

Discover new Perspectives

Data Federation in the Pharma Industry

Marc Lieber



BASEL ▪ BERN ▪ BRUGG ▪ DÜSSELDORF ▪ FRANKFURT A.M. ▪ FREIBURG I.BR. ▪ GENÈVE
HAMBURG ▪ KOPENHAGEN ▪ LAUSANNE ▪ MÜNCHEN ▪ STUTTGART ▪ WIEN ▪ ZÜRICH

trivadis
makes IT easier. ■ ■ ■

■ The Database World is changing

To manage a mix of structured, semi-structured and unstructured data



NoSQL
Databases

Not only SQL

Graph
Databases

Key Value
Stores

Wide Column
Stores

Document Store
Databases

Oracle NoSQL,
Redis, Riak KV ...

Cassandra,
Hbase ...

MarkLogic,
MongoDB ...

Property
Graphs

RDF Triple
Stores

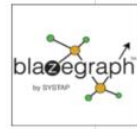


Neo4J,
Datstax Cassandra,
Oracle

Oracle Spatial&Graphs,
Allegrograph, Virtuoso,
Blazegraph, Marklogic, Enzo

■ NoSQL Graph Databases

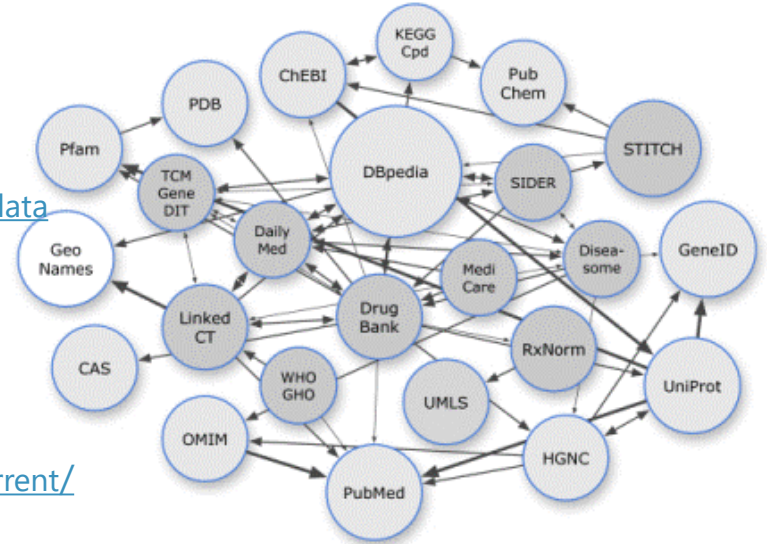
- **We just begin to explore the potential of Graph Databases**
 - Big data analytic, Social Network Analysis, data federation, Linked Open Data, meta-modelling,
 - Enables new views on data in ways that would be very difficult to do using relational data structures
 - Allows new type of queries that would be prohibitively expensive, or even impossible, to run on other databases
- Property Graphs or RDF Triple Stores ?
 - RDF is more appropriate for Data Federation and Linked Open Data technologies
 - Property Graphs are best for traversal and Graph analytics
 - Some providers are able to translate Property Graph into RDF Graphs



W3C Semantic Web technologies

1. Technologies based on ontologies that enable the proper integration of knowledge in a way that is reusable by several applications across businesses, from discovery to corporate affairs
2. Pharma Specific ontologies : Linked Open Drug Data (LODD) for sharing and interlinking data

DBPedia <http://wiki.dbpedia.org/news/dbpedia-based-rdf-dumps-wikidata>
Wikidata <https://dumps.wikimedia.org/wikidatawiki/entities/>
OpenPHACTS <https://www.openphacts.org/>
Bio2RDF <http://download.bio2rdf.org/current/release.html>
LinkedLifeData <http://linkedlifedata.com/>
Elsevier <https://www.elsevier.com/>
PubChem <ftp://ftp.ncbi.nlm.nih.gov/pubchem/RDF/>
Pathway Commons <http://www.pathwaycommons.org/archives/PC2/current/>



Linked Open Data : EBI

Current RDF resources

Services



- Data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation

UniProt Taxonomy - Homo sapiens (Human)

Images may be subject to copyright.
news.nationalgeographic.com www.bl.uk upload.wikimedia.org

Taxonomy navigation

Map to
UniProtKB (154,485)
Reviewed (20,200)
Swiss-Prot
Unreviewed (134,285)
TrEMBL
Proteomes (1)

Mnemonic	HUMAN
Taxon identifier	9606
Scientific name	Homo sapiens
Common name	Human

Your SPARQL query

```
1 PREFIX up:<http://purl.uniprot.org/core/>
2 PREFIX taxon:<http://purl.uniprot.org/taxonomy/>
3 PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
4 PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
5 PREFIX faldo:<http://biohackathon.org/resource/faldo#>
6 SELECT ?protein ?annotation ?begin ?text
7 WHERE
8 {
9   ?protein a up:Protein .
10  ?protein up:organism taxon:9606 .
11  ?protein up:annotation ?annotation .
12  ?annotation a up:Natural_Variant_Annotation .
13  ?annotation rdfs:comment ?text .
14  ?annotation up:substitution ?substitution .
15  ?annotation up:range/faldo:begin/faldo:position ?begin .
16  ?protein up:sequence ?sequence .
17  ?sequence rdf:value ?value .
18  BIND (substr(?value, ?begin, 1) as ?original) .
19  FILTER(?original = 'Y' && ?substitution = 'F') .
20 }
21
```

