

EBI RDF Platform: 6 months in

<http://www.ebi.ac.uk/rdf/>

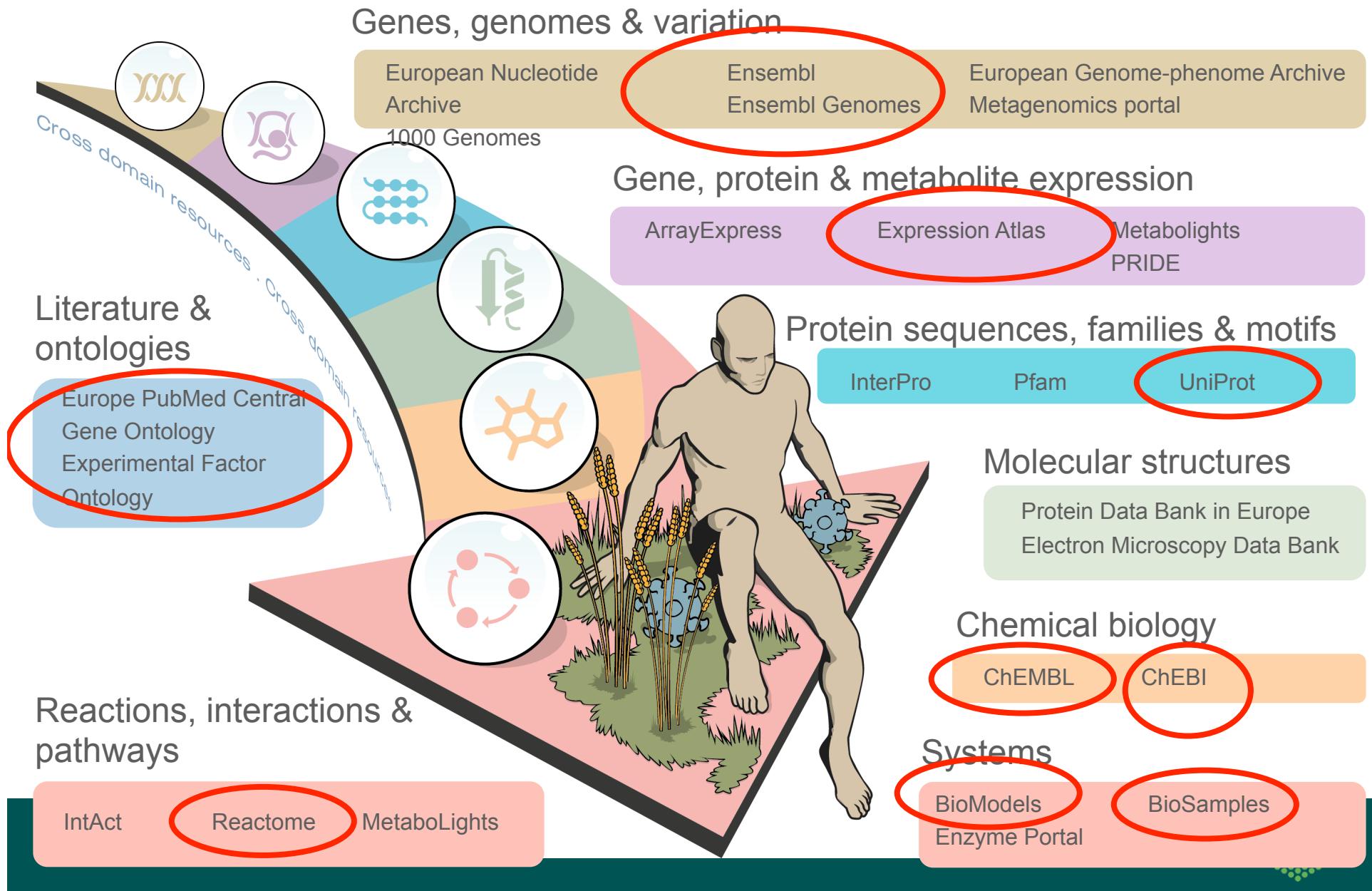
Simon Jupp <jupp@ebi.ac.uk>

What is EMBL-EBI?

- Part of the European Molecular Biology Laboratory
- International, non-profit research institute
- Europe's hub for biological data services and research
- Based in Hinxton, Cambridge



Data resources at EMBL-EBI



Why RDF at EMBL-EBI?

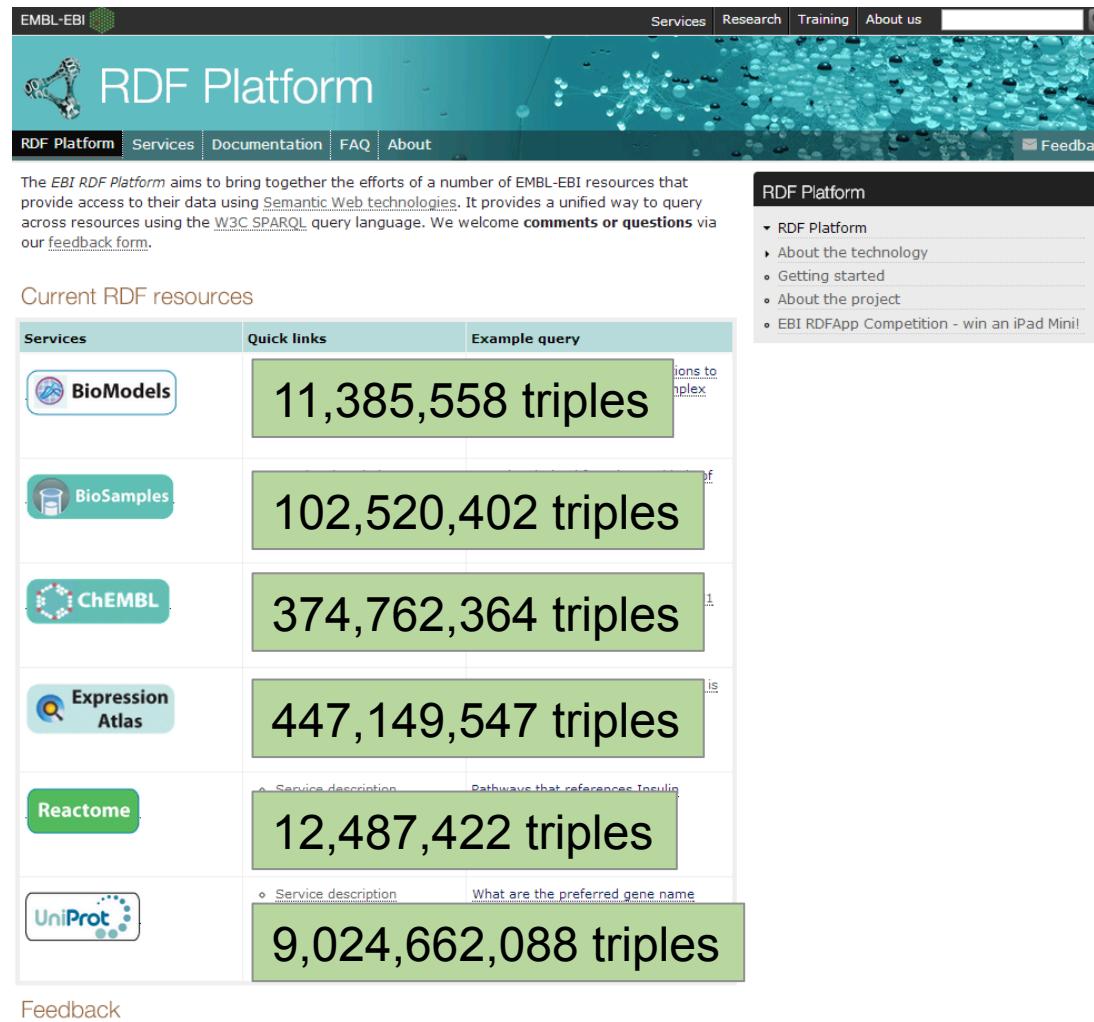
- Interest for a number of years from some of our user base particularly some industry partners
- Betas and pilots starting up within individual EBI projects without coordination
- Overall feeling that technology is maturing and life science RDF community is growing
- EBI RDF has been available elsewhere but users had concerns of stability and faithfulness to source

