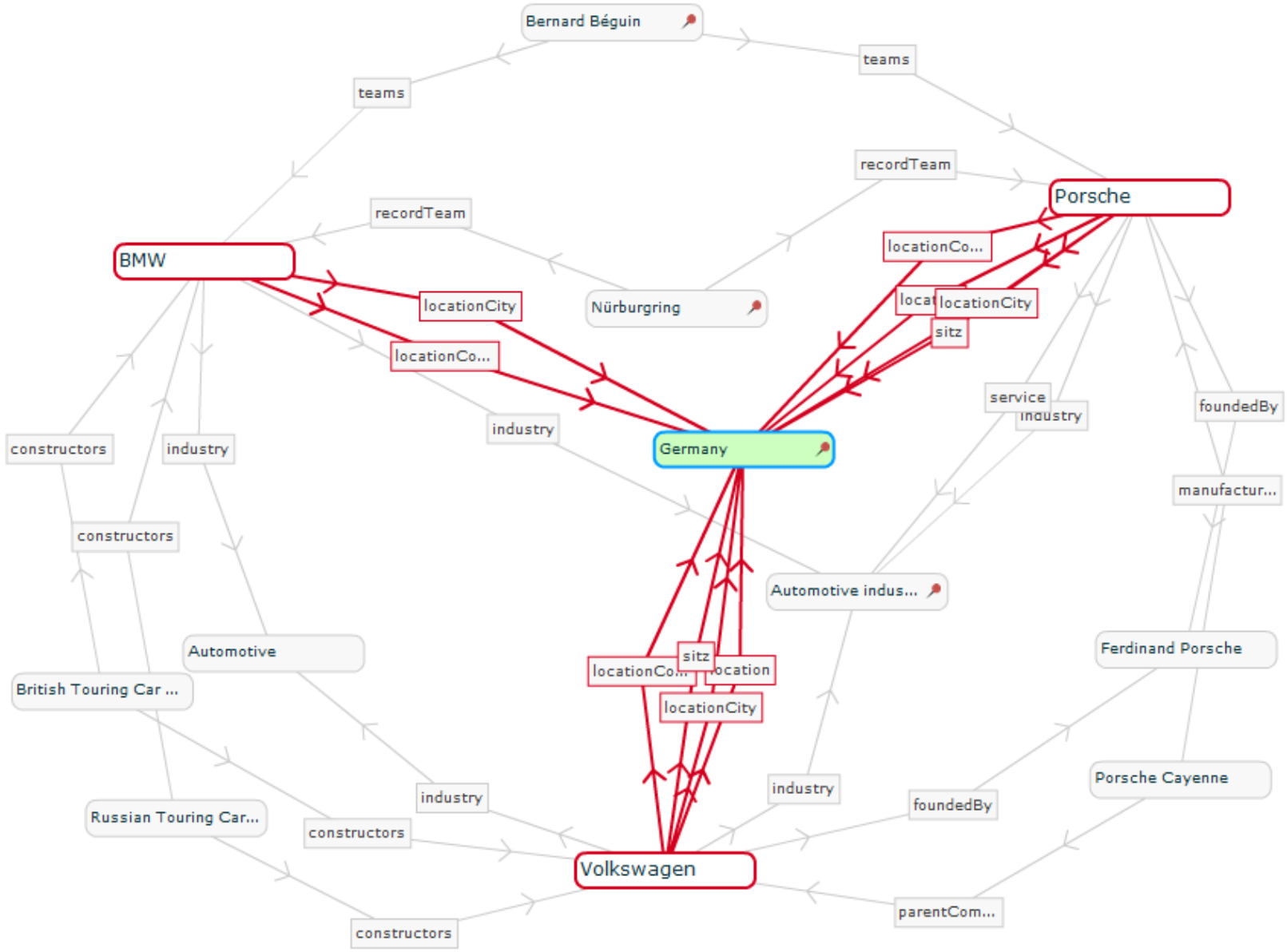


Welkin


Welkin
Load
Clear
About
Predicates
http://www.w3.org [1127]
/1999/02/22-rdf-syntax-ns [776]
/2004/02/skos/core [351]
http://purl.org [712]
http://www.imsproject.org [235]
/rdf/imsmd_lifecyclev1p2 [118]
/rdf/imsmd_generalv1p2 [117]
http://simile.mit.edu [106]
/2004/01/ontologies/ocw [106]
#linkToImage [106]
Resources
http://ocw.mit.edu
http://simile.mit.edu
http://www.imsproject.org [6]
http://www.w3.org [1]
Shigeru Miyagawa
Visualizing Cultures
http://www.w3.org/2001/vcard-rdf/3.0#FN -> Shigeru Miyagawa
http://www.w3.org/2000/01/rdf-schema#label -> Shigeru Miyagawa
John Dower
MIT OpenCourseWare
Peters, W. T.
nodes: 246
edges: 1245
drawing: 52 ms
calculation: 124 ms
In Degree
Out Degree
Clustering Coefficient
Drawing
Highlight
Parameters
Start
Circle
Scramble
Shake
Nodes
Edges
Arrows
Antialias
Background

RelFinder


RelFinder
URL
between
examples
(1) BMW
(2) Porsche
(3) Volkswagen
add
clear
Find Relations
Filter by:
relations: (31/129)
length
class
link
conne...
number of objects
num
vi
1
31/34
2
0/95
Germany
en
More Infos:
dbpedia.org
www.bundesregierung.de
Germany, officially the Federal Republic of