Submit Query

Examples

1. Select all taxa from the UniProt taxonomy
2. Select all bacterial taxa from the UniProt taxonomy
3. Select all E-Coli K12 entries and their amino acid sequences
4. Select the UniProt entries for the UniProt cross-reference category 'A4_HUMAN': (show)
5. Select a mapping of the UniProt cross-reference category '3D_structures': (show)
6. Select all cross-referenced entries that are classified in the UniProt taxonomy category '3D_structures': (show)
7. Select all UniProt entries that contain the text 'A4_HUMAN' in their protein name, that have a UniProt cross-reference to the UniProt taxonomy: (show)
8. Select the preferred name for each UniProt entry.

SPARQL End point and service calls

SPARQL allows queries on **local** data and on **remote** data accessible through a SPARQL end point

Example : call the Uniprot SPARQL end point to enrich the locally stored data

```
1 PREFIX up: <http://users.ugent.be/~tdenies/up/>
2 prefix owl: <http://www.w3.org/2002/07/owl#>
3 prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 prefix bio: <http://www.biopax.org/release/biopax-level3.owl#>
5 prefix skos: <http://www.w3.org/2004/02/skos/core#>
6 prefix : <http://www.pharma.com/pubChem/>
7 prefix up: <http://purl.uniprot.org/core/>
8 select *
9 where { ?protein skos:closeMatch|skos:exactMatch ?uniprot .
10 OPTIONAL {SERVICE <http://sparql.uniprot.org/sparql> { ?uniprot rdfs:label ?uname ; up:organism/up:scientificName
?organism }}
11 }
12
```

QUERY RESULTS

Table Raw Response

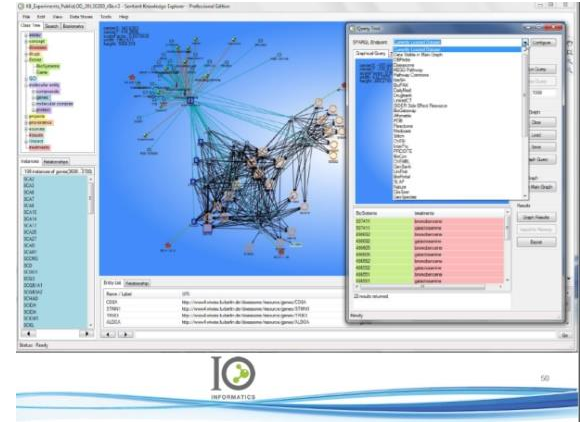
Showing 1 to 2 of 2 entries

Search: Show 50 entries

	protein	uniprot	uname	organism
1	http://rdf.ncbi.nlm.nih.gov/pubchem/prot ein/GI124375976	http://purl.uniprot.org/uniprot/P10275	"Androgen receptor"	"Homo sapiens"
2	http://rdf.ncbi.nlm.nih.gov/pubchem/com pound/CID2244	http://www.wikidata.org/entity/Q18216		

■ RDF Semantic technologies

- Use Case 1: **Find signal in large or complex data sources** (data silos)
 - For Link Analysis, Pattern discovery, detect anomalies
- Use Case 2: Build a **Semantic Metadata Layer**
- Use Case 3: **Data Integration**
 - Federate data and create a semantic data lake



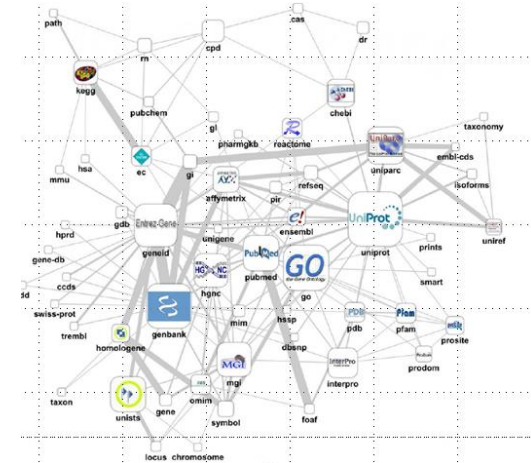
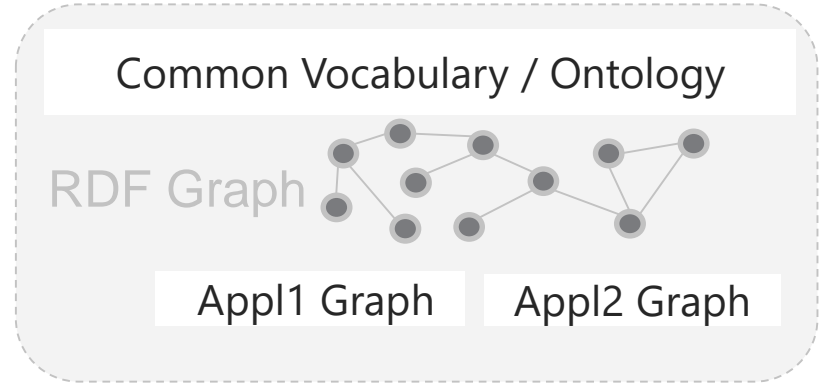
RDF Triple Stores to store the data
SPARQL to query the data, **SPARUL** to change it
Supports Reasoning/Inferencing features based on predefined rules such as
OWL, SKOS for data discovery