Components for each resource

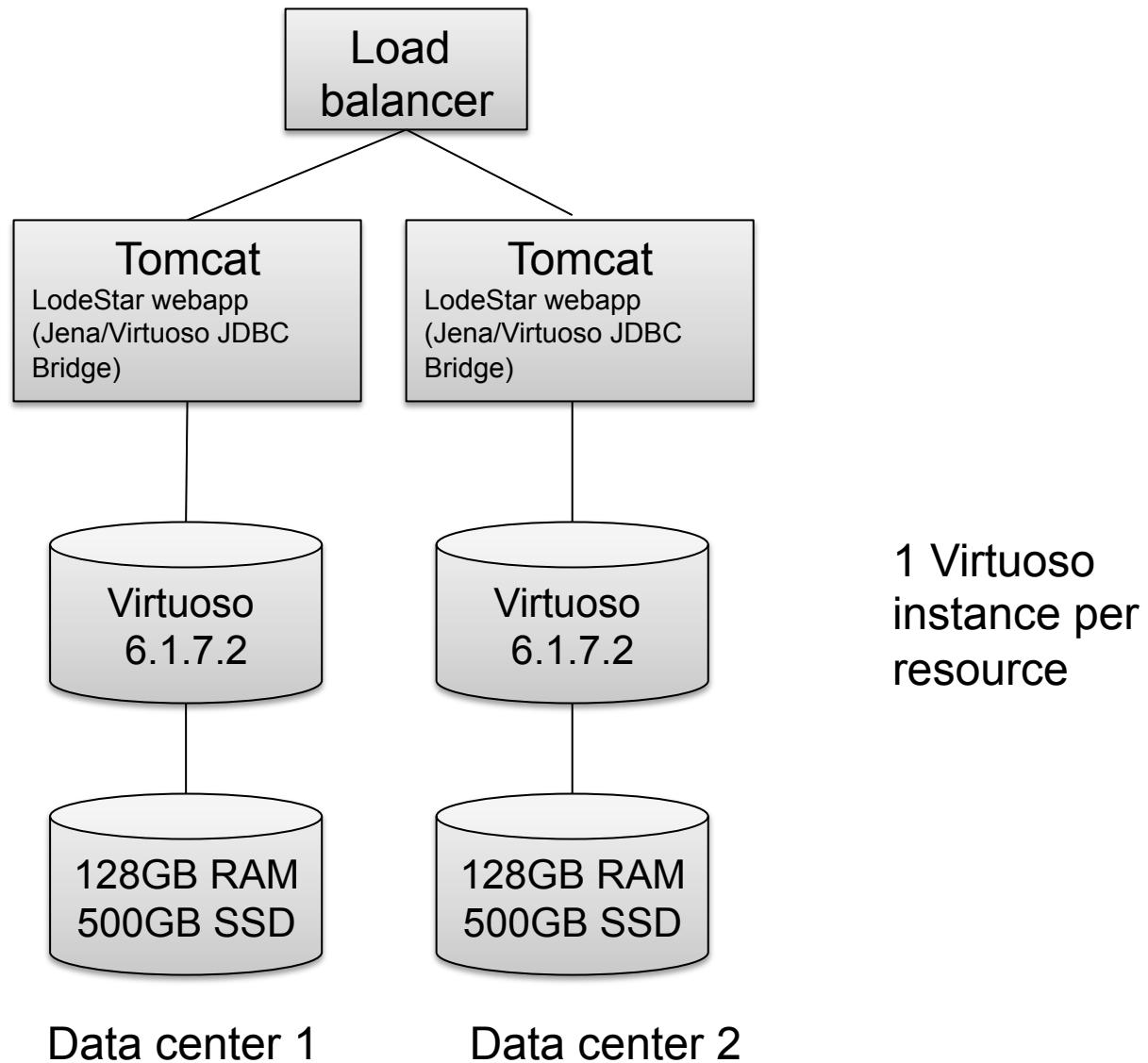
- **Stable** public SPARQL endpoint
 - Stable = production level service
- Consistent documentation with
 - Example queries
 - Schema diagram
 - Data set description (VOID)
- Bulk downloads
 - Full RDF download from FTP site
 - Hosted via Linked Data Browser (lodestar – common look and feel across resources)

The screenshot shows two views of the EMBL-EBI RDF Platform. The top view is the 'Atlas SPARQL endpoint' where a complex SPARQL query is being entered. The bottom view is the 'Atlas Linked Data' browser, showing a detailed page for the organism *Mus musculus* (Mouse), including its type (organism), description, and related entities like a fly and a person.

Numbers of triples



Technical Infrastructure



Loading times and benchmarking

- Loading tests run 8 virtual CPU core, 128GB RAM, 3TB SAN storage on flash-based Violin array.
 - RedHat Enterprise Linux version 6, virtualised on VMWare infrastructure
- Tested Sesame, OWLIM-lite, OWLIM-SE, Virtuoso 6 and 7

Dataset	Triples	OWLIM-SE	Virtuoso 7	Virtuoso 6
Expression Atlas 13.01	339,991,787	1 hr 45 mins	24 mins	25 mins
ChEMBL 14	97,717,782	32 mins	5 mins	2 mins 23 seconds
UniProt 2013_01	6,419,569,246	57 hrs 40 mins	5 hrs 11 mins	N/A
BioModels 24	5,788,106	4 mins	20 seconds	15 seconds
Reactome 42	12,927,678	9 mins	30 seconds	21 seconds

Usage analysis

- Load balancer access logs
 - No. of user
 - Number of queries
- Log every SPARQL query
 - Including execution time
- Query analysis using Jena
 - Path size
 - Used SPARQL operators
 - UNION, OPTION, GROUP, FILTER, SERVICE, SUBQUERY
 - Predicate vocabularies