Bernard Béguin
teams
teams
recordTeam
recordTeam
Porsche
BMW
locationCo...
locationCity
Nürburgring
locat
locationCity
sitz
locationCo...
service
industry
foundedBy
industry
constructors
industry
Germany
manufactur...
constructors
Automotive indus...
Automotive
sitz
locationCo...
ocation
Ferdinand Porsche
British Touring Car ...
locationCity
industry
industry
Porsche Cayenne
Russian Touring Car...
constructors
foundedBy
Volkswagen
parentCom...
constructors

RDF Data Visualization

- **RDF-spark** *is a graph-based RDF visualizer*
- **Sgvizler** *provides a javascript based SPARQL query form that passes the SPARQL result set to a powerful visualizer*
- **Visual Browser** *is a Java application that can visualize the data in RDF scheme*
- **Map4rdf** *provides your users with a nice map-based visualization of your data*

Sgvizler

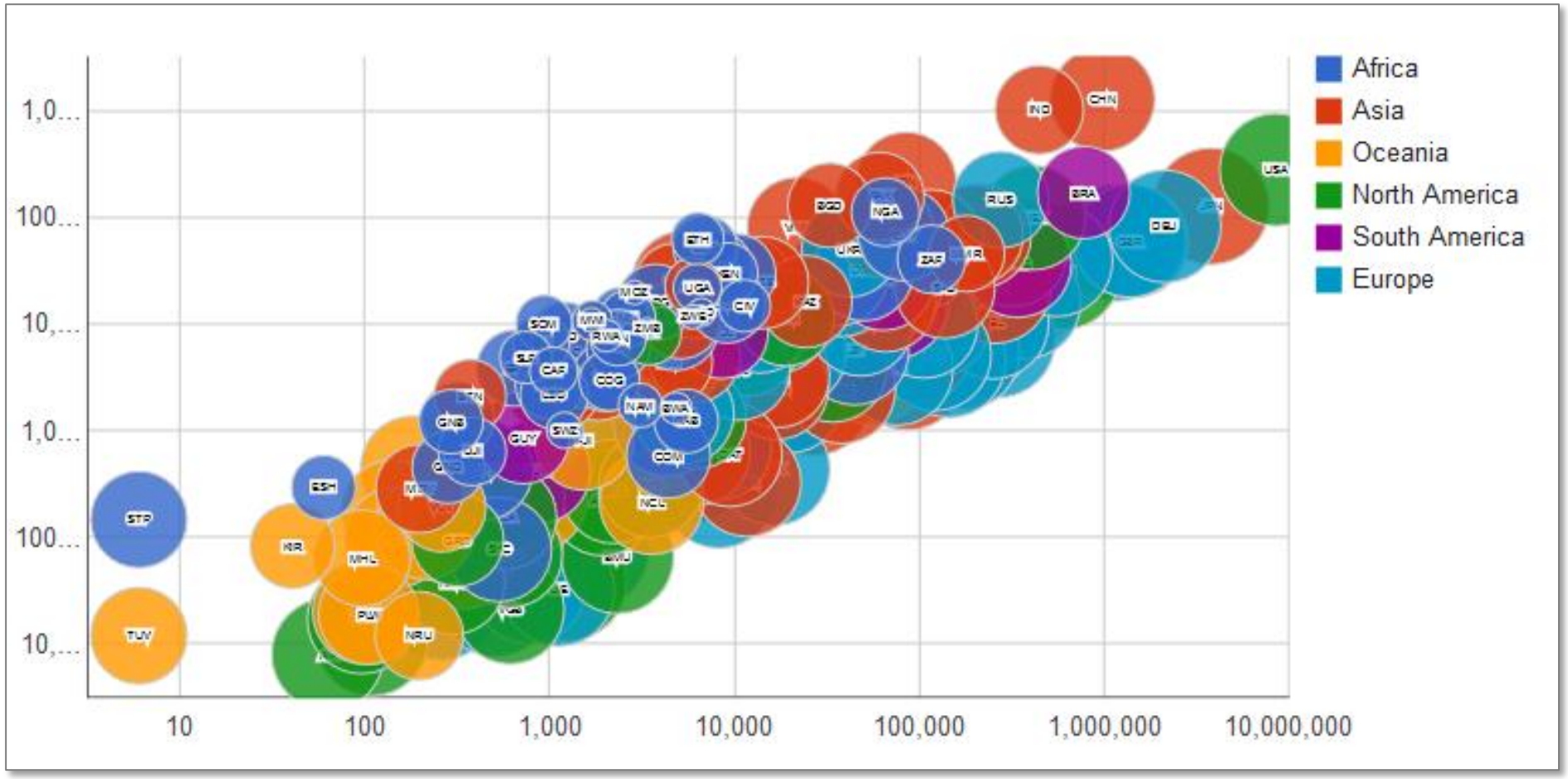
- Sgvizler is a javascript which renders the result of SPARQL SELECT queries into charts or html elements.

```
<div id="sgvzl_example_query"
   data-sgvizler-endpoint="http://sws.ifi.uio.no/sparql/world"
   data-sgvizler-query="SELECT * WHERE{[] w:hasCountryCode ?ID ; w:hasGNP ?GNP ; w:hasCountryPopulation ?Population ;
w:isCountryInContinent [ rdfs:label ?Continent ] ; w:hasLifeExpectancy ?LifeExpectancy ; }"
   data-sgvizler-chart="google.visualization.BubbleChart"
   data-sgvizler-chart-options="vAxis.logScale=true|hAxis.logScale=true|bubble.textStyle.fontSize=6"
   data-sgvizler-loglevel="2"
   style="width:800px; height:400px;"/>
```

```
SELECT * WHERE {
   [] w:hasCountryCode ?ID ;
      w:hasGNP ?GNP;
      w:hasCountryPopulation ?Population;
      w:isCountryInContinent [ rdfs:label ?Continent ];
      w:hasLifeExpectancy ?LifeExpectancy .
}
```