■ Building an ontology

Ontologies are the conceptualization of knowledge, which explains how different concepts are linked and represented

W3C RDF Semantic Technology is a natural choice for metadata and real-world facts. You add relationships between facts, documents and you can infer new facts

Ontology modeling tools: TopBraid, Protégé



Semantic Web to enrich your information layer

Semantic Technology Enables an Investigative Approach to Diverse Data from Varying Sources

In-house Comp. Intel. database

Company	Website	Mkt Cap
Bio Corp	biocorp.com	\$2.2B
Drugs123	drugs123.com	\$930M
...

Web news

On Tuesday, **Drugs123 Inc.** announced **phase 1** development of their newest **sleep aid** therapeutic, **Narcoleptol**.

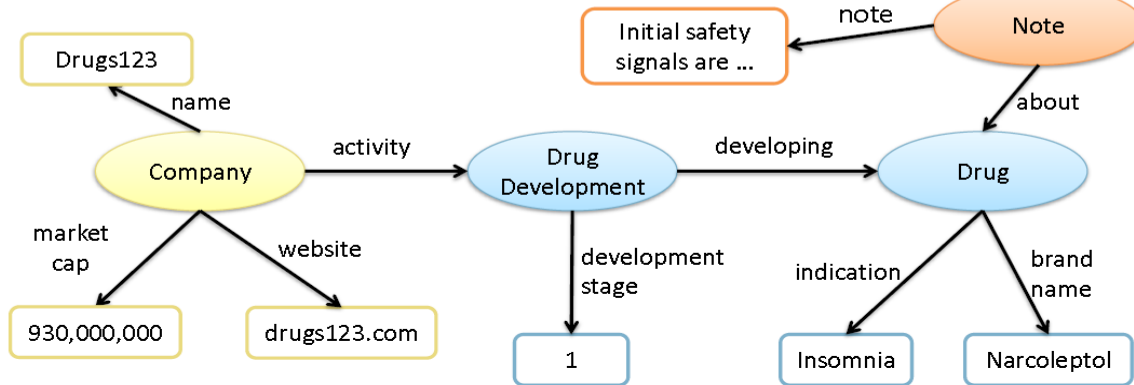
Ad-hoc notes, forms, etc.

Review Notes

Drug:

Reviewed on:

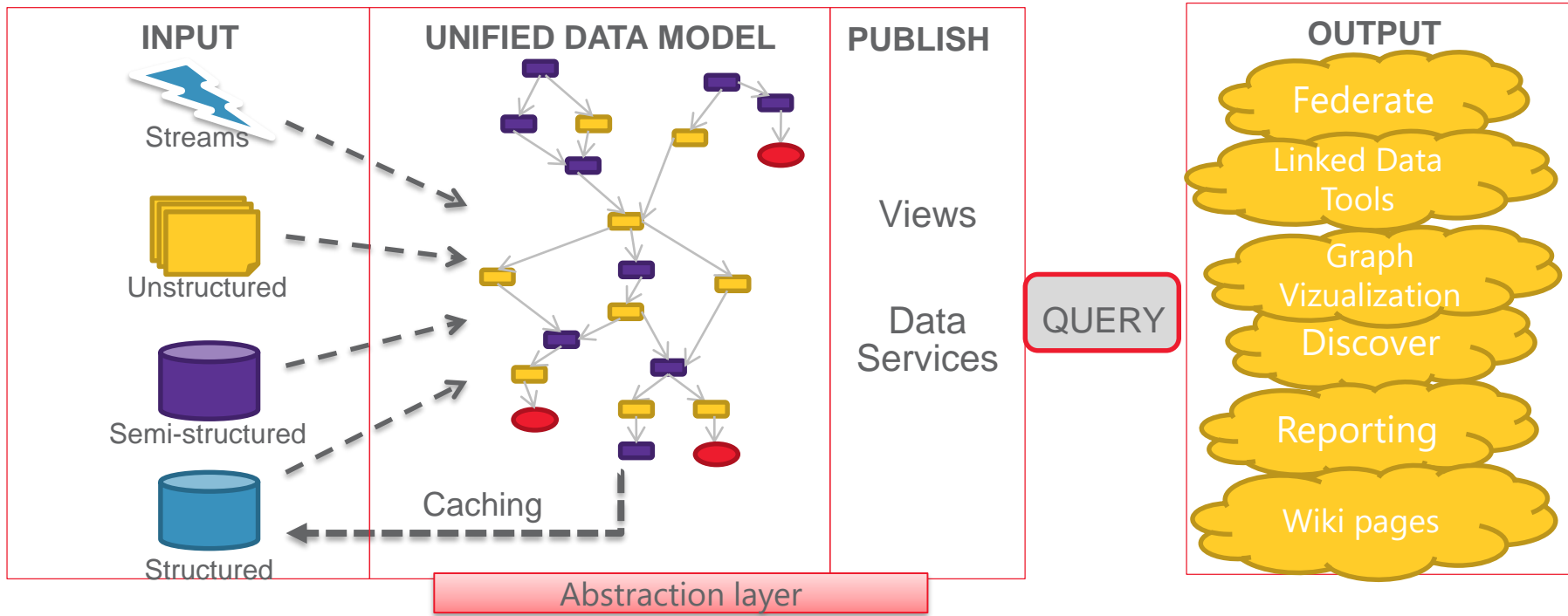
Notes:



©2013 Cambridge Semantics Inc. All rights reserved.

■ Build a Semantic Metadata Layer

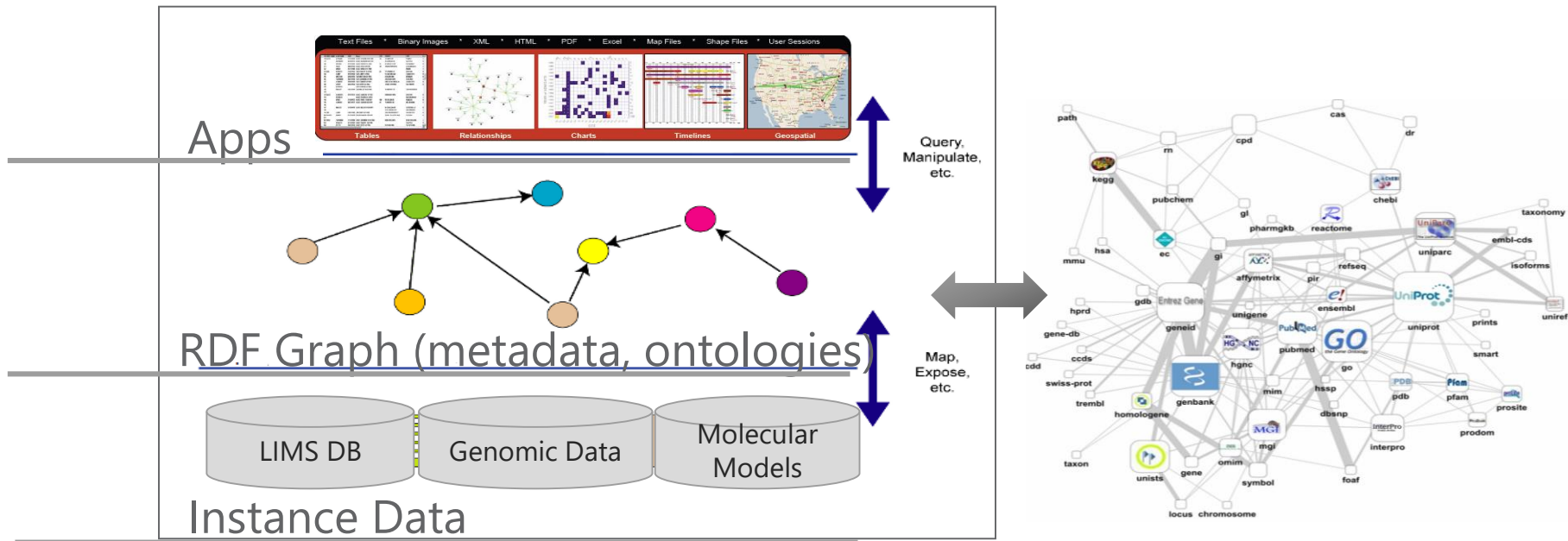
Ingest and semantify, index and query



■ Master Data Management / Reference Data / enterprise vocabulary

- The RDF graph model can align the entities and semantics in the graph with the semantics of an enterprise vocabulary or ontology.
- One Master & Meta-terminology (countries, region, product, Genes, proteins, ...)
- Applications consume the master data in a flexible way through SPARQL end points, Java APIs, database views or webservice
- This ensures that application developers code to a common, semantically consistent vocabulary when performing federated queries
- A unique benefit of RDF graphs is that resources enable data integration between different and even disparate data sets
- Integration is possible because each resource has a globally unique universal resource identifier (URI). [<http://rdf.cdisc.org/send-terminology#C85493.C85564>](http://rdf.cdisc.org/send-terminology#C85493.C85564)

Integration and Discovery of Disparate Life Science Data



RDF graph is an enterprise metadata framework. The metadata graph associates underlying instance data to other data resources based on their semantics.

