Discover new Perspectives

Data Federation in the Pharma Industry

Marc Lieber

The Database World is changing ….

hadoop

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

NoSQL Databases

Not only SQL

RDBMS

ORACLE DATABASE

- **Graph Databases**
  - **Property Graphs** — Neo4J, Datastax Cassandra, Oracle
  - **RDF Triple Stores** — Oracle Spatial&Graphs, Allegrograph, Virtuoso, Blazegraph, Marklogic, Enzo
- **Key Value Stores** — Oracle NoSQL, Redis, Riak KV …
- **Wide Column Stores** — Cassandra, Hbase …
- **Document Store Databases** — MarkLogic, MongoDB …

W3C Semantic Web

trivadis makes IT easier.

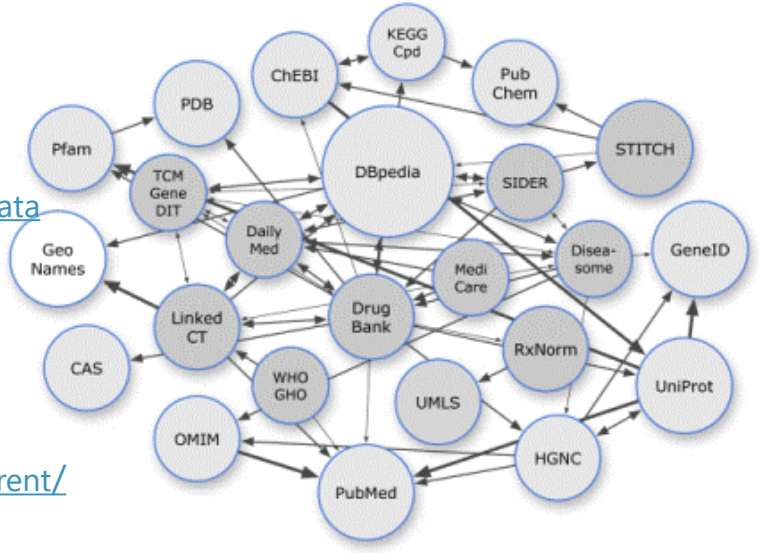
NoSQL Graph Databases

- **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

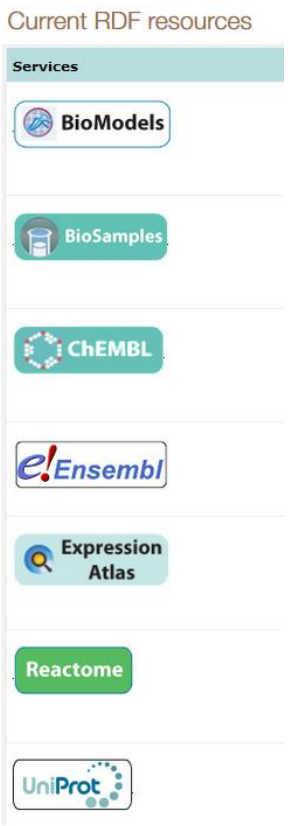
W3C Semantic Web technologies

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**

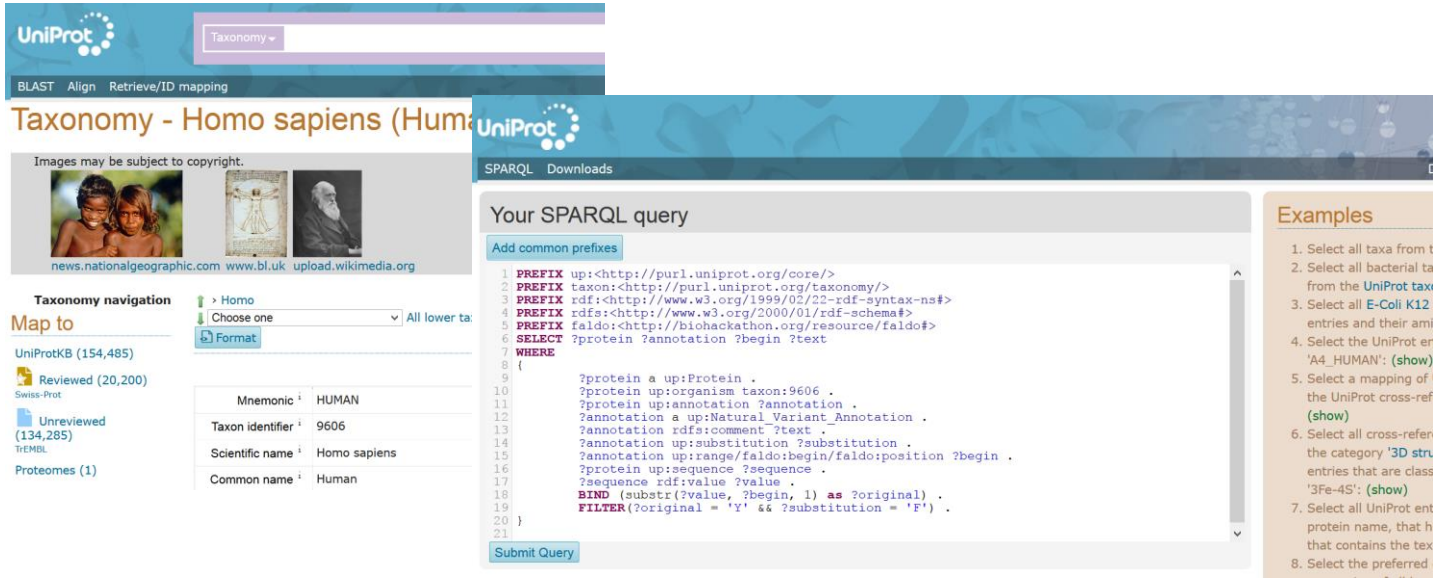
DBPedia http://wiki.dbpedia.org/news/dbpedia-based-rdf-dumps-wikidata
Wikidata https://dumps.wikimedia.org/wikidatawiki/entities/
OpenPHACTS https://www.openphacts.org/
Bio2RDF http://download.bio2rdf.org/current/release.html
LinkedLifeData http://linkedlifedata.com/
Elsevier https://www.elsevier.com/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