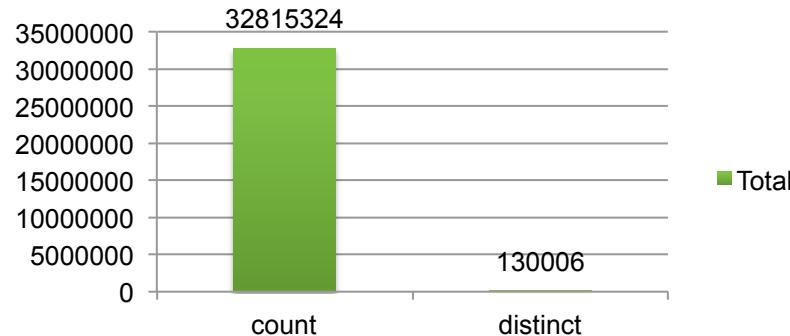
Usage

- ~500 unique visitors per month
- ~55 millions hits to site
- Most usage coming from UK
 - Internal pipelines and API calls that use the RDF platform
- Some datasets being downloaded and hosted externally behind other public APIs
 - OpenPhacts project and ChEMBL RDF
- Most resources already have good modes of access

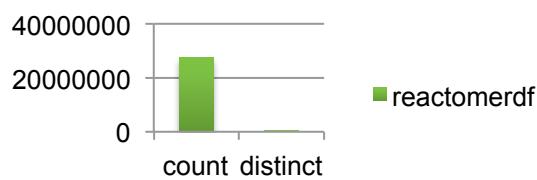


Number of queries (6 months)

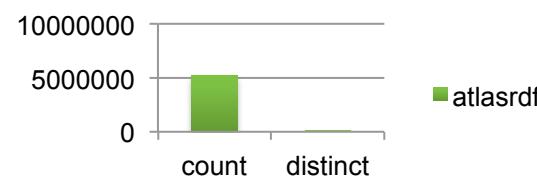
Total



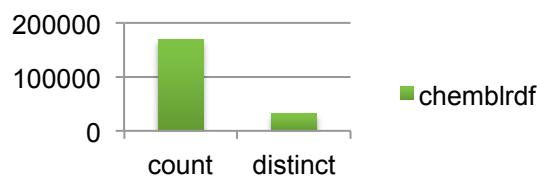
reactomerdf*



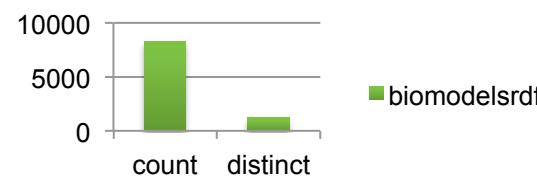
atlasrdf



chemblrdf

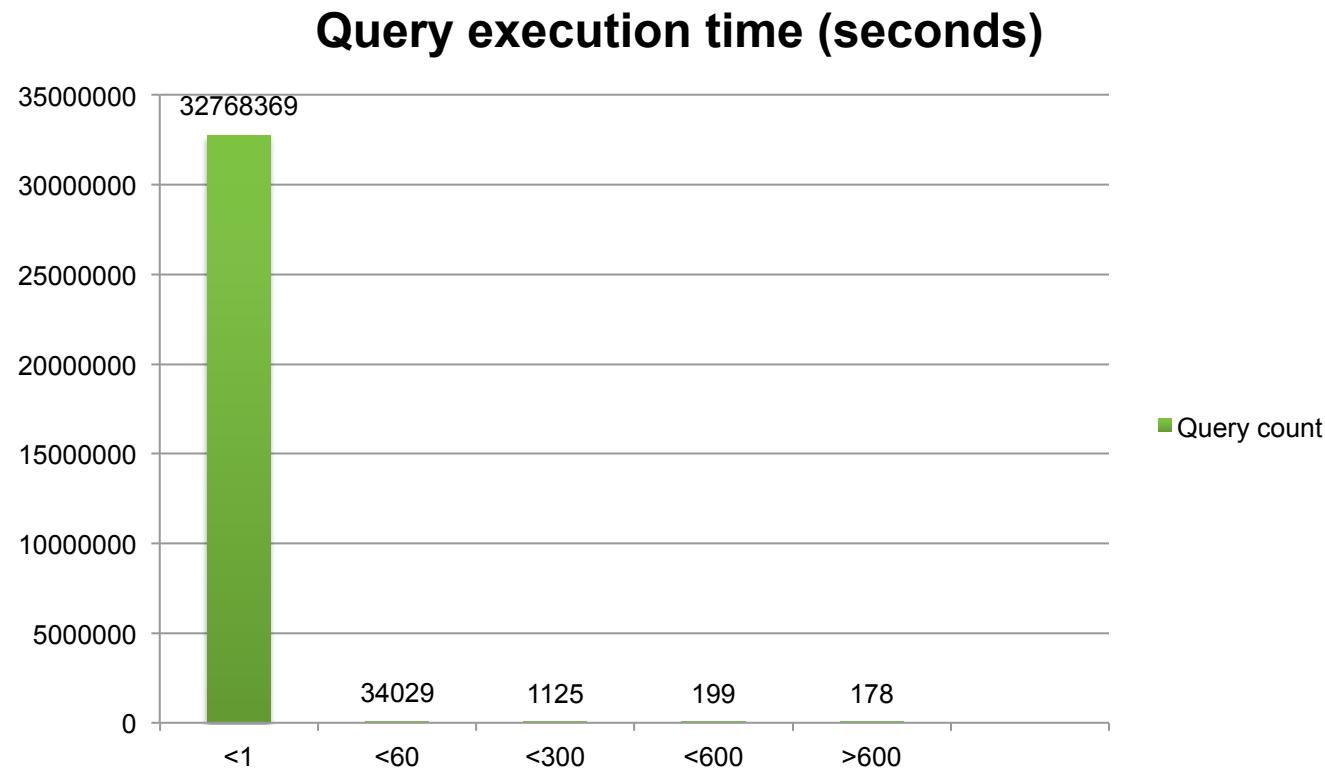


biomodelsrdf



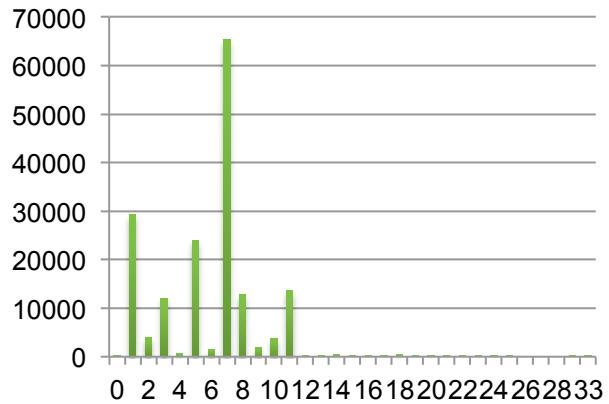
*Excludes ~90 million queries generated
By internal pipeline between Dec 1-4th 2013