■ Semantic metadata integration, linked data

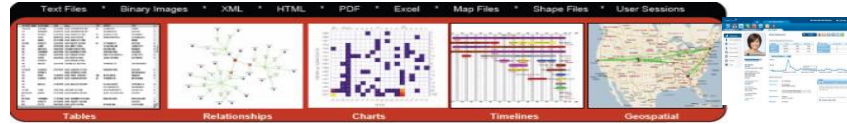
Benefits:

- **Flexible Data Modeling:** incorporates new kinds of data and relationships
- **Data Integration:** allows semantic and relational data assets to be interrogated together for first time for greater discovery
- **New set of services:** extends accessibility and usability of enterprise data
→ The linking of resources enables interoperability between apps that exchange data
- **Better Analysis:** enables discovery of unknown relationships based on semantics; visualization of relationships

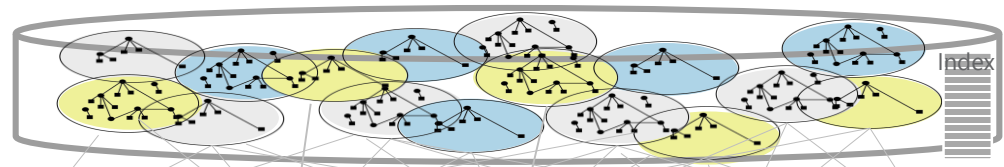
Health Care Enterprise metadata framework

Harmonizing the Electronic Health Care Ecosystem using a Triple store

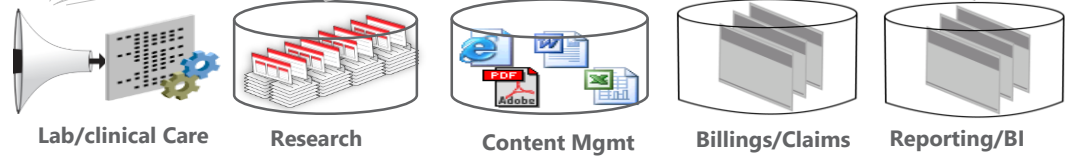
Enterprise-wide, Patient-centric, longitudinal Record System



Domain Ontologies
(business metadata + Ontologies)



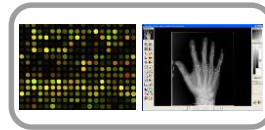
Data Servers



Data Sources / Data Types



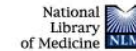
Social Media



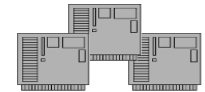
Medical Devices



Lab Information Systems



Subscription Services

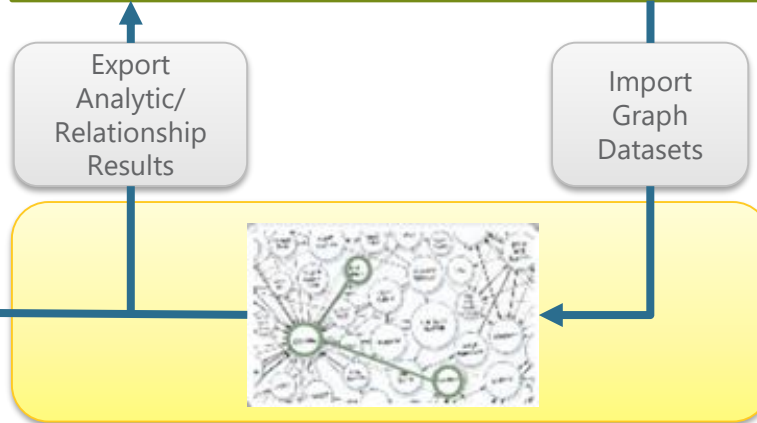
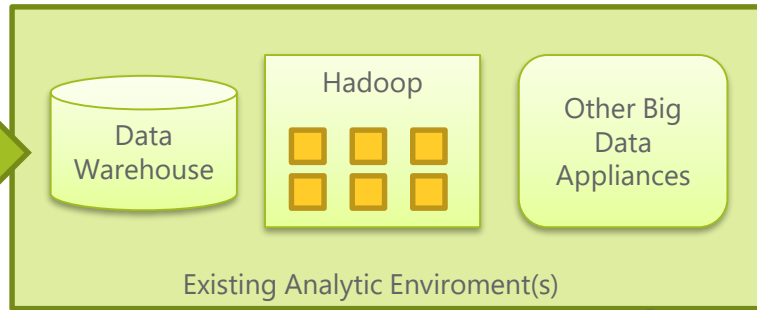


