



Advances and Progress in Drug Design, London

Herman van Vlijmen | Discovery Sciences | 15-16 Feb 2016

Judith Hinton Andrew, *Rock Composite 22*
Artwork from The Creative Center

What is Linked Data?

Linked data

From Wikipedia, the free encyclopedia



WIKIPEDIA
The Free Encyclopedia

In [computing](#), **linked data** (often capitalized as **Linked Data**) is a method of publishing [structured data](#) so that it can be [interlinked](#) and become more useful through [semantic queries](#). It builds upon standard Web technologies such as [HTTP](#), [RDF](#) and [URIs](#), but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. [This enables data from different sources to be connected and queried.](#)^[1]

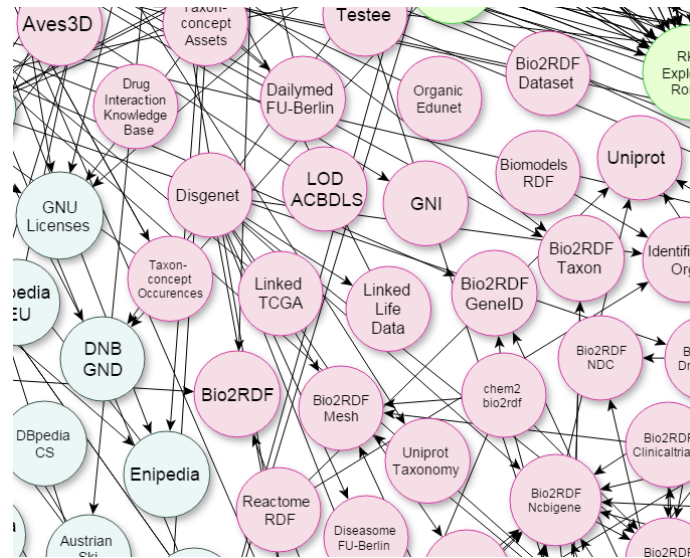
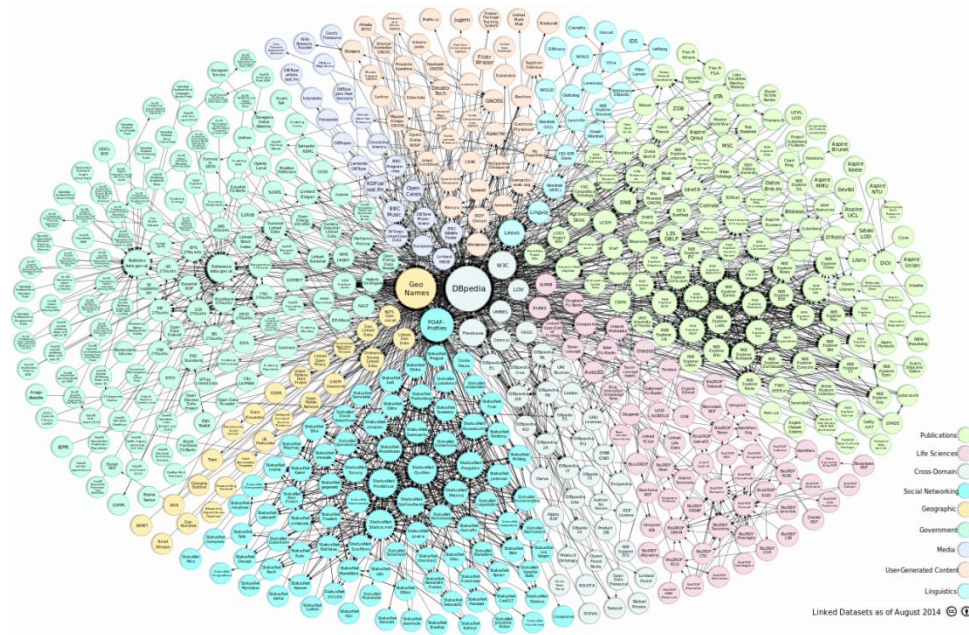
Semantic query

From Wikipedia, the free encyclopedia

Semantic queries allow for queries and analytics of associative and contextual nature. Semantic queries enable the retrieval of [both explicitly and implicitly derived](#) information based on syntactic, semantic and structural information contained in data. They are designed to deliver [precise results](#) (possibly the distinctive selection of one single piece of information) or to answer [more fuzzy and wide open](#) questions through [pattern matching and digital reasoning](#).

Semantic queries work on [named graphs](#), [linked-data](#) or [triples](#). This enables the query to process the actual relationships between information and [infer the answers from the network of data](#). This is in contrast to [semantic search](#), which uses [semantics](#) (the science of meaning) in [unstructured text](#) to produce a better search result (see [Natural language processing](#)).