Query execution time

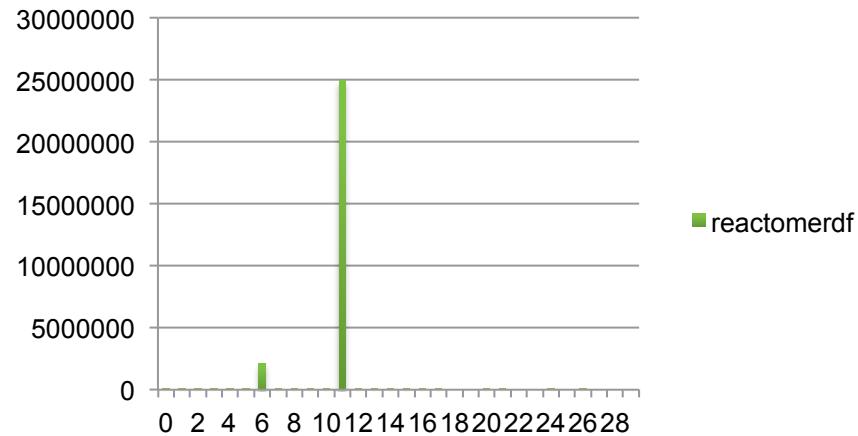


Query path size

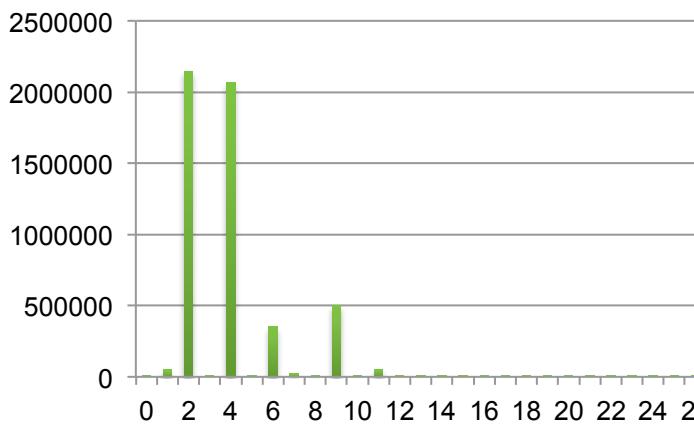
chemblrdf



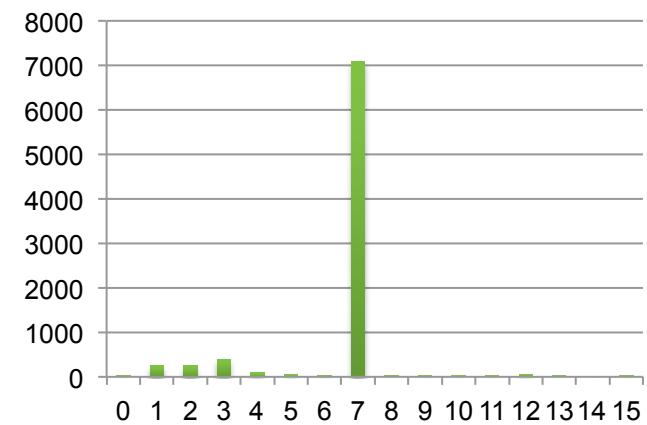
reactomerdf



atlasrdf

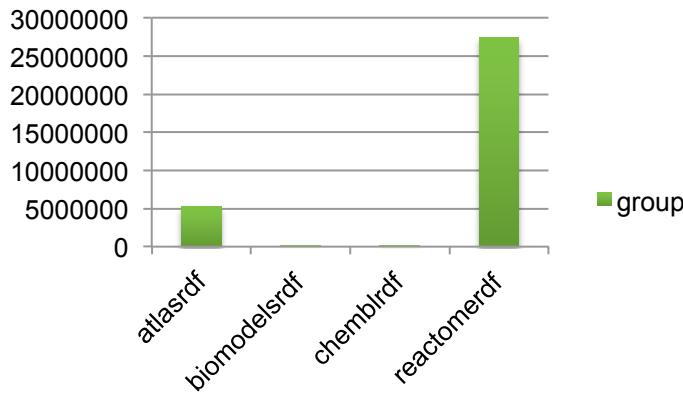


biomodelsrdf

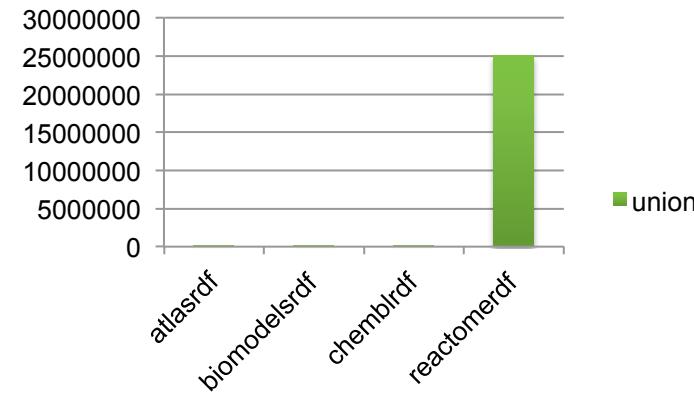


SPARQL elements

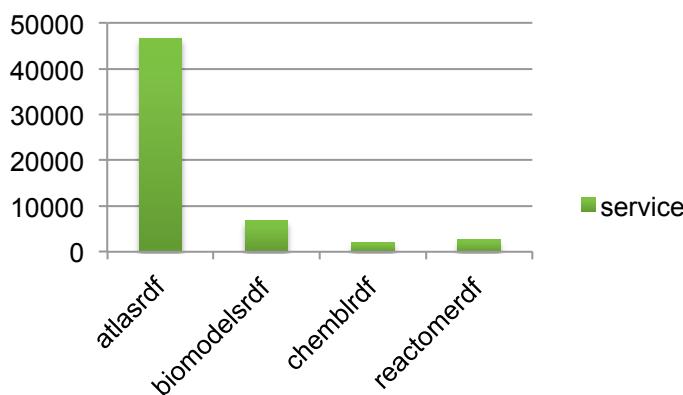
group



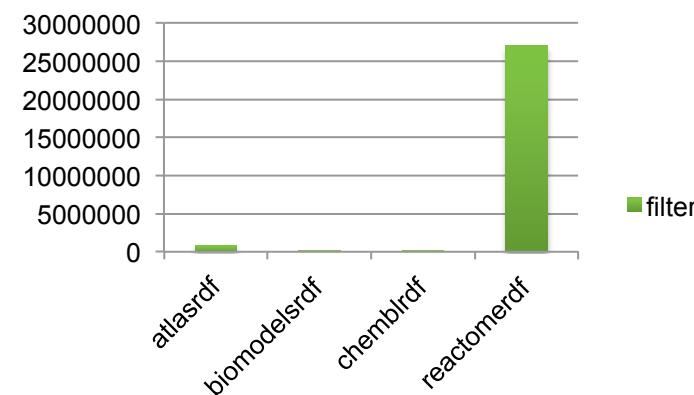
union



service

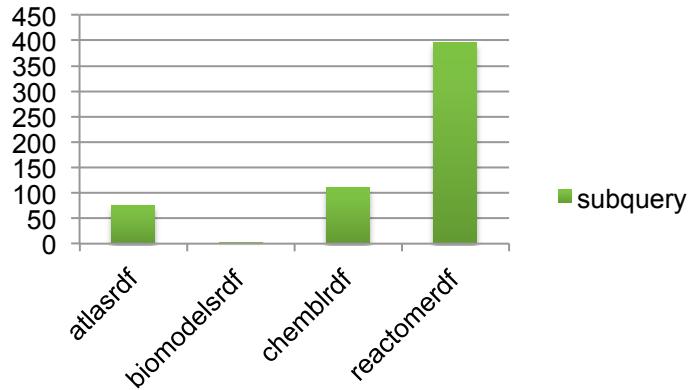


filter

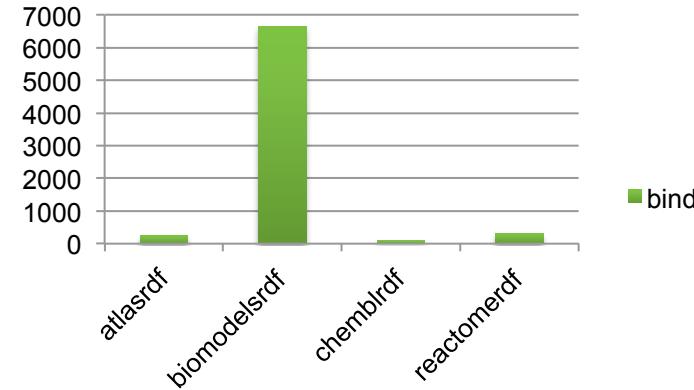


SPARQL elements cont...