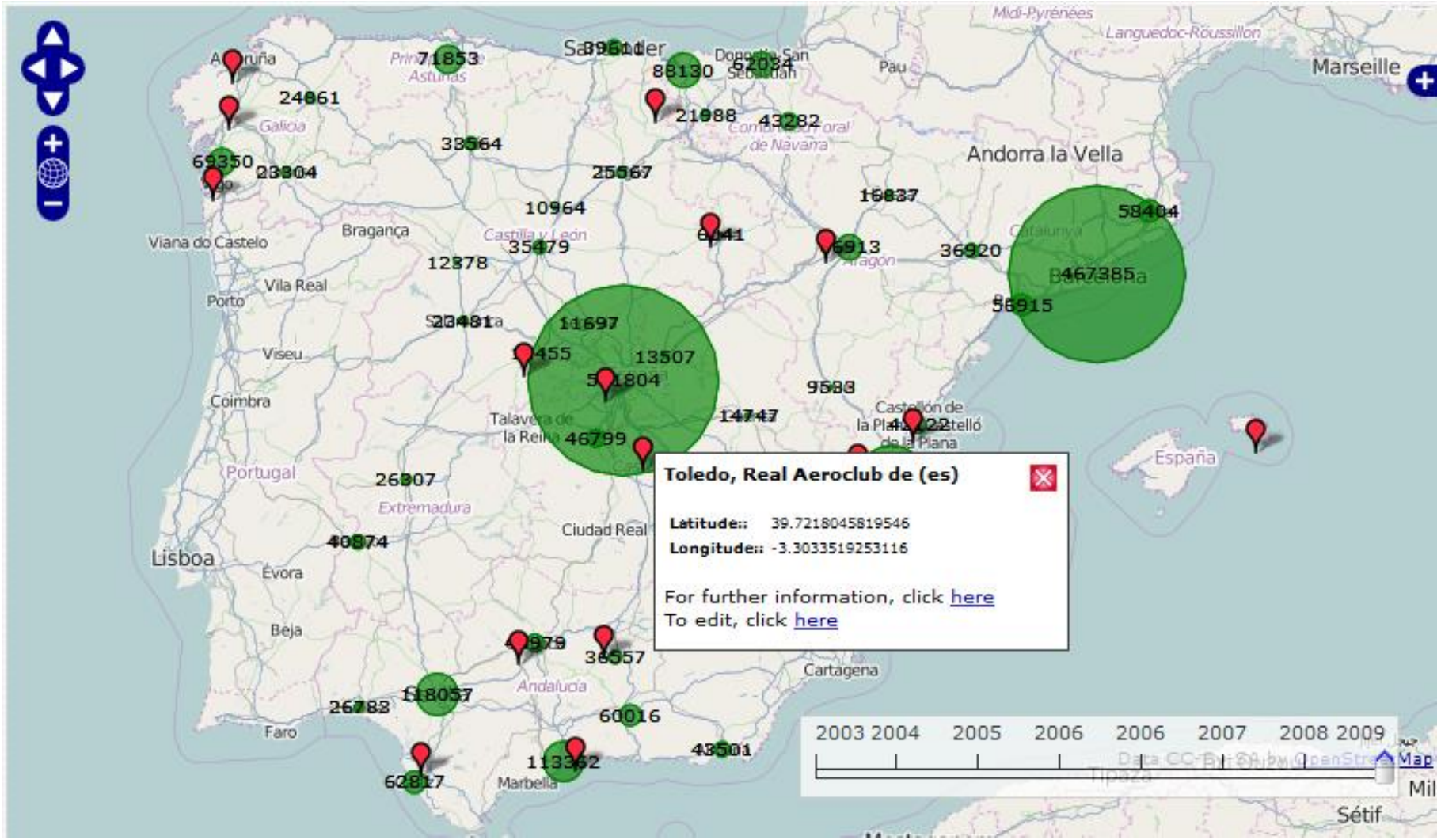
Sgvizler

Result



Map4Rdf

geo.linkeddata.es



Map4Rdf

Toledo, Real Aeroclub de at geo.linkeddata.es

http://geo.linkeddata.es/resource/Aeroclub/Toledo%2C%20Real%20Aeroclub%20de

Property	Value
geo:geometry	■ <http://geo.linkeddata.es/resource/wgs84/39.7218045819546_-3.3033519253116>
rdfs:label	■ Toledo, Real Aeroclub de (es)
rdf:type	■ geoes:Aeroclub