Linked Open Date : EBI



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



trivadis makes IT easier.

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

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

QUERY RESULTS

Table | Raw Response

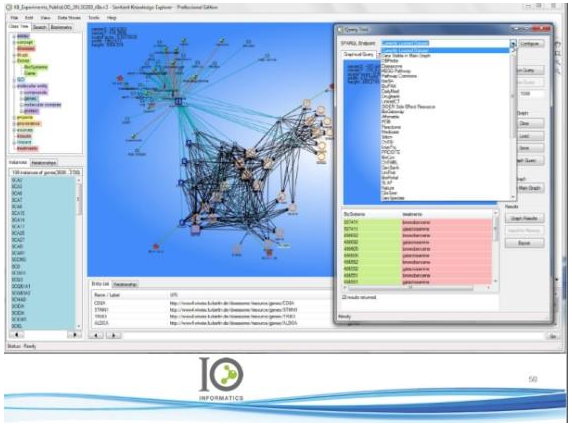
Showing 1 to 2 of 2 entries

Search: Show 50 entries

	protein	uniprot	uname	organism
1	<http://rdf.ncbi.nlm.nih.gov/pubchem/protein/GI124375976>	<http://purl.uniprot.org/uniprot/P10275>	"Androgen receptor"	"Homo sapiens"
2	<http://rdf.ncbi.nlm.nih.gov/pubchem/compound/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



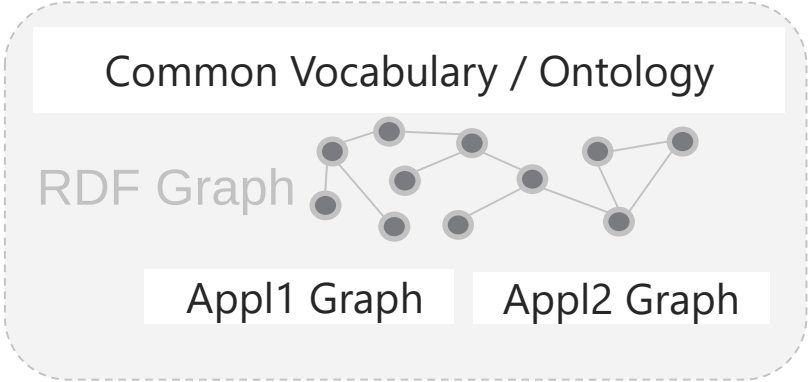
RDF Triple Stores to store the data
SPARQL to query the data, **SPARUL** to change it
Supports Reasoning/Inferencing features based on predefined rules such as OWL,SKOS for data discovery