Legacy Patient Records

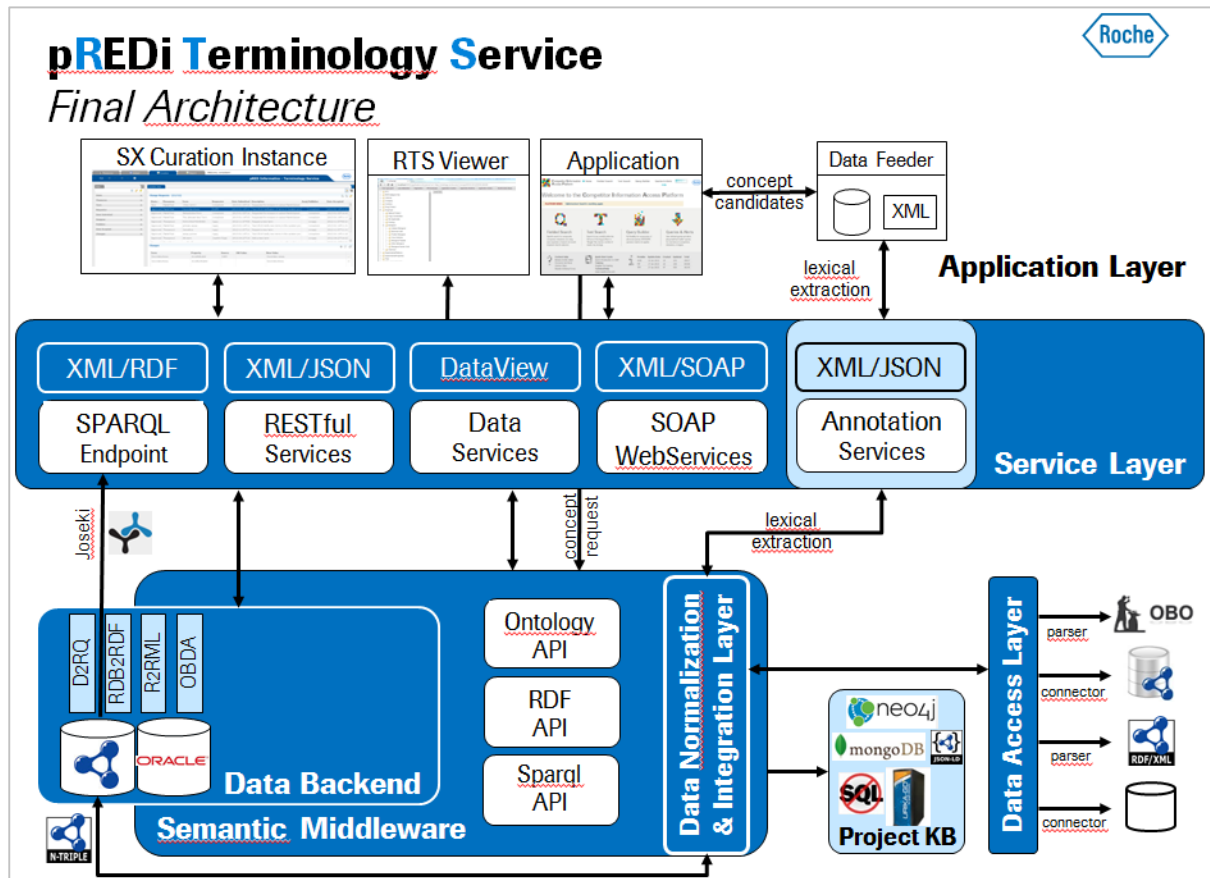
Semantic Web technologies complements Data Warehouse/Hadoop by offloading Graph Analytics



- Unstructured data (PDFs, MS Word, pictures)
- XML documents,
- Multimedia content
- Web content
- Satellite and medical imagery
- Maps and geographic information
- Sensor data
- Semantic web structures



Semantic Metadata Layer at Roche



- Oracle Spatial & Graph
- Metadata layer using SKOS and SKOS-XL
- REST Api
- SPARQL End Point
- Curation platform
- Validation, Versioning, Security

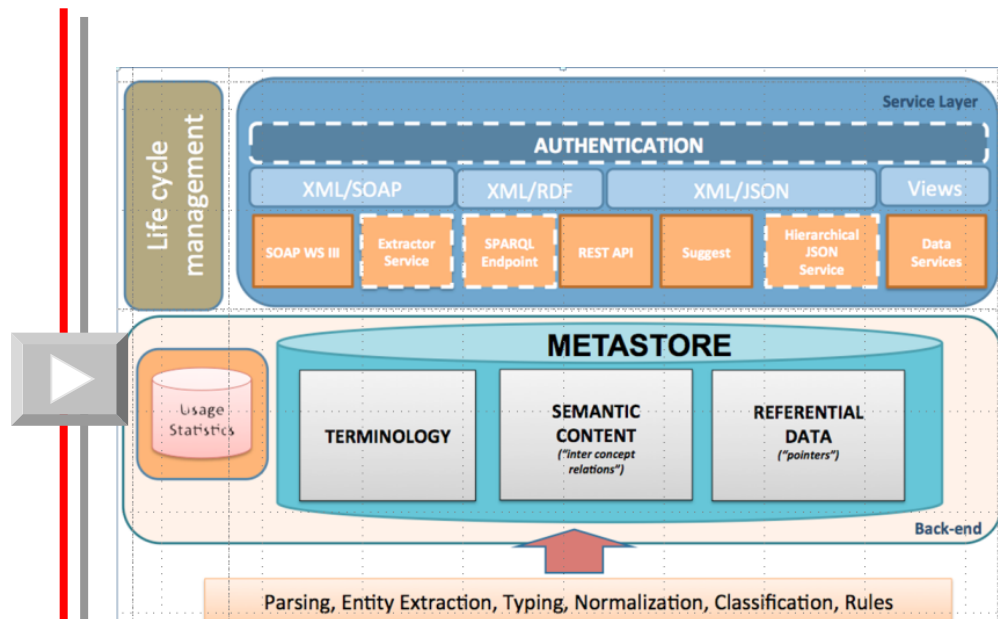
Semantic search on scientific data

Business Challenge

- Link database information on genes, proteins, metabolic pathways, compounds, ligands, etc. to original sources.
- Increase productivity for accessing, sharing, searching, navigating, cross-linking, analyzing internal /external data