Metadata

Anon_0

rdf:type	<http://www.w3.org/2004/03/trix/rdfg-1/Graph>
foaf:primaryTopic	<http://geo.linkeddata.es/resource/Aeroclub/Toledo%2C%20Real%20Aeroclub%20de>
dcterms:creator	<http://geo.linkeddata.es/resource/Organizaci%C3%B3n/InstitutoGeogr%C3%A1ficoNacionalDeEspa%C3%B1a>
dcterms:publisher	<http://geo.linkeddata.es>
dc:rights	La información geográfica digital comprendida en el Equipamiento Geográfico de Referencia Nacional (artículo 1.1 de la Orden FOM/956/2008) así como los Metadatos de los datos geográficos y servicios del IGN-CNIG, no requieren la aceptación de licencia y su uso será, en cualquier caso, libre y gratuito, siempre que se mencione al Instituto Geográfico Nacional como propietario de los datos.
dcterms:spatial	<http://geo.linkeddata.es/resource/Pa%C3%ADs/Espa%C3%B1a>
prv:createdBy	Anon_1 (more)

expand all

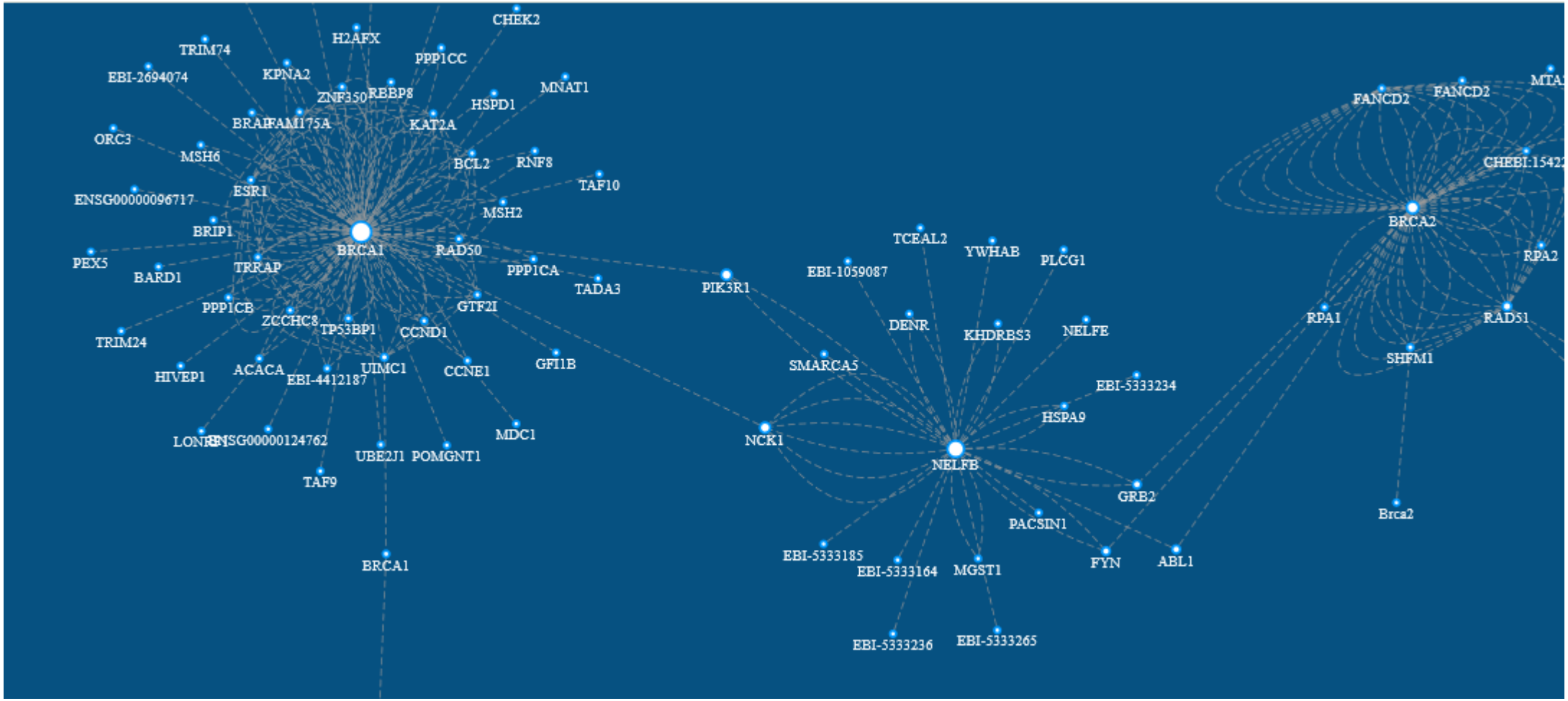
This page shows information obtained from the SPARQL endpoint at http://geo.linkeddata.es/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Visual Tools

- **D3**
- **Simile Exhibit**
- **Echarts**
- **Google Visualization API**
- **Data Publica**
- **GeoAPI**
- **Cytoscape**