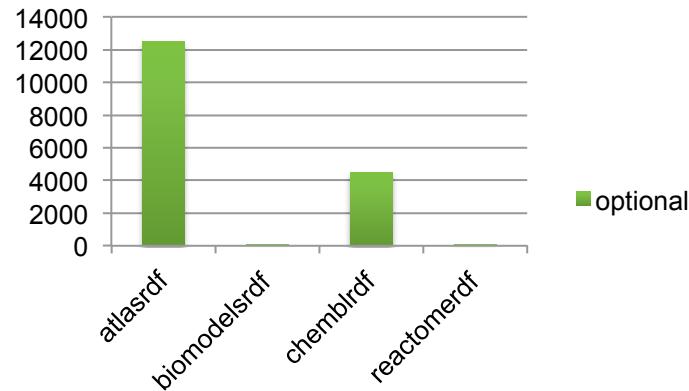
subquery



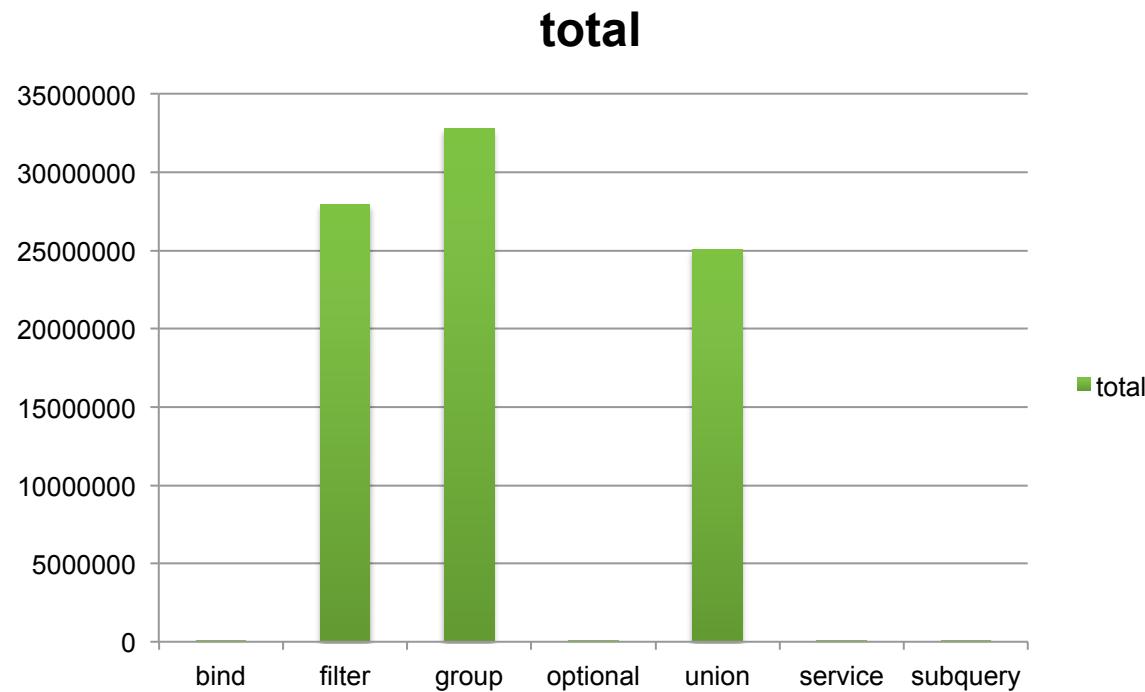
bind



optional

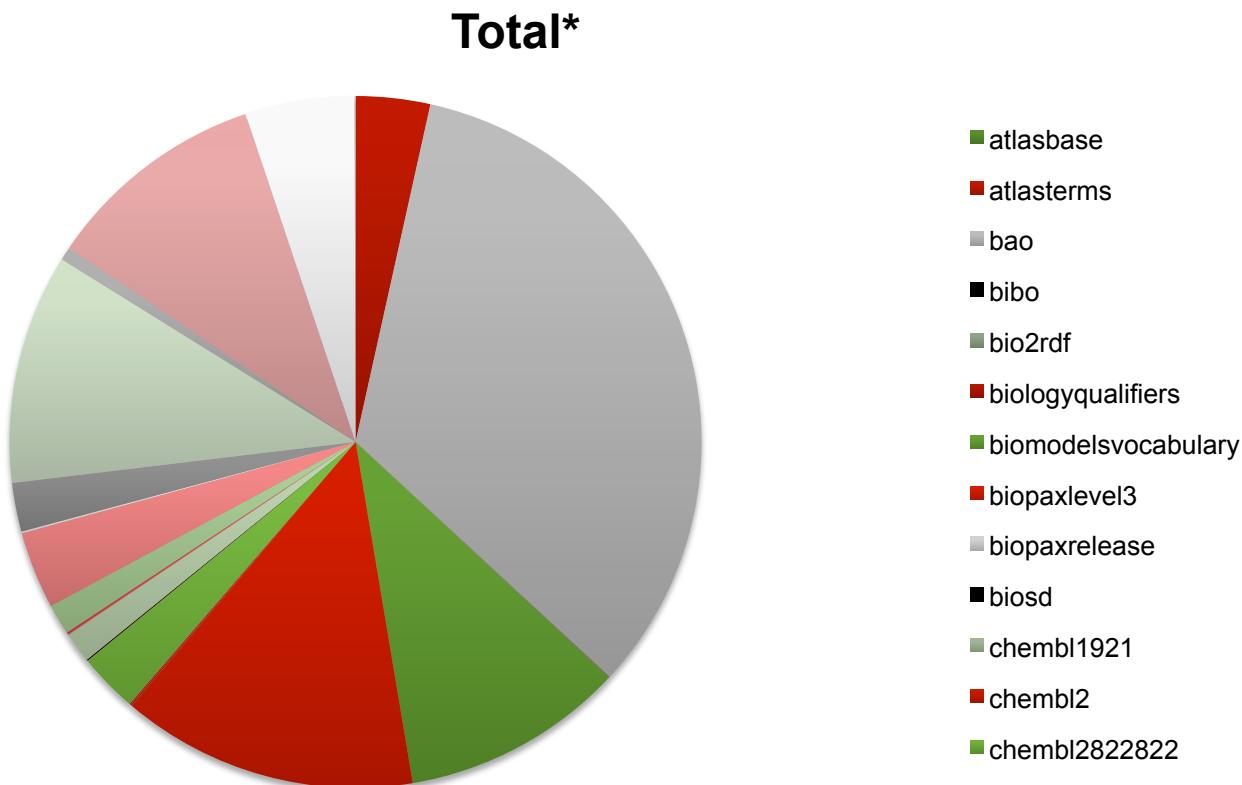


SPARQL elements



Predicate vocabularies

- 47 different vocabularies in total
- 577 relations



* Excludes rdf vocabulary

Performance problems

- Federated querying
 - One triple store vs multiple federate stores
- We use lots of ontologies (3.9 million OWL class usage)
 - Transitive queries still hard to answer
 - Get metformin associated pathways with differentially expressed genes, find any proteins that are targets for known diabetes drugs
- Some big datasets coming in 2014
 - Ensembl gene database. 78 different species
 - Many billions of triples

How can we help