What is Linked Data?



"LOD Cloud 2014" by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak - <http://lod-cloud.net/>.
Licensed under CC BY-SA 3.0 via Commons - https://commons.wikimedia.org/wiki/File:LOD_Cloud_2014.svg#/media/File:LOD_Cloud_2014.svg

How can data be linked?

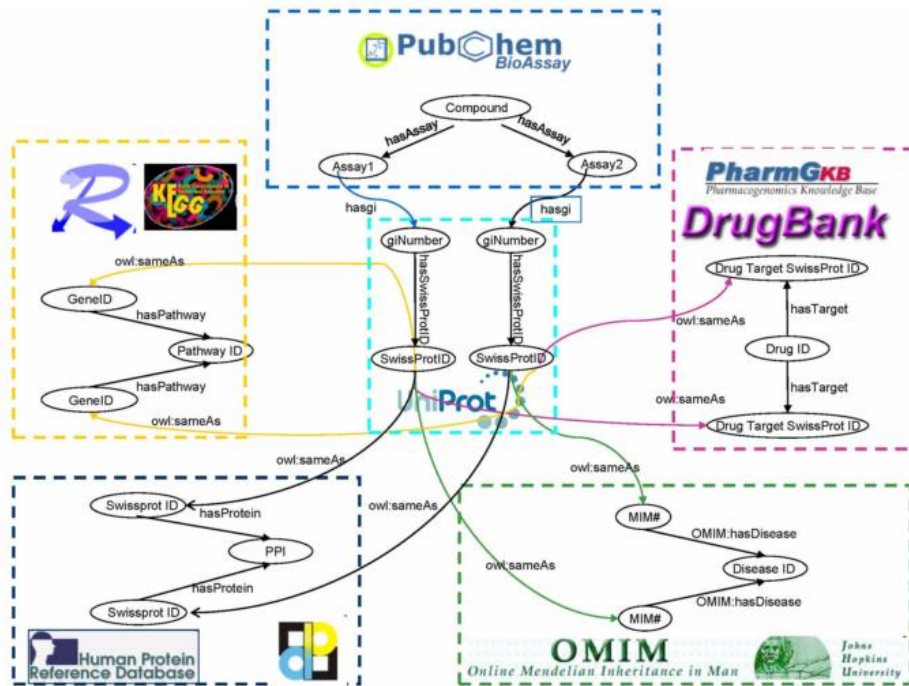
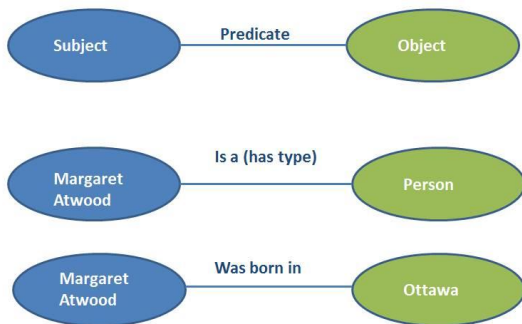


Figure 4 Class links for polypharmacology. Includes the classes: Bioassay, Drug Target, Pathway, Protein-Protein Interaction, and Disease. Some classes include more than one data source. Two nodes in different classes are linked through two paths. For instance, drug *X* is linked to compound *Y* if targets *A* and *B* of drug *X* are linked to assays *A* and *B* of compound *Y* via UNIPROT ID.

- Requires linking to standards:
common “concepts”
 - Names, units, chemical structures, etc
- Data storage format
 - Triples, graphs
- Query tools
 - SPARQL
- Provenance
 - Original data source

RDF triples

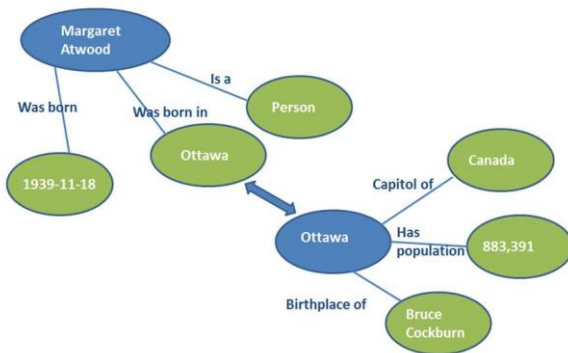
Resource Description Framework (RDF) Triples