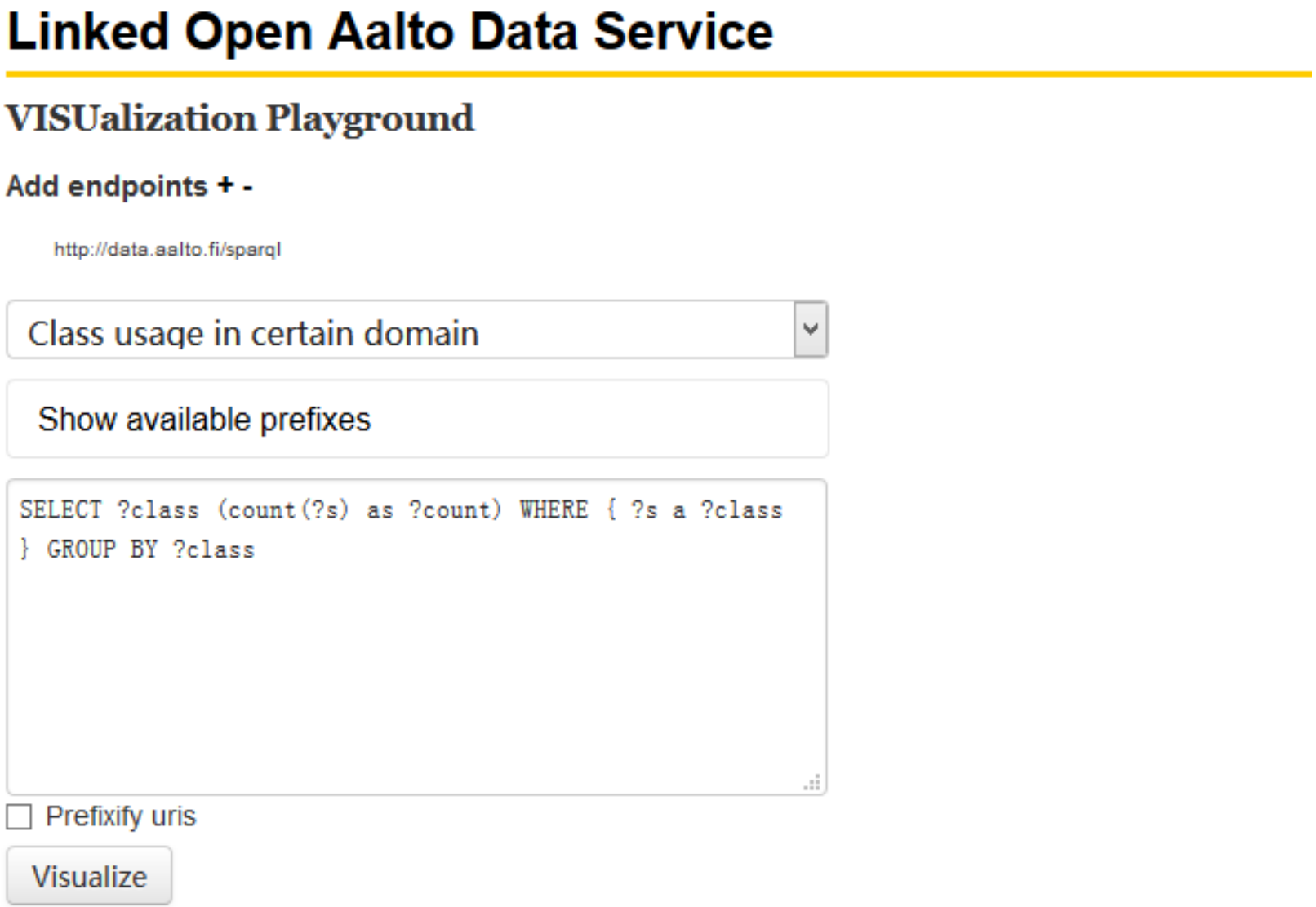
Cytoscape

An open source software platform for visualizing complex networks and integrating these with any type of attribute data.