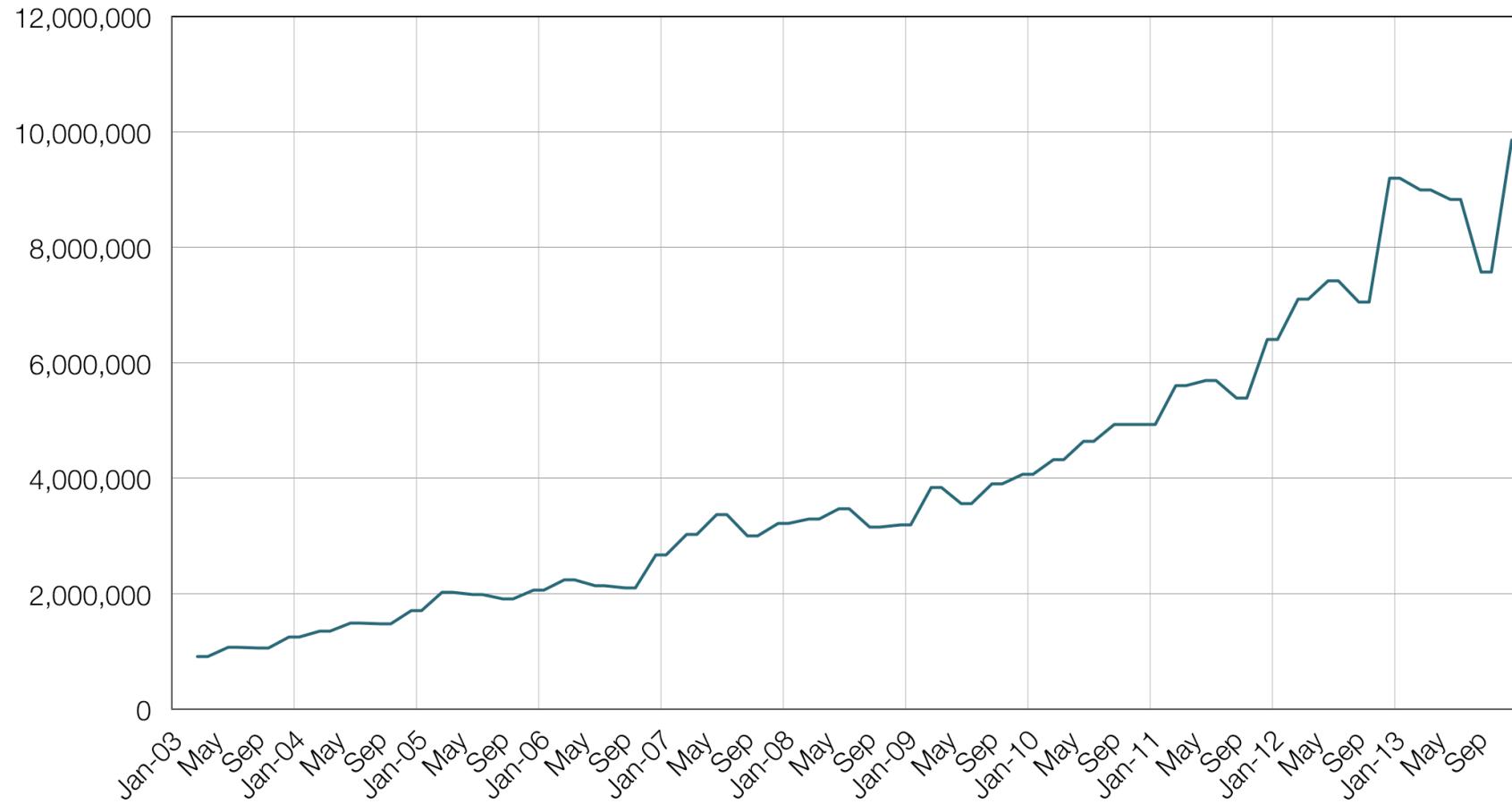
- All datasets are publicly available from FTP site
 - Currently ~1 billion triples
 - More datasets planned in 2014 expect triples to rise
- We have complex biological queries currently difficult to ask
 - We need better inference support
 - We need federated queries to scale
- Open for collaboration on projects to help develop and push this technology



Acknowledgments and Funding

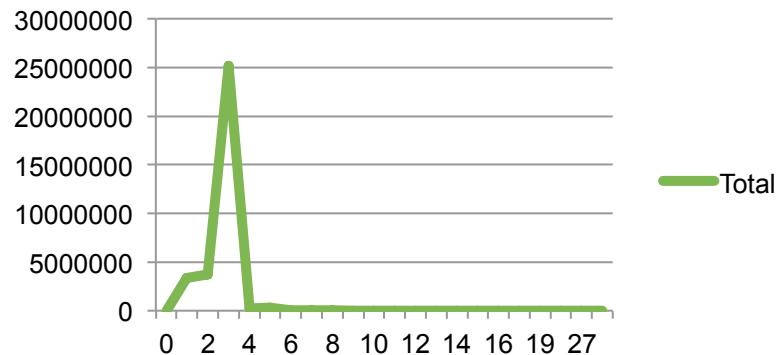
- RDF platform: James Malone, Andy Jenkinson, Mark Davies, Marco Brandizi, Sarala Wimalaratne, Leyla Garcia, Jerven Bolleman
- OntoText and OpenLink
- EMBL
- European Commission:
 - BioMedBridges [284209]
 - Diachron [601043]
- OpenPhacts
- NCBO through NIH NCBC grant U54-HG004028