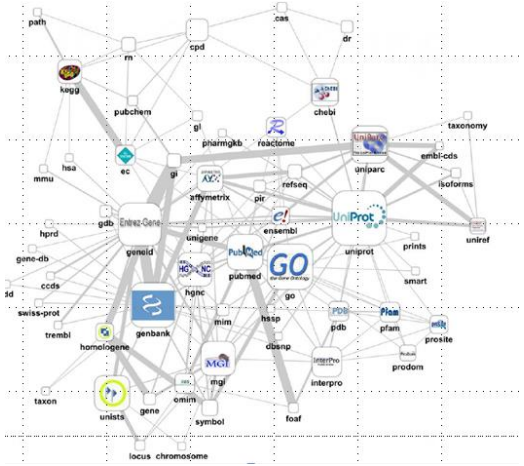
Building an ontology

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

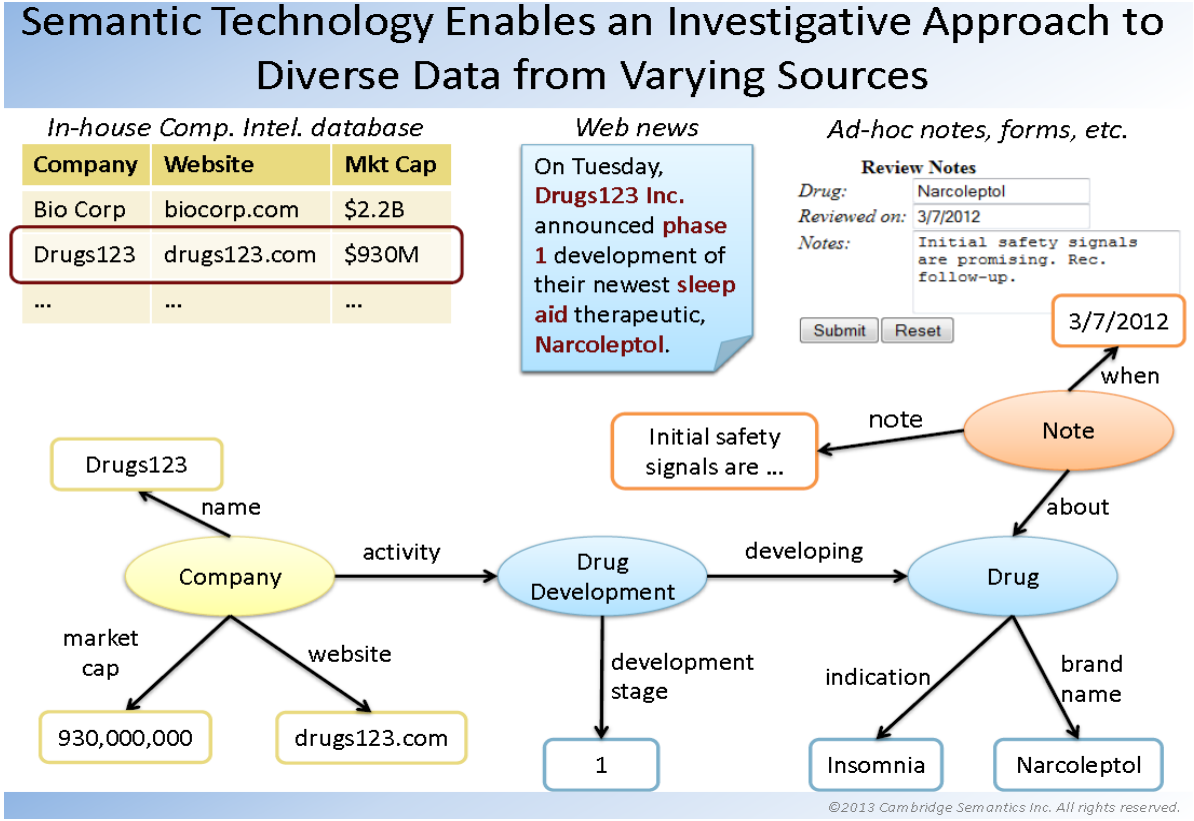
Ontology modeling tools: TopBraid, Protégé





trivadis makes IT easier.