Hitchens - LOD & Libraries - Dec. 2013

Basic format

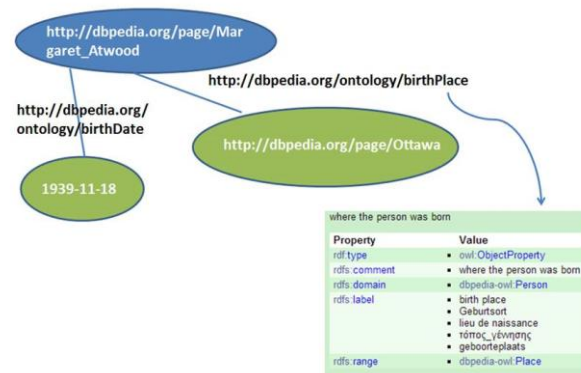
Joining Graphs



Hitchens - LOD & Libraries - Dec. 2013

Linking data sets

Use URIs



Hitchens - LOD & Libraries - Dec. 2013

12

Concept standards

Examples of Linked Data challenges

STANDARD_TYPE	UNIT_COUNT	STANDARD_TYPE	STANDARD_UNITS	COUNT (*)
AC50	7	IC50	nM	829448
Activity	421	IC50	ug.mL-1	41000
EC50	39	IC50	ug/ml	2038
IC50	46	IC50	ug ml-1	509
ID50	42	IC50	mg kg-1	295
Ki	23	IC50	molar ratio	178
Log IC50	4	IC50	ug	117
Log Ki	7	IC50	%	113
Potency	11	IC50	uM well-1	52
log IC50	0	IC50	p.p.m.	51
		IC50	ppm	36
		IC50	uM-1	25
		IC50	nM kg-1	25
		IC50	milliequivalent	22
		IC50	kJ m-2	20

>5000 types

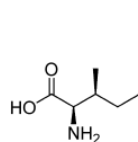
Implemented using the Quantities, Dimension, Units, Types
Ontology (<http://www.qudt.org/>)

~ 100 units

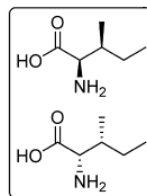
Data types and units for
pharmacological activity in ChEMBL

Names & Taxonomyⁱ

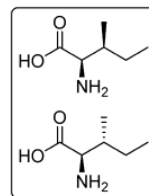
Protein names ⁱ	<p>Recommended name: Tyrosine-protein kinase BTK (EC:2.7.10.2)</p> <p>Alternative name(s):</p> <ul style="list-style-type: none"> Agammaglobulinemia tyrosine kinase <ul style="list-style-type: none"> Short name:ATK B-cell progenitor kinase <ul style="list-style-type: none"> Short name:BPK Bruton tyrosine kinase
Gene names ⁱ	<p>Name:BTk</p> <p>Synonyms:AGMX1, ATK, BPK</p>



Single Known
Stereoisomer

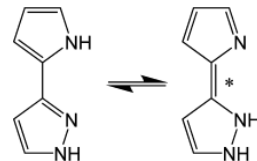


Mixture of
Enantiomers



Mixture of
Diastereomers

Stereochemistry



Tautomerism

Data Standards are Essential

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



Used in Open PHACTS:

Basic Semantic web standards

- SPARQL 1.1, RDF(S), SKOS

Dataset descriptions

- Vocabulary of Interlinked Datasets (VoID)
- VoID linkset descriptions

QUDT

- Quantities, Units, Dimensions and Types

Provenance

- W3C PROV, PAV, Nanopublications

<http://xkcd.com/927/>

Why do we need Linked Data?

- Multiple data sources can be queried at once
 - For example: In-house data, ChEMBL, Pubchem, Thomson-Reuters, DrugBank, GOSTAR, all have compound pharmacology data
 - Time savings
 - Certain to get full picture from private, public, and commercial data
- Complex questions can be asked relatively easily
 - Databases from multiple domains, e.g. compounds, diseases, genes, pathways, etc.
 - Scientists will ask things they would not ask otherwise
- Completely new type of analysis
 - Network based queries, semantic reasoning: not possible previously

Answering more complex questions

Scientific competency questions as the basis for semantically enriched open pharmacological space development

Kamal Azzaoui¹, Edgar Jacoby¹⁴, Stefan Senger², Emiliano Cuadrado Rodríguez³, Mabel Loza³, Barbara Zdrazil⁴, Marta Pinto⁴, Antony J. Williams⁵, Victor de la Torre⁶, Jordi Mestres⁷, Manuel Pastor⁷, Olivier Taboureau⁸, Matthias Rarey⁹, Christine Chichester¹⁰, Steve Pettifer¹¹, Niklas Blomberg^{12,a}, Lee Harland¹³, Bryn Williams-Jones¹³ and Gerhard F. Ecker⁴