Requests per day, 2003-2013

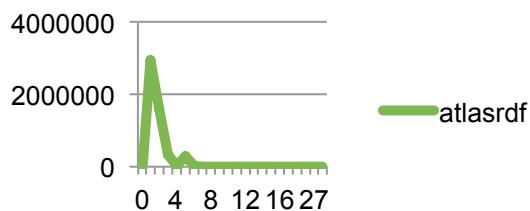


Selected variables

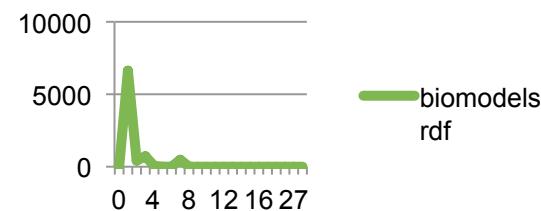
Total



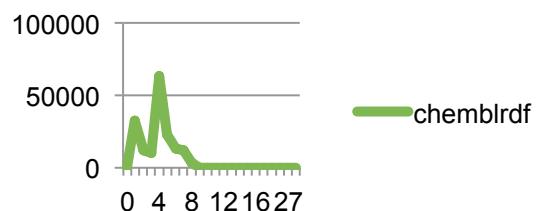
atlasrdf



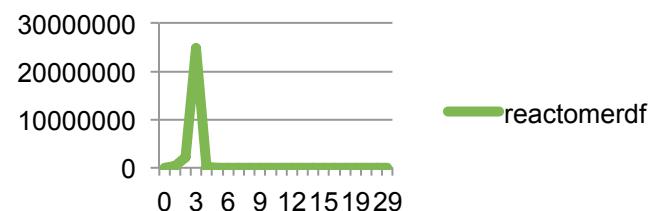
biomodelsrdf



chemblrdf

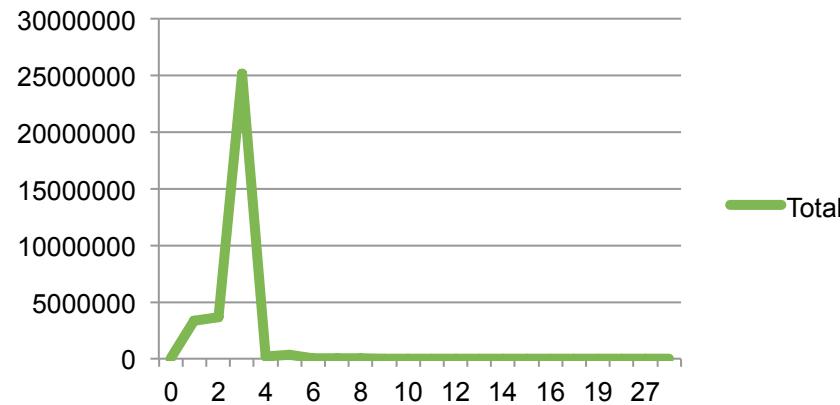


reactomerdf

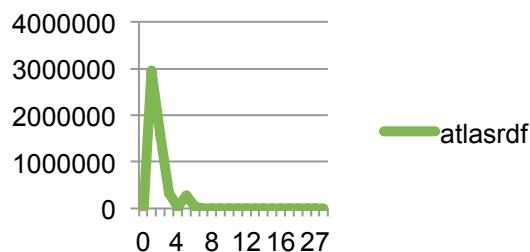


Query variables

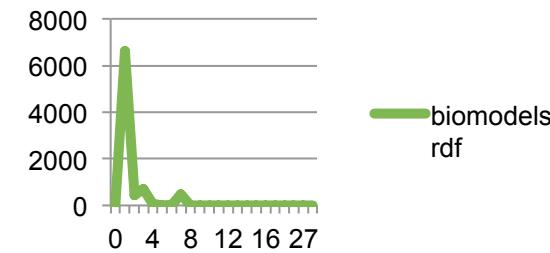
Total



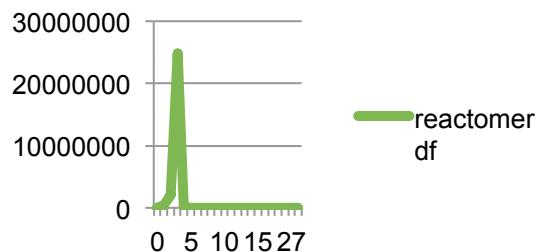
atlasrdf



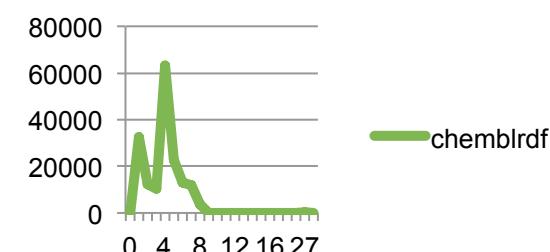
biomodelsrdf



reactomerdf



chemblrdf



RDF platform Released October 2013

The EBI RDF Platform aims to bring together the efforts of a number of EMBL-EBI resources that provide access to their data using Semantic Web technologies. It provides a unified way to query across resources using the W3C SPARQL query language. We welcome comments or questions via our feedback form.

Current RDF resources

Services	Quick links	Example query
BioModels	<ul style="list-style-type: none">◦ Service description◦ SPARQL endpoint◦ Documentation◦ RDF download	All model elements with annotations to acetylcholine-gated channel complex (GO:0005892)
BioSamples	<ul style="list-style-type: none">◦ Service description◦ SPARQL endpoint◦ Documentation◦ RDF download	Samples derived from known kinds of listeria organism
ChEMBL	<ul style="list-style-type: none">◦ Service description◦ SPARQL endpoint◦ Documentation◦ RDF download	Find drug-like (but currently not approved) molecules which bind 7TM1 GPCRs with high affinity
Expression Atlas	<ul style="list-style-type: none">◦ Service description◦ SPARQL endpoint◦ Documentation◦ RDF download	Under what experimental conditions is Ensembl gene ENSG00000129991 (TNNI3) expressed?
Reactome	<ul style="list-style-type: none">◦ Service description◦ SPARQL endpoint◦ Documentation◦ RDF download	Pathways that references Insulin (P01308)
UniProt	<ul style="list-style-type: none">◦ Service description◦ SPARQL endpoint◦ Documentation◦ RDF download	What are the preferred gene name and disease annotations of all human UniProt entries that are known to be involved in a disease?