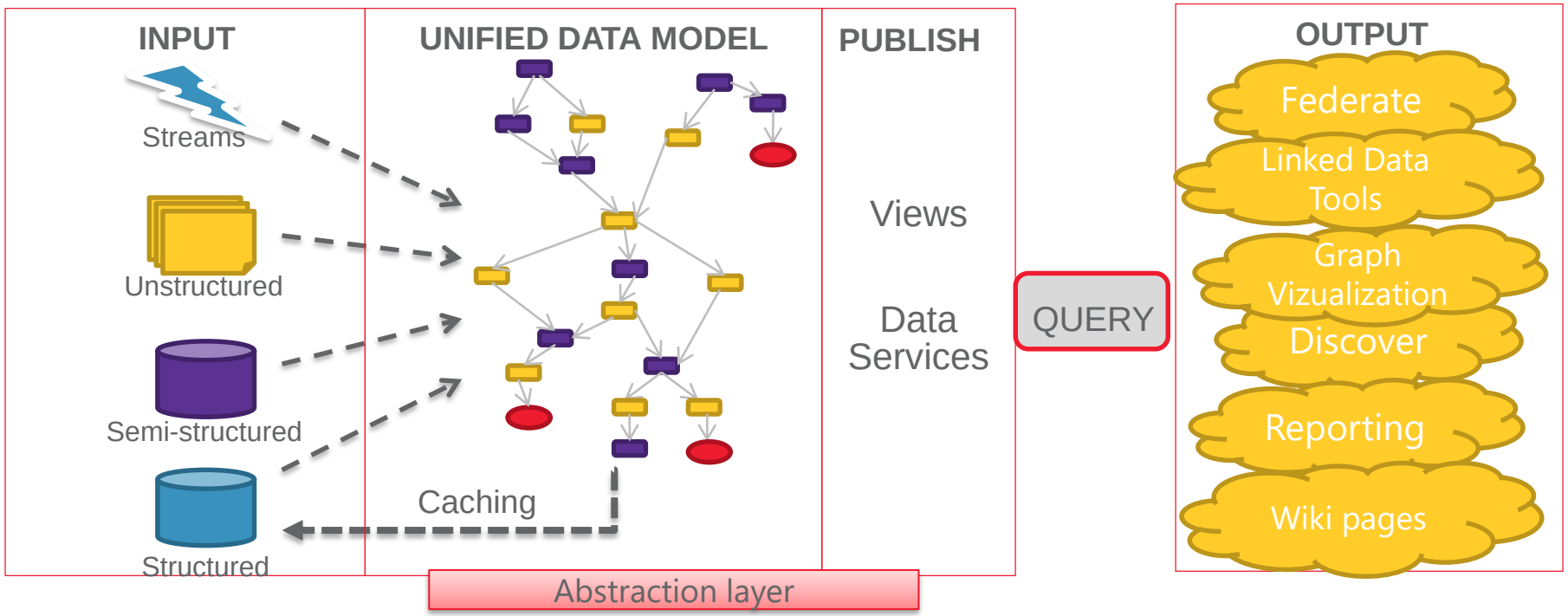
Semantic Web to enrich your information layer



trivadis makes IT easier.

Build a Semantic Metadata Layer

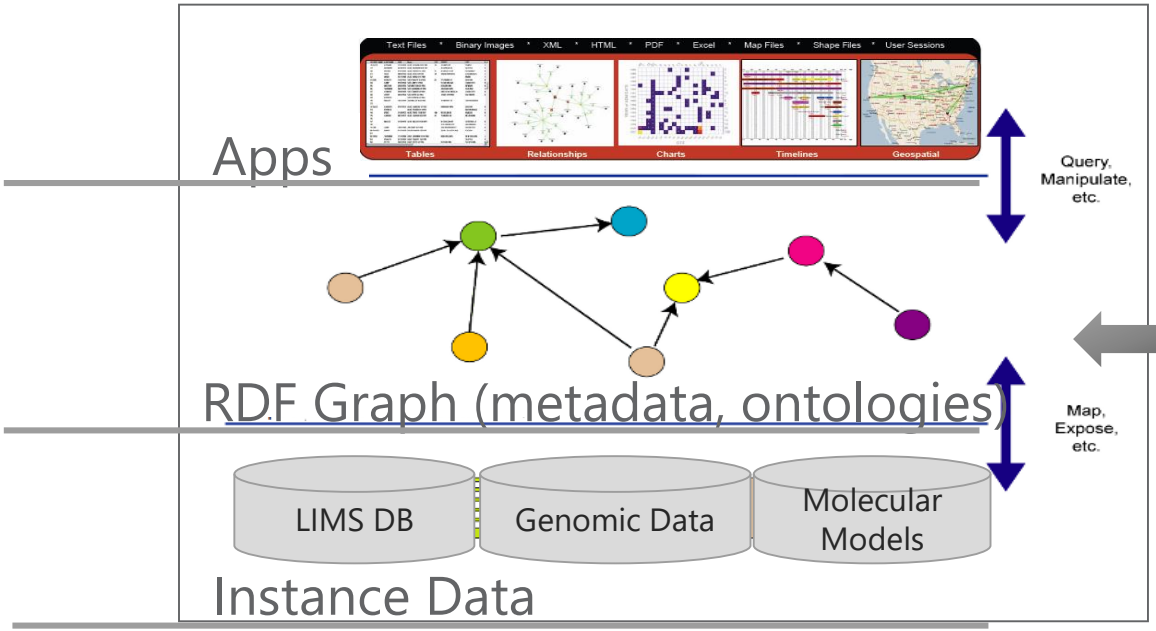
Ingest and semantify, index and query

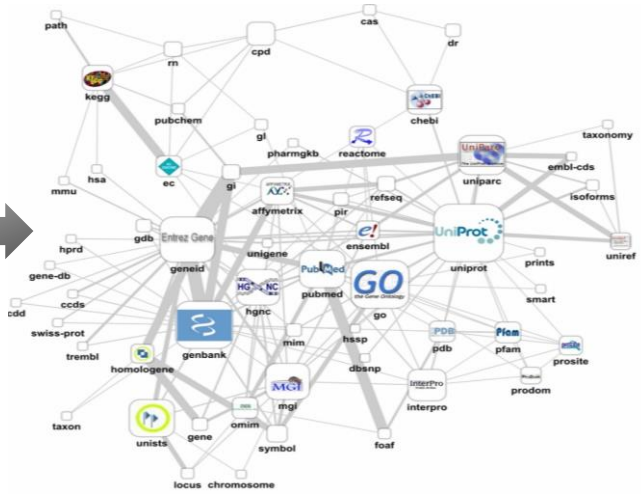


Master Data Management / Reference Data / enterprise vocabulary

- 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 though SPARQL end points, Java APIs, database views or webservices
- 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>