Solution

- Semantic integration layer using RDF graph
- Rich domain-specific terminology (biology, chemistry and medicine) 1.6 M terms
- Terminology Hub: 8 GB of referential data (ontologies) that cross-reference various data repositories.
- About 140 million triples



■ Query RDF Data with Oracle

1. SPARQL
 - SQL SEM_MATCH
 - SPARQL end point
2. SQL
3. PL/SQL
4. Java
5. REST APIs
6. SOAP Webservices

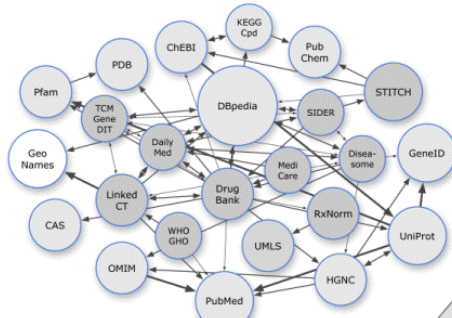
```
SELECT cname, pop, o, neighbor
FROM TABLE (SEM_MATCH (
  'PREFIX : <http://www.semwebtech.org/mondial/10/meta#>
  SELECT ?x ?cname ?pop ?gdp ?o ?neighbor
  WHERE {?x rdf:type :Country .
        ?x :name ?cname filter (sameTerm(?cname,"Switzerland")) .
        ?x :population ?pop .
        ?x :gdpTotal ?gdp .
        ?x :neighbor ?o .
        ?o :name ?neighbor}',
  SEM_Models('VIRT_MODEL_MONDIAL'),
  SEM_Rulebases('',null, null, null,null));
```

Query Result x

SQL | All Rows Fetched: 5 in 0.192 seconds

	CNAME	POP	O	NEIGHBOR
1	Switzerland	7207060	http://www.semwebtech.org/mondial/10/countries/D/	Germany
2	Switzerland	7207060	http://www.semwebtech.org/mondial/10/countries/I/	Italy
3	Switzerland	7207060	http://www.semwebtech.org/mondial/10/countries/F/	France
4	Switzerland	7207060	http://www.semwebtech.org/mondial/10/countries/FL/	Liechtenstein
5	Switzerland	7207060	http://www.semwebtech.org/mondial/10/countries/A/	Austria