Feedback



Services > Gene Expression Atlas

Gene Expression Atlas

The Functional Genomics Production Team are pleased to announce the beta publication of gene expression data from the Gene Expression Atlas as RDF Linked Open Data. The Expression Atlas Linked Dataset is an alternative API to the Gene Expression Atlas data. The purpose of this API is to enable richer queries over the data, it also supports federated queries over other linked datasets, including ChEMBL, Reactome, BioModels, BioSample, Uniprot, BioPortal and Bio2RDF.

The primary interface to the Expression Atlas RDF data is via the [SPARQL endpoint](#).

Gene Expression Atlas

- [SPARQL endpoint](#)
- [Atlas documentation](#)

Full VOID dataset description at <http://rdf.ebi.ac.uk/dataset/atlas/13.07>

Title	Gene Expression Atlas RDF
Description	RDF representation of all the experiments loaded into the Gene Expression Atlas database.
Version	13.07
Issued	July 19 2013
Number of triples	447149547

Contact

If you are interested in this work, please sign up to our [EFO user mailing list](#) where we will announce updates.

If you have questions or would like more information, you can email Simon Jupp [jupp](mailto:jupp@ebi.ac.uk) [at] ebi.ac.uk and James Malone [malone](mailto:malone@ebi.ac.uk) [at] ebi.ac.uk directly.

Atlas RDF Related Links

- [Experimental Factor Ontology \(EFO\)](#)
- [Gene Expression Atlas](#)
- [ArrayExpress](#)
- [Functional Genomics Production Team](#)



EMBL-EBI 

Services Research Training About us

RDF Platform



RDF Platform Services Documentation FAQ About Feedback

Services > Atlas > Atlas SPARQL endpoint

Expression Atlas SPARQL Endpoint

Enter SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/atlas/>

SELECT DISTINCT ?experiment ?description WHERE
{?experiment
a atlasterms:Experiment ;
dcterms:description ?description ;
atlasterms:hasAssay
[atlasterms:hasSample
[atlasterms:hasSampleCharacteristic
[ atlasterms:propertyType ?propertyType ;
atlasterms:propertyValue ?propertyValue]
]
]
filter regex (?description, "diabetes", "i")}
```

RDFS inference?

Output:

Results per page:

Example Queries

- [Query 1](#)
Get experiments where the sample description contains diabetes
- [Query 2](#)
Get differentially expressed genes where factor is asthma
- [Query 3](#)
Show expression for ENSG00000129991 (TNNI3)
- [Query 4](#)
Show expression for ENSG00000129991 (TNNI3) with its GO annotations from Uniprot
(Federated query to http://beta.sparql.uniprot.org/sparql)
- [Query 5](#)
For the genes differentially expressed in asthma, get the gene products associated to a Reactome pathway
- [Query 6](#)

EMBL-EBI



www.ebi.ac.uk/rdf/services/atlas/describe?uri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FNCBITaxon_10090

EMBL-EBI 

Services Research Training About us  

RDF Platform

RDF Platform Services Documentation FAQ About 

Services > Atlas > Atlas Linked Data

About: [Mus musculus](#) 

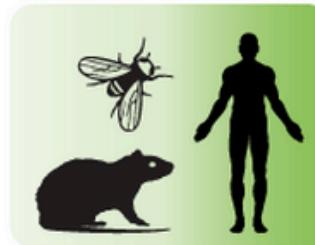
http://purl.obolibrary.org/obo/NCBITaxon_10090

Identifier : "10090"

Type: [organism](#)

A material entity that is an individual living system, such as animal, plant, bacteria or virus, that is capable of replicating or reproducing, growth and maintenance in the right environment. An organism may be unicellular or made up, like humans, of many billions of cells divided into specialized tissues and organs.

[more types...](#)



Expression Atlas RDF

- [Dataset home](#)
 - [SPARQL endpoint](#)
 - [Documentation](#)

Related to

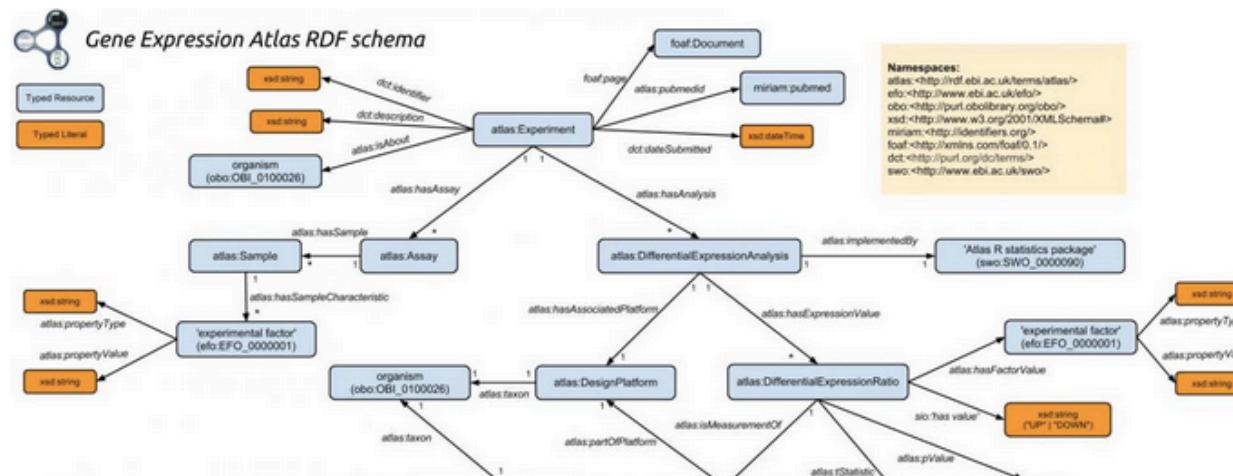
subClassOf

..

Atlas documentation

The schema for the atlas datasets is depicted below. It shows how resources are connected in the RDF graph. Each resource is types according to the atlas terms ontology. The stable URI for the latest version of this ontology is <http://rdf.ebi.ac.uk/terms/atlas>.

This ontology acts as our internal schema and provides a basic description of the resources in the Atlas RDF dataset. In order to facilitate integration of the Atlas RDF data with other datasets, we provide an additional ontology that maps the atlas terms ontology to several external ontology development projects. We currently map to the Semantic Science Integration Ontology (SIO), the Ontology for Biomedical Investigations (OBI), the Experimental Factor Ontology (EFO) and the ontology of bioinformatics operations, topics, types of data including identifiers, and formats (EDAM). A table of the mappings can be seen here. The full atlas terms mapping ontology can be found at <http://rdf.ebi.ac.uk/terms/atlas-mapping>. We are happy to extend our mapping to other relevant ontologies and are keen to discuss the integration of this dataset with other similar resources.



Documentation

Atlas documentation

- Atlas contact
- Atlas download
- Atlas examples
- AtlasRDF R package

BioModels documentation

ChEMBL documentation

Reactome documentation

UniProt documentation

Using the SPARQL Endpoints

Example SPARQL queries

Programmatic access

Technical documentation