Drug Discovery Today • Volume 18, Numbers 17/18 • September 2013

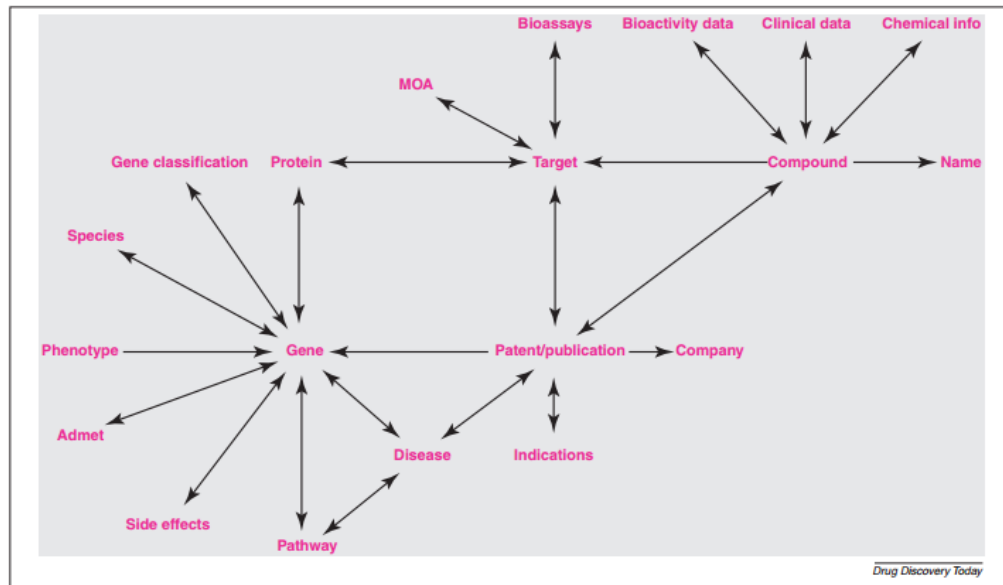


FIGURE 2

Network of data associations needed to answer the top-ranked scientific competency questions. The network reflects a cartoon that summarizes the data associations that are needed to target the top 20 research questions.

TABLE 1

The top 20 research questions

<i>Question number</i>	<i>Question</i>
Cluster I	
Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse
Q2	Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?
Q3	Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives
Q4	For a given interaction profile – give me similar compounds
Q5	The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X
Q6	A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature
Q7	Give me all active compounds on a given target with the relevant assay data
Q8	Identify all known protein–protein interaction inhibitors
Q9	For a given compound, give me the interaction profile with targets
Q10	For a given compound, summarize all ‘similar compounds’ and their activities
Q11	Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)
Cluster II	
Q12	For my given compound, which targets have been patented in the context of Alzheimer’s disease?
Q13	Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?
Q14	Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What

Answering more complex questions

TODAY

What are the Janssen compounds active in this Janssen assay?

What is the difference in gene expression profile between tumor and normal tissue?

I have a CDK2 lead compound. Is there anything known in PubMed on toxicity of CDK2 inhibitors?

WITH LINKED DATA

Give me all **internal/commercial/public** data on compounds that are active on my target **and other closely related** targets.

Given the differences in expression profiles between these tissues, **give me the compounds with biochemical activity profiles that resemble the difference profile most**

Given my CDK2 lead compound, what are the **most likely mechanisms** by which this **compound class** could cause toxicity

New types of analysis with Linked Data

TODAY

Search PubMed for potential target-disease association:
“bcl2 and schizophrenia”

Search a gene disease association database like
DISGENET for possible genes/proteins that can serve as
biomarkers for colorectal cancer

WITH LINKED DATA

Show me **all possible direct and indirect** links between
bcl2 and schizophrenia, **ranked by level of scientific data
support**

Based on **all data that I have access to**, provide a
prioritized list of potential biomarkers for colorectal
cancer **that satisfy specific tissue constraints** and are
obtainable from blood, urine, or stool

New types of analysis with Linked Data:

TODAY

Search PubMed for potential target-disease association:
“bcl2 and schizophrenia”

Search a gene disease association database like
DISGENET for possible genes/proteins that can serve as
biomarkers for colorectal cancer