BIO-Linked Scientific Data

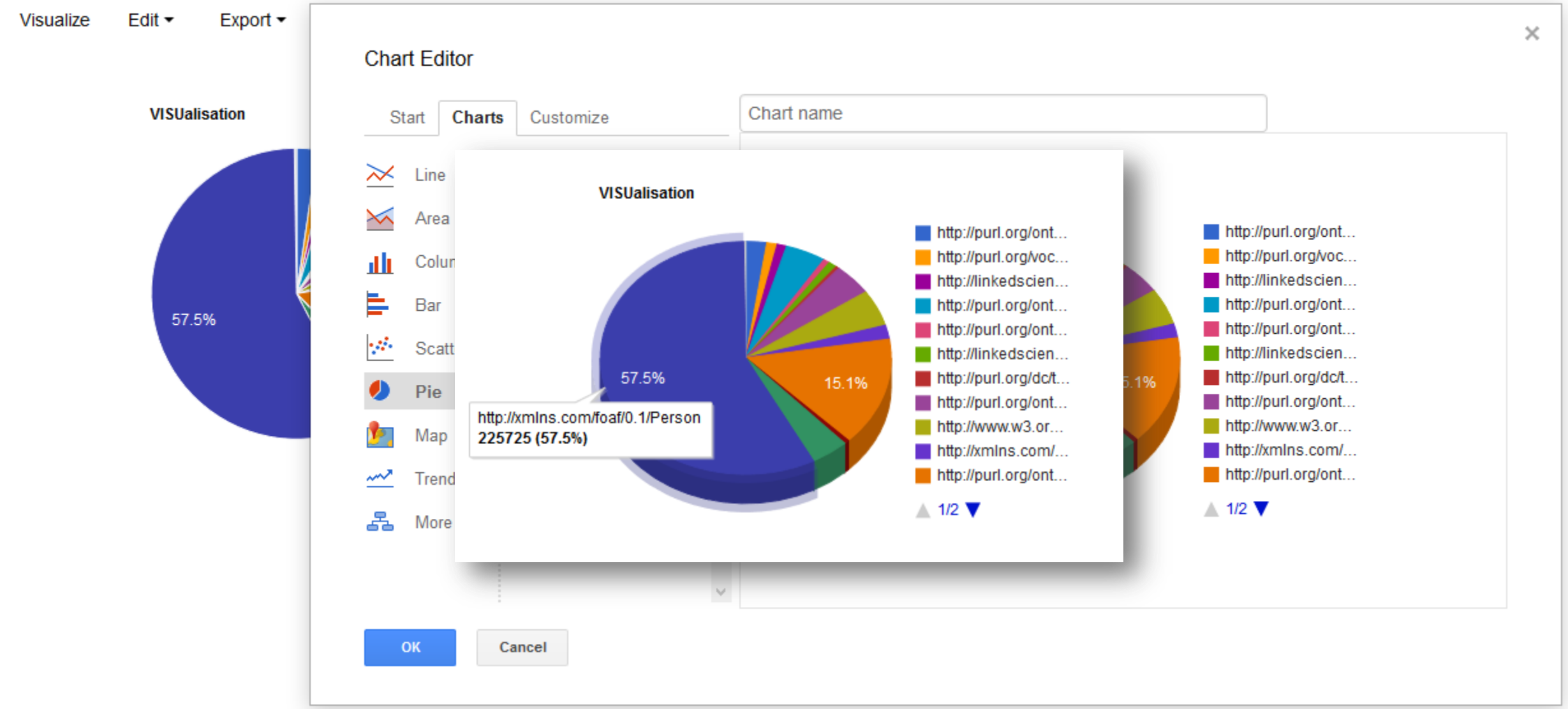
VISUalization Playground



Aalto University

Linked Open Aalto Data Service

VISUalization Playground

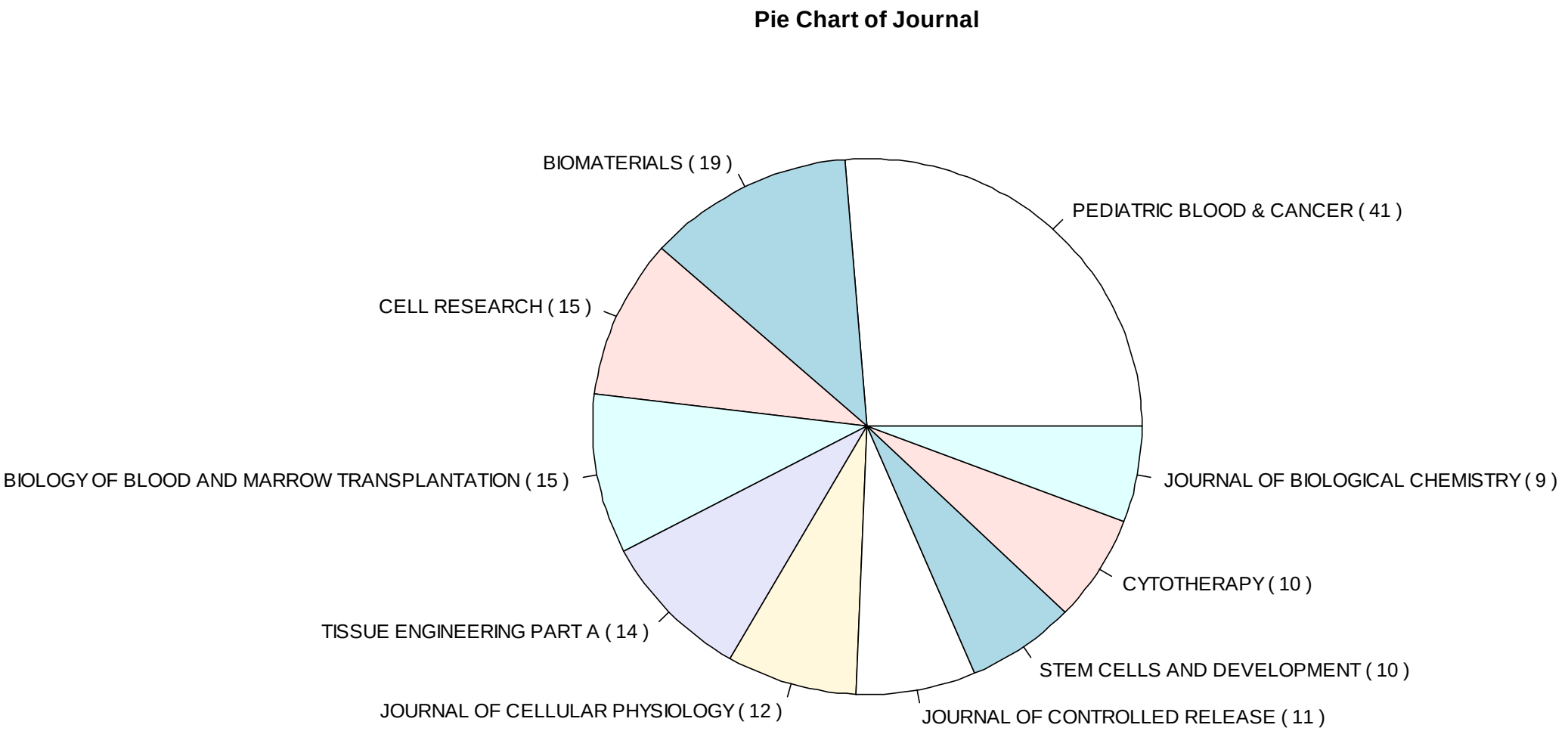


R script

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques.

```
library(SPARQL)
library(igraph)
library(network)
library(ergm)

# Live DBpedia endpoint
endpoint <- 'http://localhost:8890/sparql'
options <- NULL

prefix <- c("db","http://dbpedia.org/resource/")

sparql_prefix <- "PREFIX bibo: <http://purl.org/ontology/bibo/>
                  PREFIX dc: <http://purl.org/dc/elements/1.1/>
                  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
"

q <- paste(sparql_prefix,
           'select ?journal (count(distinct ?s) as ?count)
            where {graph <http://www.cba.ac.cn/stemcell/graph/article> {
              ?s dc:title ?title ; bibo:presentedAt ?journal .}
            } order by desc(?count) limit 10')

res <- SPARQL(url=endpoint,query=q,ns=prefix)$results
# iconv(res,"utf-8","gbk")

slices <- res$count
lbls <- res$journal
lals <- paste(lbls, slices)
pie(slices,labels=lbls,main="Pie Chart of Journal")
```