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

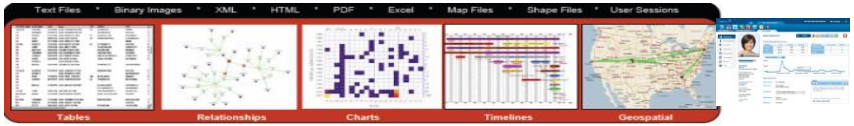
• **New set of services**: extends accessibility and usability of enterprise data
→ The linking of resources enables interoperability between apps that exchange data

• **Better Analysis:** enables discovery of unknown relationships based on semantics; visualization of relationships

Health Care Enterprise metadata framework

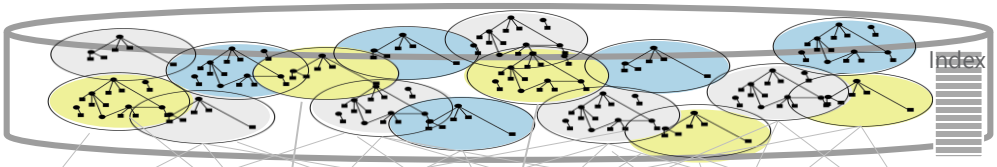
Harmonizing the Electronic Health Care Ecosystem using a Triple store

Enterprise-wide, Patient-centric, longitudinal Record System

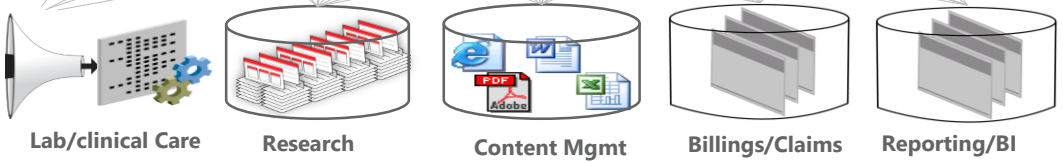


Domain Ontologies

(business metadata + Ontologies)



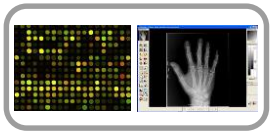
Data Servers



Lab/clinical Care | Research | Content Mgmt | Billings/Claims | Reporting/BI

Data Sources / Data Types

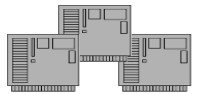
Social Media



Medical Devices

Lab Information Systems

Subscription Services

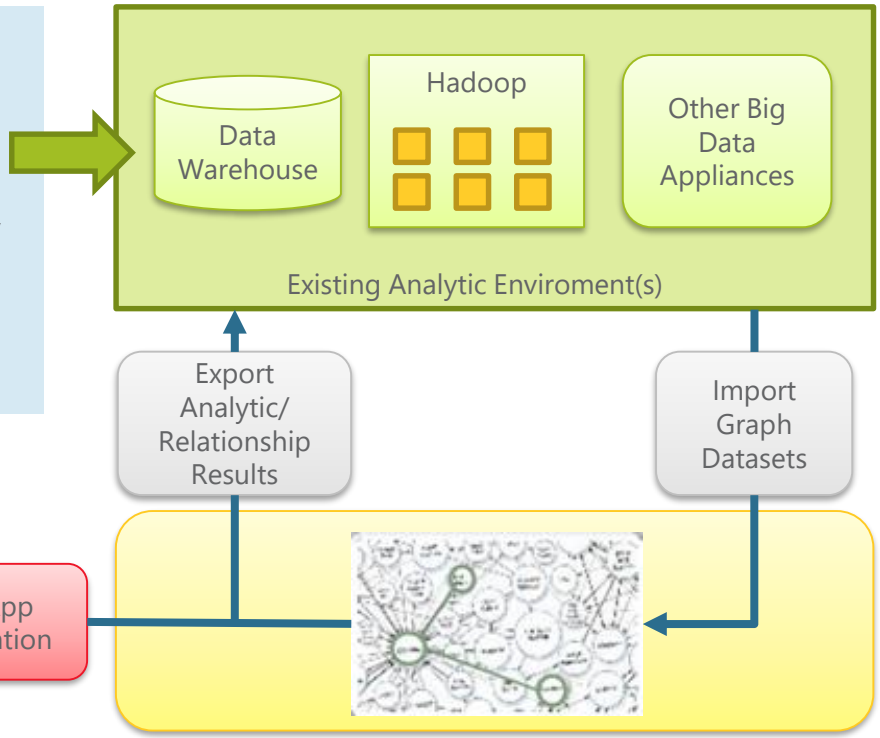


Legacy Patient Records

trivadis makes IT easier.