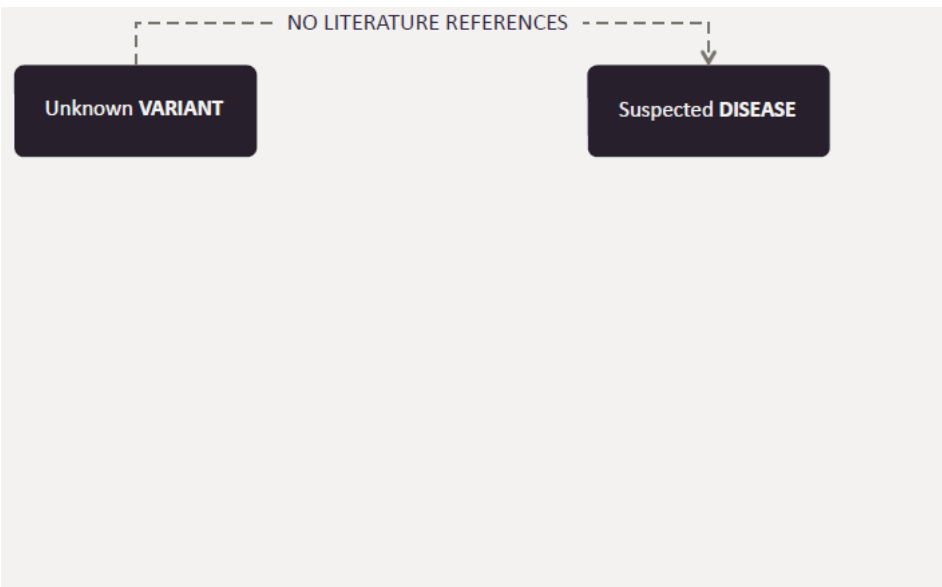
WITH LINKED DATA

Show me **all possible direct and indirect** links between
bcl2 and schizophrenia, **ranked by level of scientific data
support**

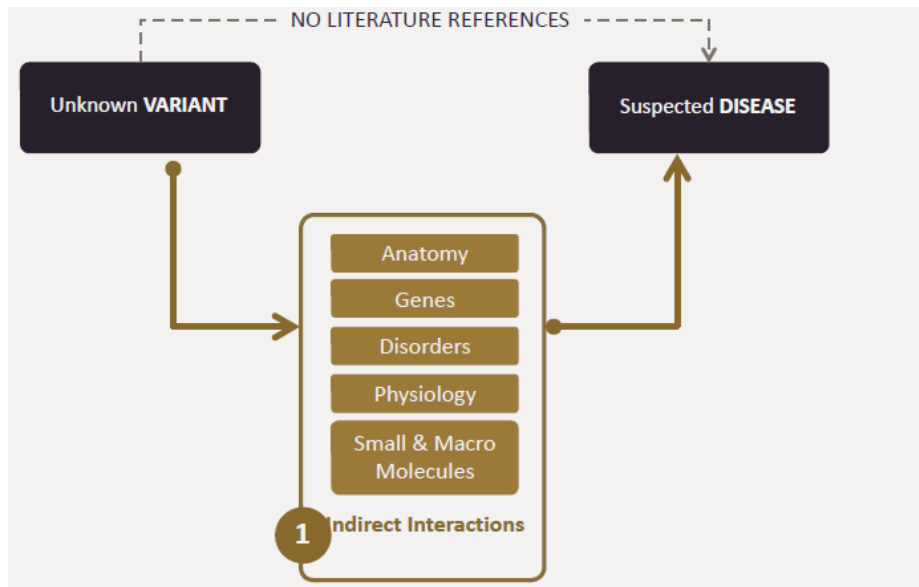
Based on **all data that I have access to**, provide a
prioritized list of potential biomarkers for colorectal
cancer **that satisfy specific tissue constraints** and are
obtainable from blood, urine, or stool

Example:

Gene variant disease association workflow

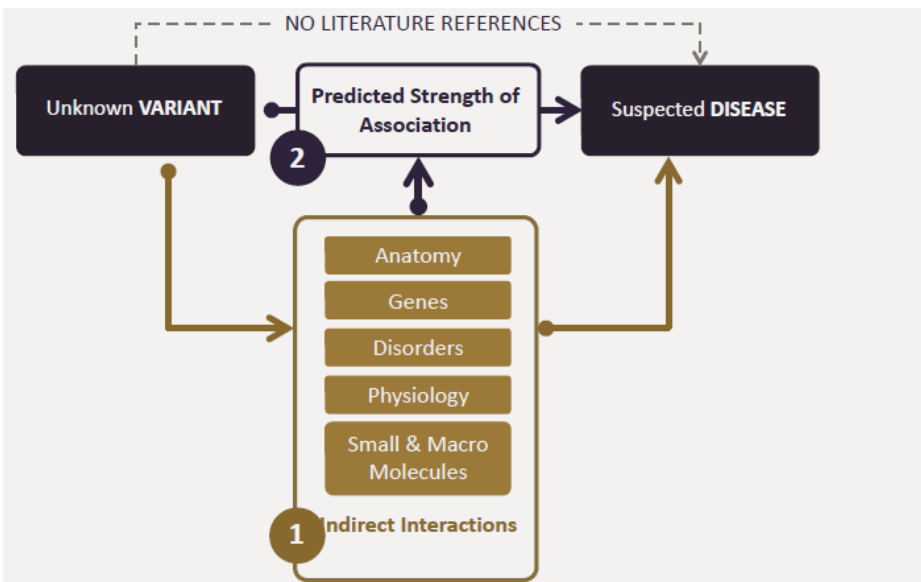


Step 1

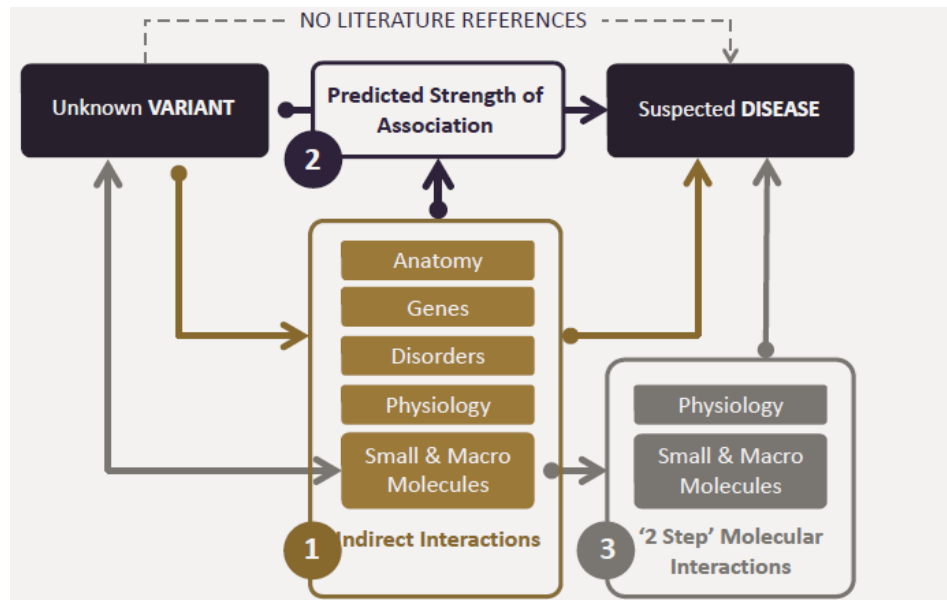


Step 2

Gene variant disease association workflow



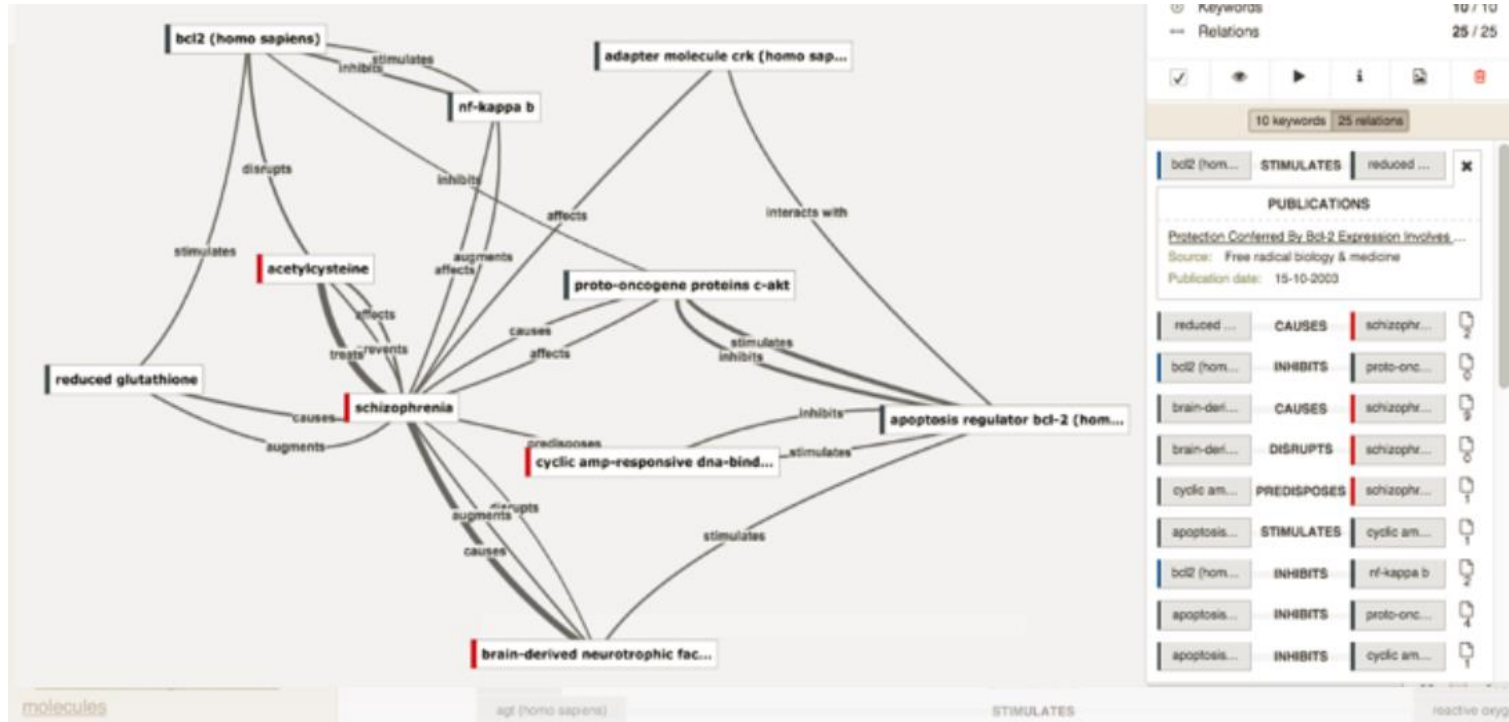
Step 3



Step 4

Gene variant disease association workflow

bcl2 - schizophrenia



What is going on with Linked Data?

Some examples

- World Wide Web Consortium (W3C) continues to develop standards for Semantic Web
- Open PHACTS: EU sponsored IMI project to develop Linked Data database and semantic applications in biomedical field
- ELIXIR: sustainable European infrastructure for biological information. Interoperability of data is key objective
- Development of advanced Linked Data analysis tools
 - For example: Euretoss, Cambridge Semantics, Ontoforce
- Pharma and Biotech companies are actively integrating internal with public and commercial databases with data companies and public-private consortia



Mission: Integrate multiple research biomedical data resources into a single open, sustainable and free access point

The Open PHACTS Foundation is a registered charity dedicated to sustaining and developing the Open PHACTS Discovery Platform after completion of the IMI project