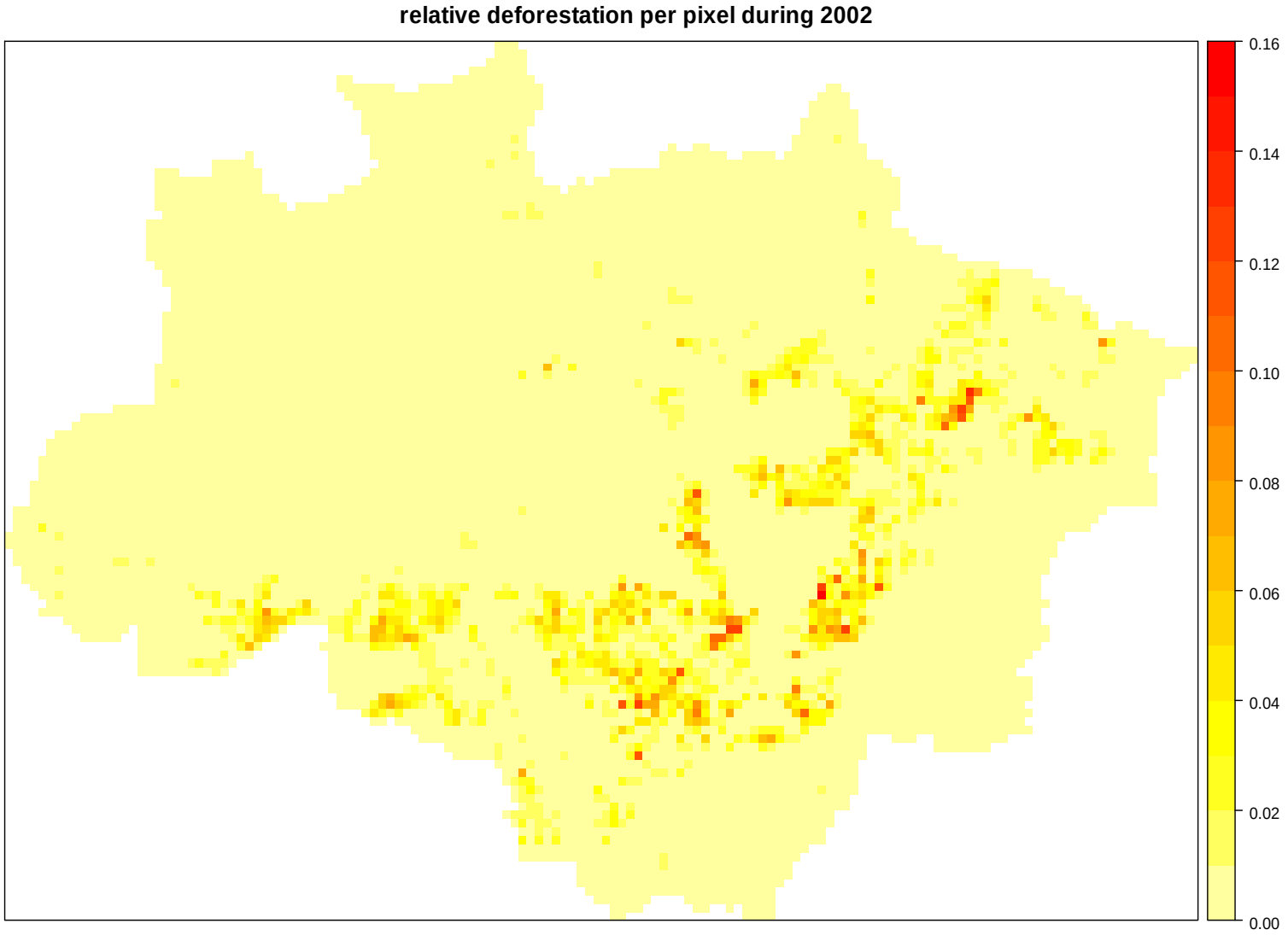
R - Pie Chart

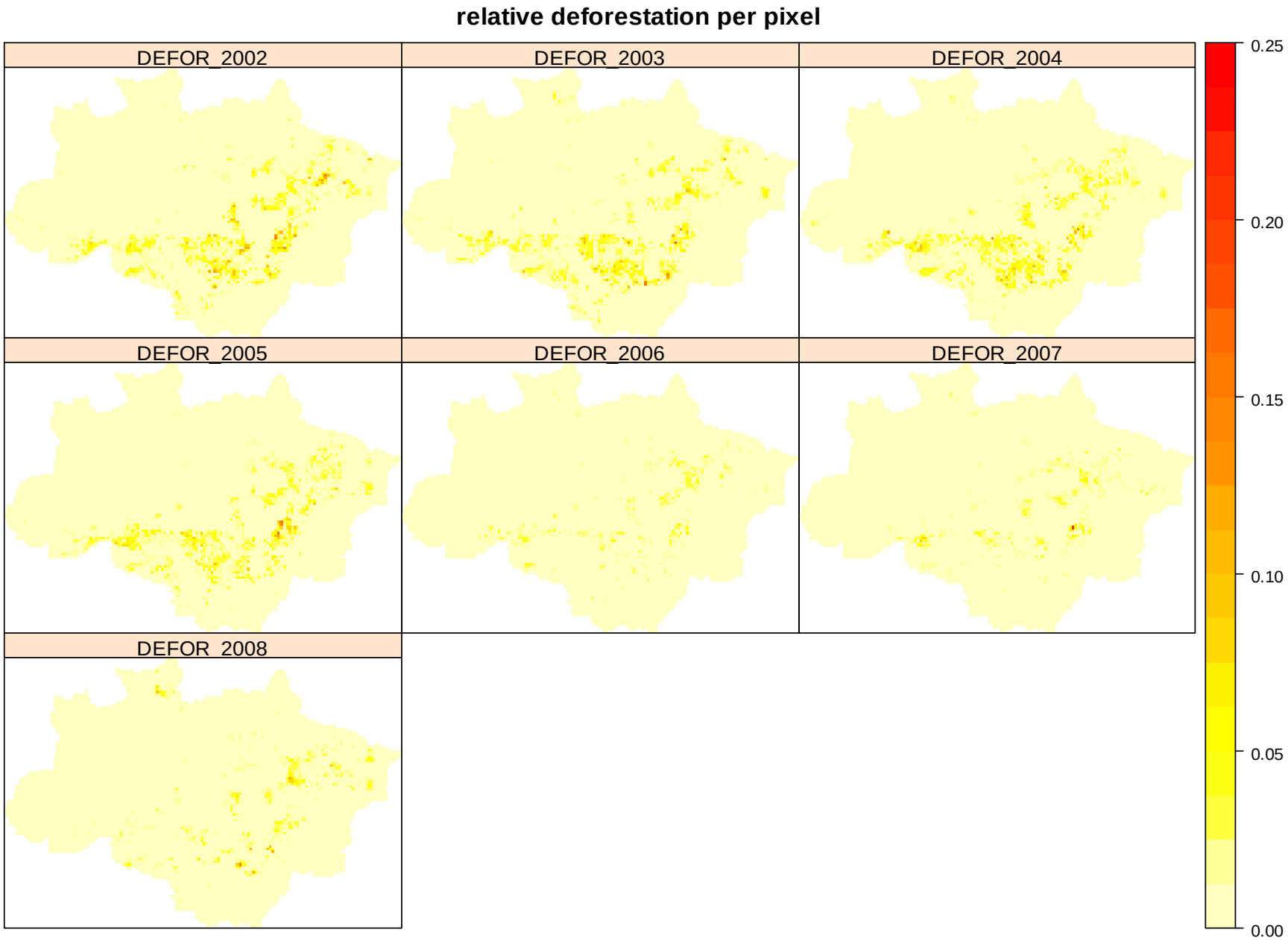

Pie Chart of Journal
BIOMATERIALS (19)
PEDIATRIC BLOOD & CANCER (41)
CELL RESEARCH (15)
BIOLOGY OF BLOOD AND MARROW TRANSPLANTATION (15)
JOURNAL OF BIOLOGICAL CHEMISTRY (9)
CYTOTHERAPY (10)
TISSUE ENGINEERING PART A (14)
STEM CELLS AND DEVELOPMENT (10)
JOURNAL OF CELLULAR PHYSIOLOGY (12)
JOURNAL OF CONTROLLED RELEASE (11)

R - Brazilian Amazon Rainforest

```
library(SPARQL)
library(sp)
endpoint <- "http://spatial.linkedscience.org/sparql"
q <- "SELECT ?cell ?row ?col ?polygon
 WHERE {
 ?cell a <http://linkedscience.org/lsv/ns#Item> ;
 <http://spatial.linkedscience.org/context/amazon/Lin> ?row ;
 <http://spatial.linkedscience.org/context/amazon/Col> ?col ;
 <http://observedchange.com/tisc/ns#geometry> ?polygon .
 }"

res <- SPARQL(url=endpoint, q)$results
for(var in c("DEFOR_2002", "DEFOR_2003", "DEFOR_2004", "DEFOR_2005", "DEFOR_2006",
  "DEFOR_2007","DEFOR_2008")) {
  tmp_q <- paste("SELECT ?cell ?",var,"\n
  WHERE { \n ?cell a <http://linkedscience.org/lsv/ns#Item> ;\n
  <http://spatial.linkedscience.org/context/amazon/",var,"> ?",var," .\n }\n",sep="")