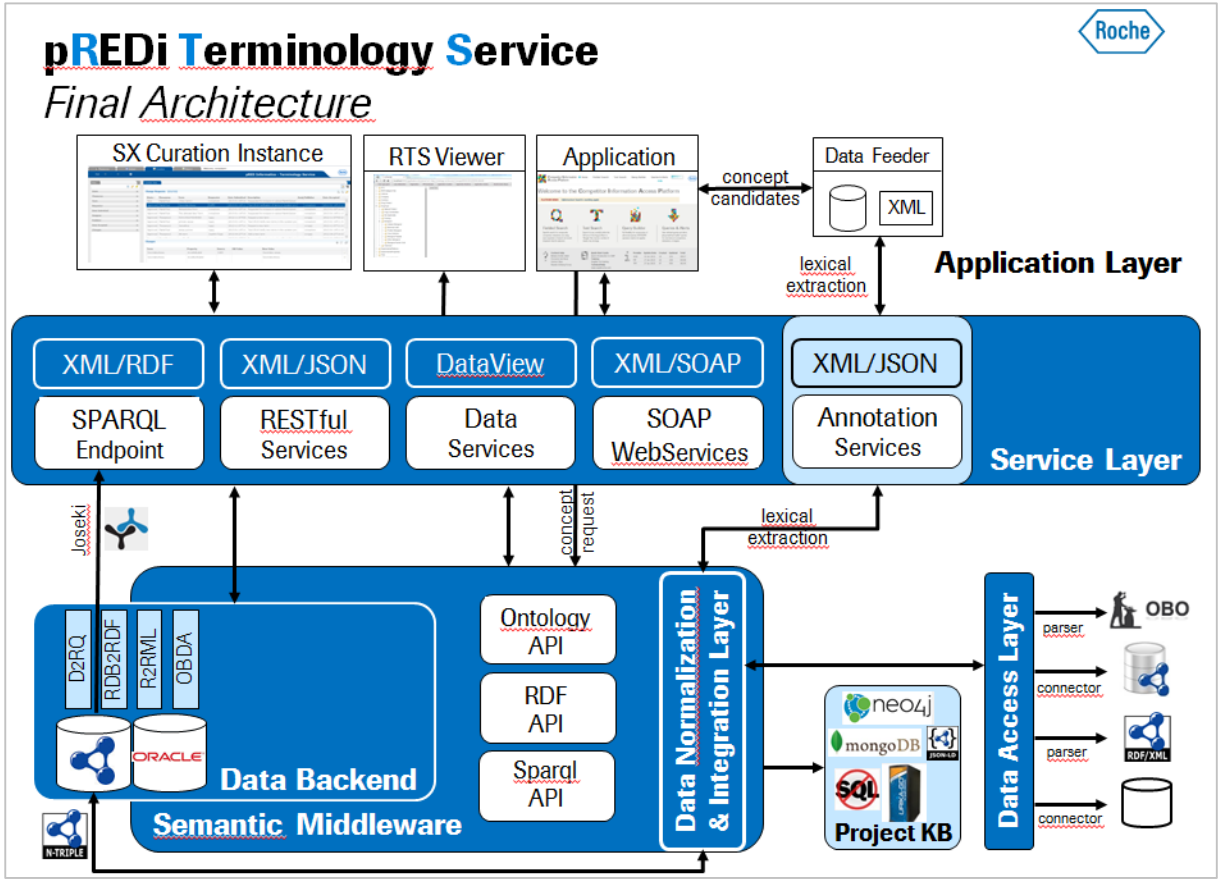
Semantic Web technologies complements Data Warehouse/Hadoop by offloading Graph Analytics

- 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



trivadis makes IT easier.