Data federation & Semantic Knowledge Hub

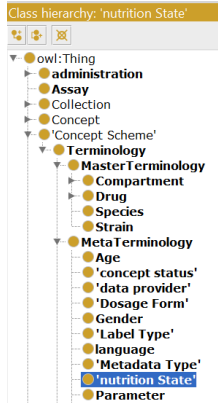


Linked Open Data

SPARQL Endpoint



Reporting + visualisation



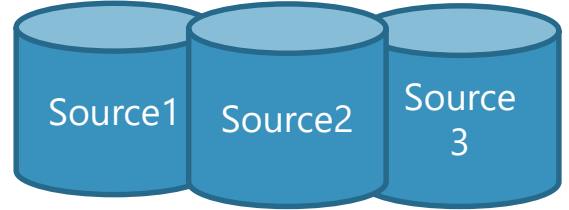
Metamodel ontology



Store n Data sources in RDF + Metamodel

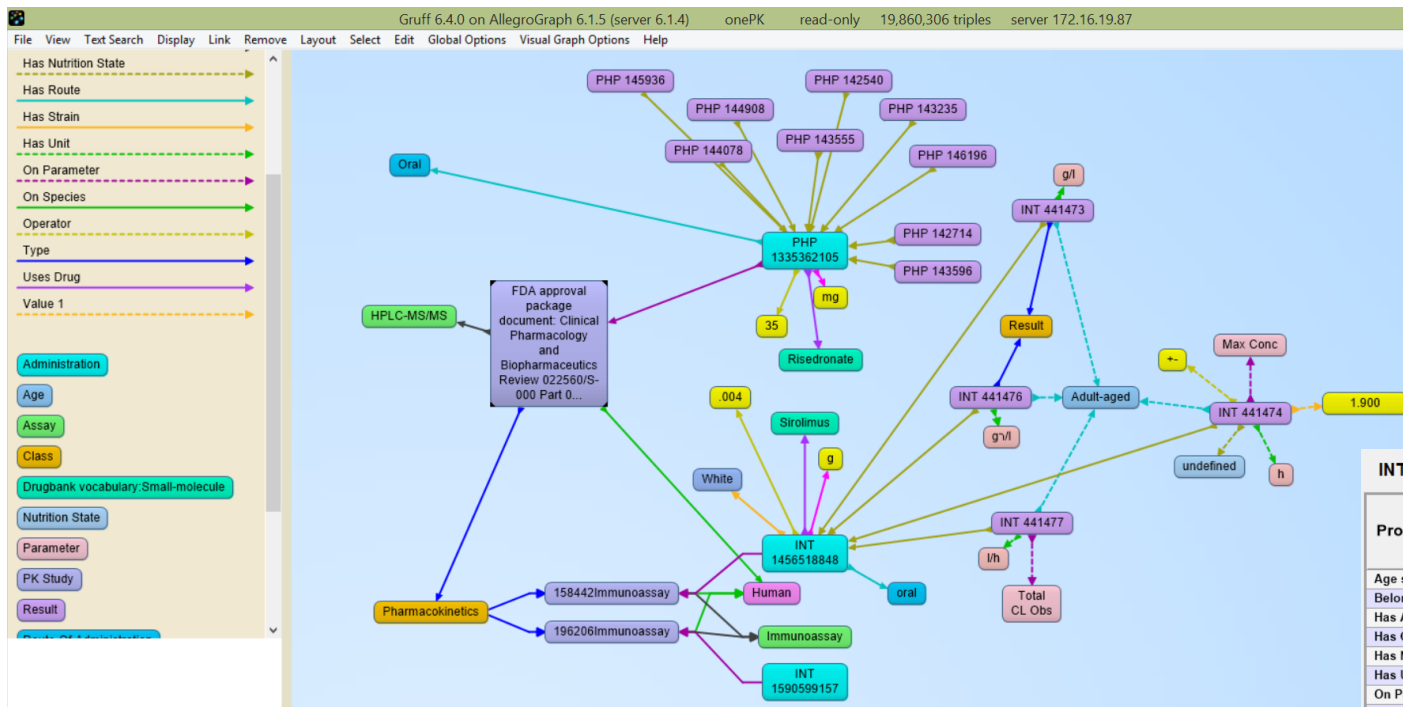
On the fly
ontop
Relational → RDF

ORACLE 12c DATABASE



persist

Allegrograph Gruff on onePK



INT 441476

Property	Value
Age source id	9
Belongs To Administration	INT 1456518848
Has Age	Adult-aged
Has Gender	undefined
Has Nutrition State	undefined
Has Unit	g/l
On Parameter	AUC All
Operator	+-
Result ID	441476
Type	Result
Value 1	.000
Value 2	.000
Weight	undefined

Reporting on Linked Open Data

- onepK demo uses ELDA Epimorphics (Open Source)

- REST based syntax

<http://.../ontologies/onePK/pkstudies>

- Hides SPARQL

http://../ontologies/onePK/administration/PHP_1427975312

Elda Standalone **List**

The screenshot displays the Elda Standalone web interface. On the left, a 'Search Results' panel shows three entries for FDA approval package documents. The first entry is for 'Clinical Pharmacology and Biopharmaceutics Review 021260/S-000, page:11 PDF 1179k' with a hasid of 3234442097. The second entry is for 'Approval Package 087863, page:8 PDF 5931k' with a hasid of 1182050252. The third entry is for 'Clinical Pharmacology and Biopharmaceutics Review 020716, page:7 PDF 736k' with a hasid of 694328565. On the right, an 'Item' view for 'PHP_1427975312' is shown, displaying various properties such as 'admin ID', 'concentration', 'dose_unit', 'dose_value', 'drug_source_id', 'formulation', 'time_duration', 'type', 'admin has study', 'has dose unit', 'has route', 'has strain', and 'uses drug'. The interface includes a search bar, a 'Show Search Form' button, and a 'Sort by' dropdown menu.

Search Results

FDA approval package document: Clinical Pharmacology and Biopharmaceutics Review 021260/S-000, page:11 PDF 1179k

hasid	3234442097
type	PKStudy
has assay	PHP_2870945051
on species	9606
study has administration	PHP_285809240

FDA approval package document: Approval Package 087863, page:8 PDF 5931k

hasid	1182050252
type	PKStudy
has assay	PHP_2870945051
on species	9606
study has administration	PHP_2915238968

FDA approval package document: Clinical Pharmacology and Biopharmaceutics Review 020716, page:7 PDF 736k

hasid	694328565
-------	-----------

On This Page

Results 1 to 10 of 12145

- > FDA approval package document: Clinical
- > FDA approval package document: Approval
- > FDA approval package document: Clinical
- > FDA approval package document: Medical
- > FDA approval package document: Approval
- > FDA approval package document:
- > FDA approval package document: Clinical
- > FDA approval package document: Clinical
- > FDA approval document: ANNEX1, page:39
- > FDA approval package document: Approval
- > next >

Sort by

- label
- hasid
- type
- has assay
- on species
- study has administration

View

> basic > short

Item

PHP_1427975312

http://localhost:8087/ontologies/onePK/administration#PHP_1427975312

admin ID	1427975312
concentration	undefined
dose_unit	ug
dose_value	75
drug_source_id	
formulation	undefined
time_duration	Single
type	Administration
admin has study	PHP_2400385764
has dose unit	C85494_C48152
has route	4945CA56B0D71011E0530201A8C066D3
has strain	495E38E4F55E42E8E0530201A8C03EFF
uses drug	drugbank:DB00586

- Other tools :
 - Poolparty
 - Fluidops
 - Metaphacts

Questions & Answers...

Marc Lieber
Principal consultant
Tel. +41 79 457 97 61
Marc.lieber@trivadis.com