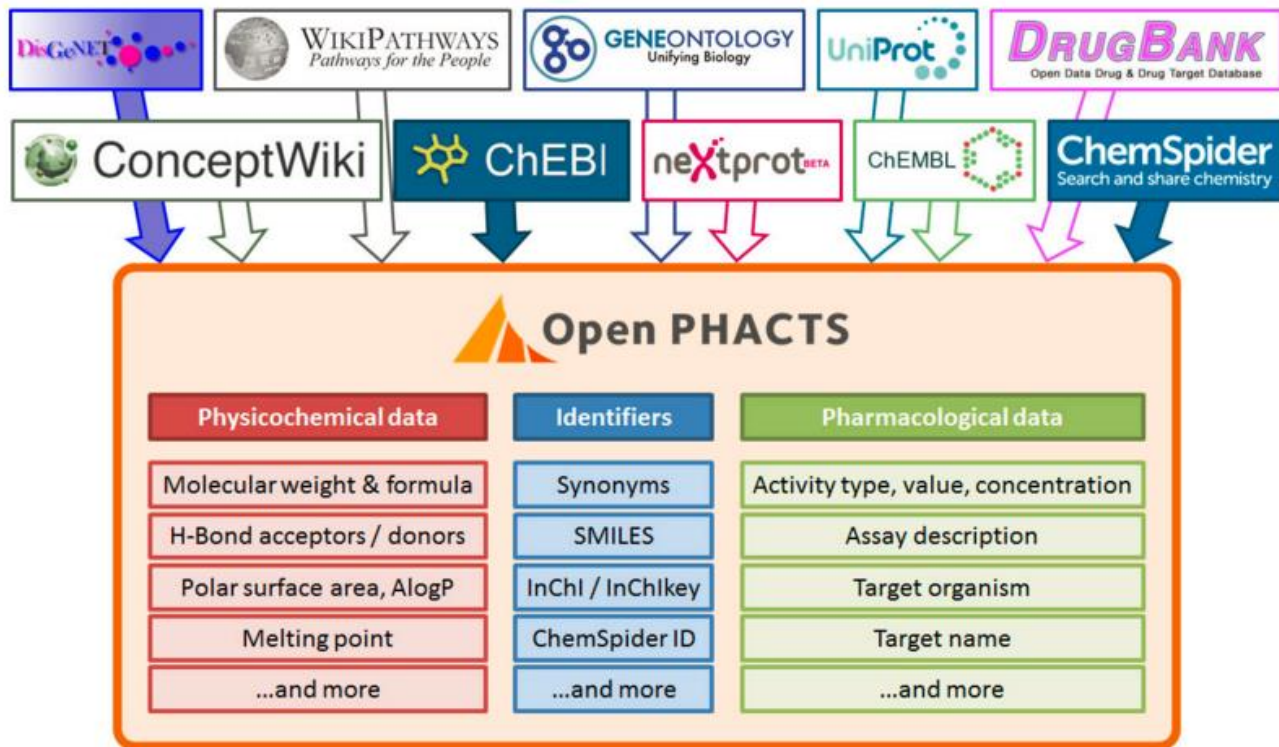
The diagram illustrates the Discovery Platform architecture. At the bottom, 'Public Content' and 'Commercial' sources feed into 'RDF' processing blocks. These blocks are connected to a 'Data Cache (Virtuoso Triple Store)'. The 'Data Cache' is linked to a 'Semantic Workbench' and a 'Linked Data' layer. The 'Linked Data' layer is connected to the 'Discovery Platform'. A 'Data Import' process is shown on the right, feeding into the 'Discovery Platform'. The 'Discovery Platform' is a large box at the top right. The 'Linked Data' layer is a box at the top left. The 'Semantic Workbench' is a box in the middle left. The 'Data Cache' is a box in the middle left. The 'RDF' blocks are at the bottom. The 'Public Content' and 'Commercial' labels are at the bottom. The 'Data Import' label is on the right. The 'Discovery Platform' label is at the top right.

```
PI & Data
```

```
      name :  
      reqiv_target :  
      url_name :  
  
      reqiv_target = cvtStr(target_name);  
      ops.target_orgname target_orgname ;  
      ops.target assay reqiv assay ;  
      void InDataset -> {cvto/cdoto/kamobi.com}  
      ops.target assay owl.inverseof(chembi) hasTarget  
      reqiv assay chembi.orgname Tassay.orgname ;  
      chembi.hasDescription Tassay.description  
      ops.assayofActivity Tassay.activity_url ;  
      ops.assayofActivity owl.inverseof(chembi)andose  
      Activity_url chembi.type Tbio_type ;
```

Apps

Open PHACTS data sources





Explorer

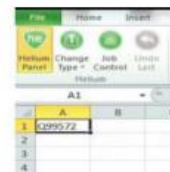
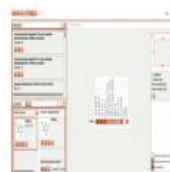
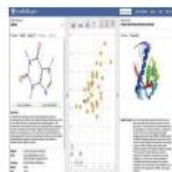
Explorer2

ChemBioNavigator

Target Dossier

Pharmatrek

Helium



MOE

Collector

Cytophacts

Utopia

Garfield

SciBite



KNIME

Mol. Data Sheets

PipelinePilot

scinav.it

Taverna



Applications that use
the Open PHACTS API

<http://www.openphactsfoundation.org/apps.html>

Open PHACTS consortium partners



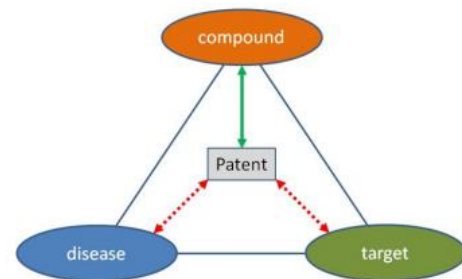
Consortium partners



Associated partners

Recent Open PHACTS developments: Patent Info

- Huge amount of knowledge in patent corpus, most of which will never be published elsewhere, but potentially great value to drug discovery
- SureChEMBL system (EBI) already extracts compounds from these documents
- Open PHACTS consortium funded project to also extract gene/disease information (EMBL-EBI and SciBite)
- ~4 million patents in total, 260 million annotations (patent-compound, patent-gene or patent-disease associations)
- Example use cases:
 - For a given target, give me all the compounds that are linked to this target through patents
 - For a given disease, give me all the targets that are linked to this disease through patents
 - Tell me how reliable these links are



Acknowledgements

- Janssen
 - Edgar Jacoby
 - Jean-Marc Neefs
 - Dmitrii Rassokhin
- Open PHACTS and Open PHACTS Foundation
- Euretos
 - Albert Mons
 - Arie Baak

Backup slides