Semantic Metadata Layer at Roche



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

Semantic search on scientific data

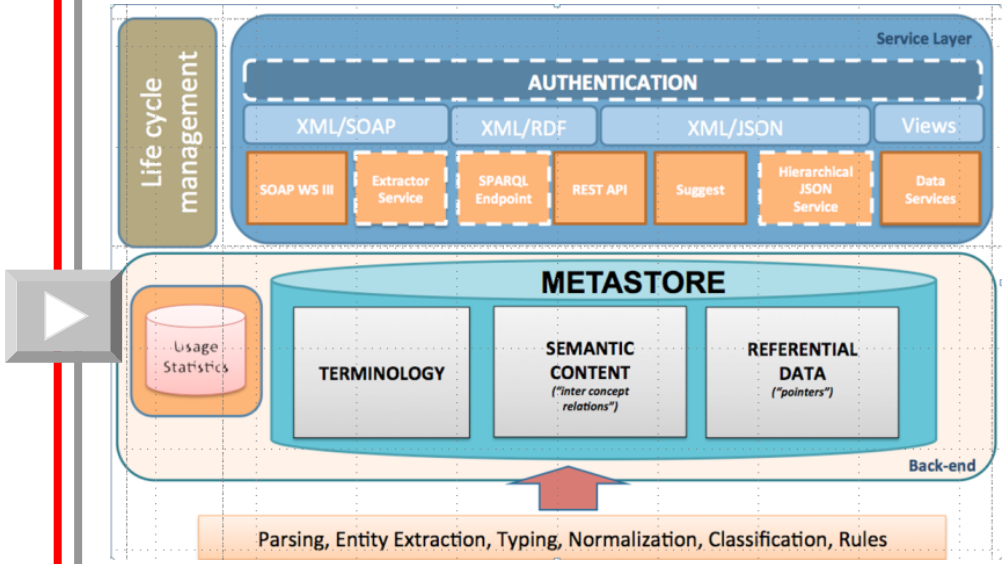
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH

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



trivadis makes IT easier.