  cat(tmp_q)
  res <- merge(res, SPARQL(endpoint, tmp_q)$results, by="cell")
}
amazon <- res
amazon$row <- -res$row
coordinates(amazon) <- ~ col+row
gridded(amazon) <- TRUE
spplot(amazon,"DEFOR_2002",col.regions=rev(heat.colors(17))[-1], at=(0:16)/100,
       main="relative deforestation per pixel during 2002")
spplot(amazon, c("DEFOR_2002", "DEFOR_2003", "DEFOR_2004", "DEFOR_2005",
                 "DEFOR_2006", "DEFOR_2007","DEFOR_2008"),
       col.regions=rev(heat.colors(26))[-1], at=(0:20)/80,as.table=T,
       main="relative deforestation per pixel")
```

R - Brazilian Amazon Rainforest (2002)




relative deforestation per pixel
DEFOR_2002
DEFOR_2003
DEFOR_2004
DEFOR_2005
DEFOR_2006
DEFOR_2007
DEFOR_2008
0.25
0.20
0.15
0.10
0.05
0.00

OpenLink Virtuoso Open-Source Edition

- **Install**
- **Use it for linked data**
- **Install VAD plugin packages**
- **Visualization interface (Demo datasets: Map, Article)**
  - **Sparql endpoint**
  - **iSparql**
  - **Facet view**
  - **Pivot viewer**

OPENLINK SOFTWARE
Making Technology Work For You®

Give a little, Get a little!