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**

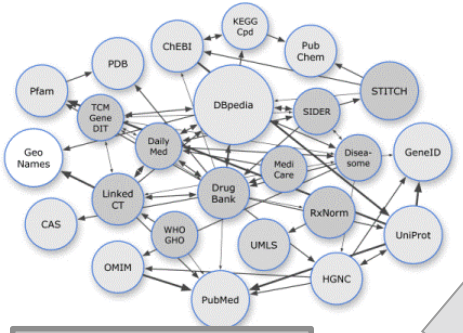
```
SELECT cname, pop, o, neighbor
  FROM TABLE(SEM_MATCH(
    'PREFIX : <http://www.semwebtech.org/mondial/10/meta#>
SELECT ?x ?cname ?pop ?gdp ?o ?neighbor
WHERE {?x rdf:type :Country .
      ?x :name ?cname filter (sameTerm(?cname,"Switzerland")) .
      ?x :population ?pop .
      ?x :gdpTotal ?gdp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

uery Result ×

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

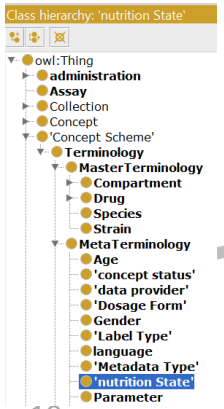
Reporting + vizualisation

ORACLE DATABASE 12c

On the fly

ontop
Relational → RDF

Local DWH



Metamodel ontology

Triple Store

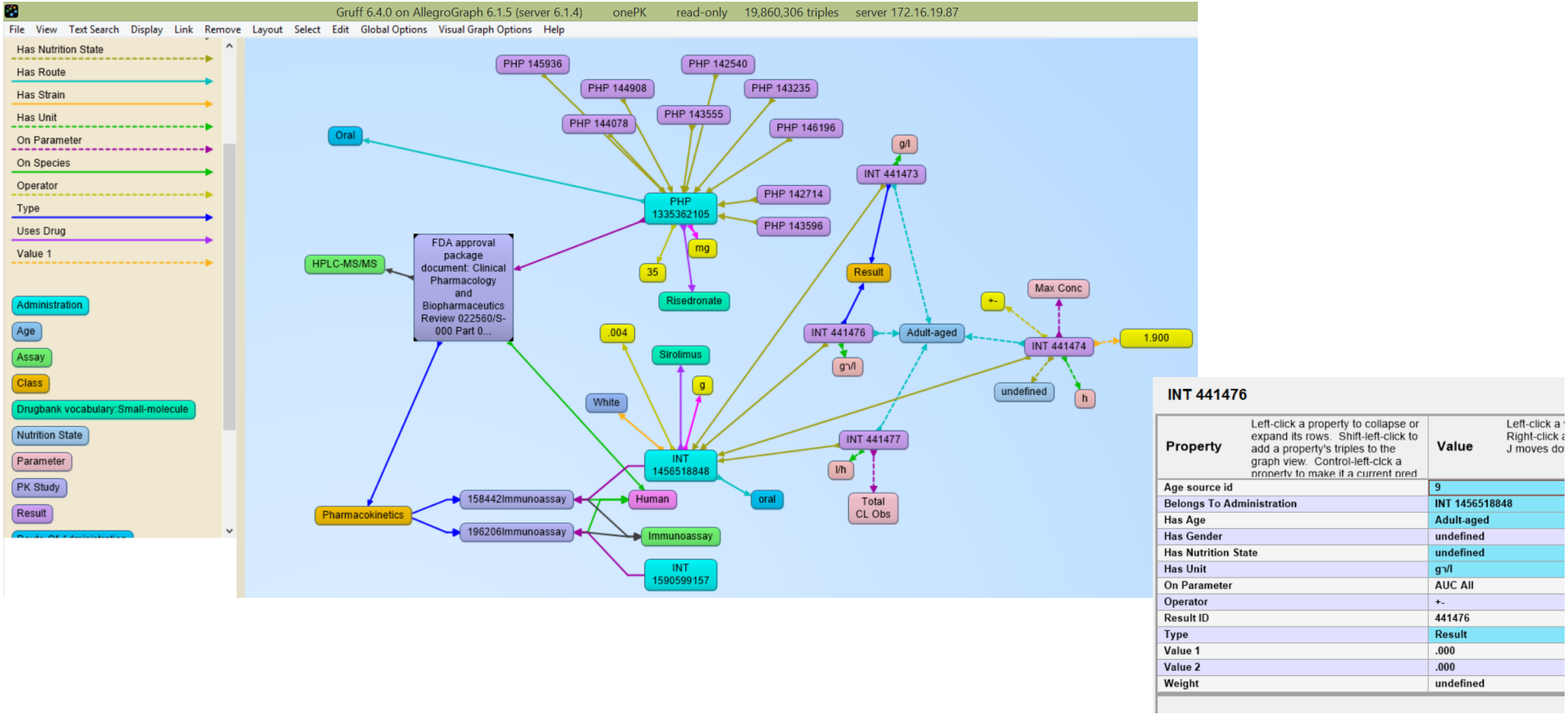
AllegroGraph

persist

Source1 Source2 Source 3

Store n Data sources in RDF + Metamodel

Allegrograph Gruff on onePK



INT 441476

Property	Value
Left-click a property to collapse or expand its rows. Shift-left-click to add a property's triples to the graph view. Control-left-click a property to make it a current pred	Left-click a … Right-click a … J moves do…
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

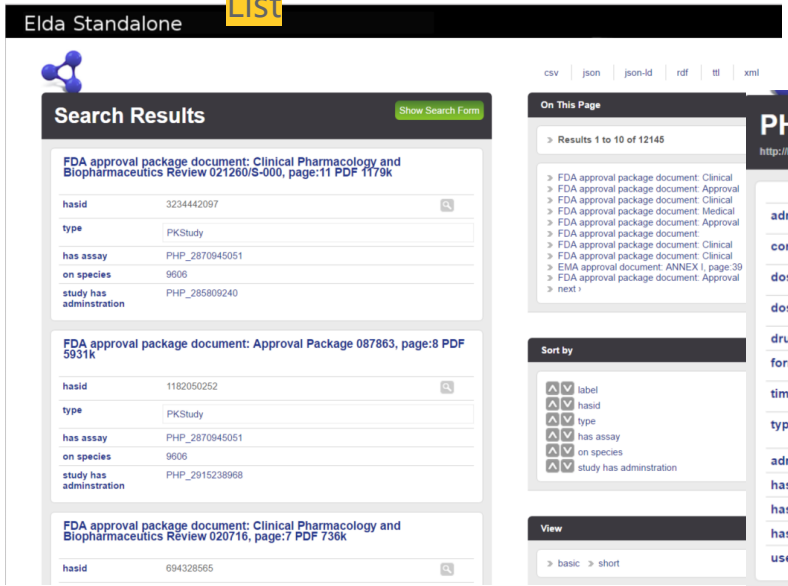
Reporting on Linked Open Data

- onepK demo uses ELDA Epimorphics (Open Source)
  - REST based syntax
  - Hides SPARQL

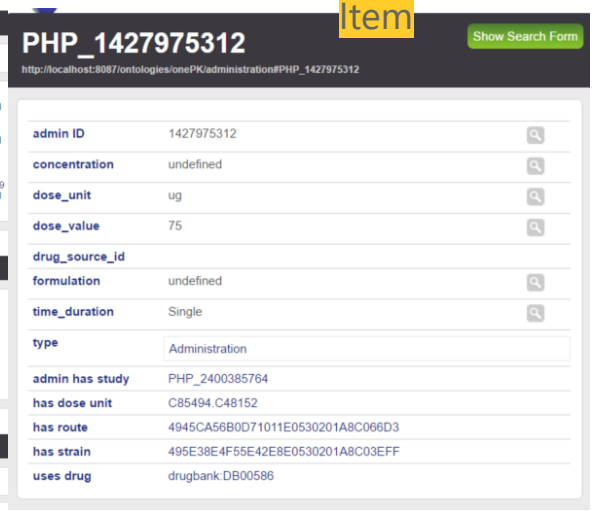
http://.../ontologies/onePK/pkstudies

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

List



Item



- Other tools :
  - Poolparty
  - Fluidops
  - Metaphacts

Questions & Answers…

Marc Lieber
Principal consultant

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

trivadis
makes IT easier.