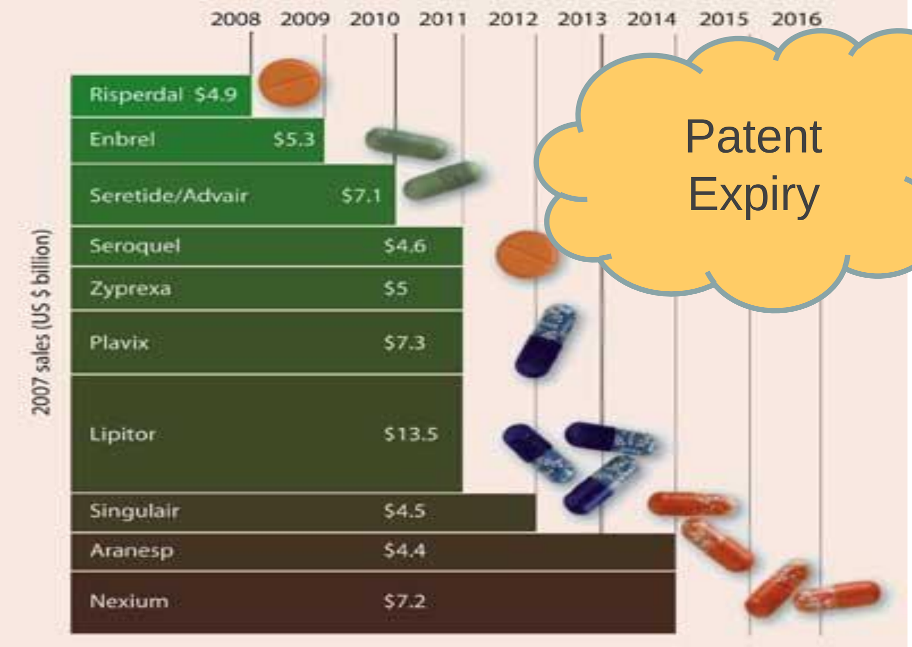
Open PHACTS
Linked Open Data for Drug Discovery

Herman van Vlijmen – Janssen Pharmaceutica
Orri Erling – OpenLink Software

Linked Open Data congress, Hilversum, 25 June 2014

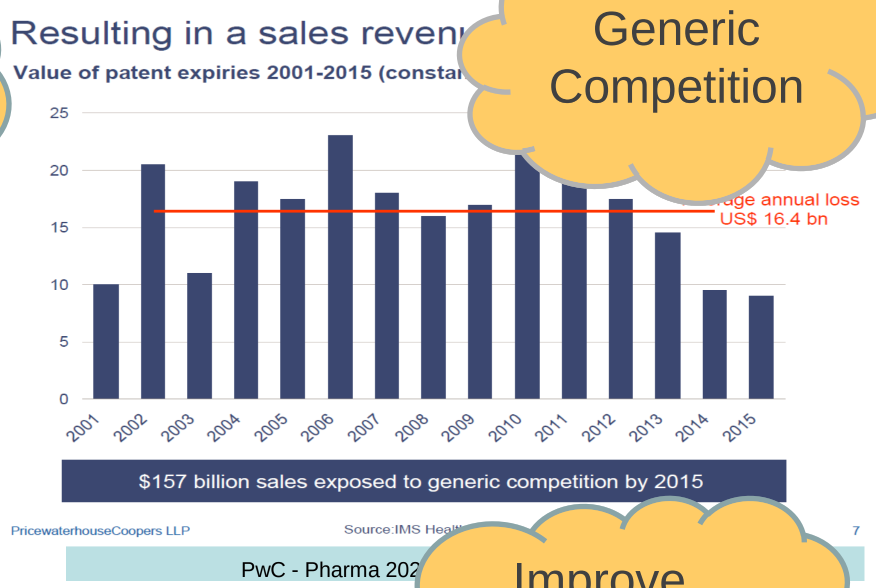
Open PHACTS
Open Pharmacological Space

The Pharma Industry
Challenges in a Changing Landscape



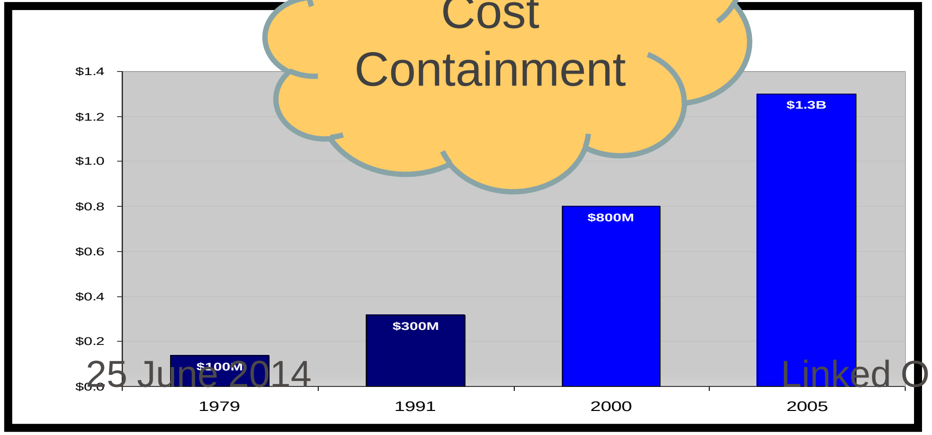
Patent Expiry

http://www.rsc.org/chemistryworld/Is … esOnThePatentCliff.asp

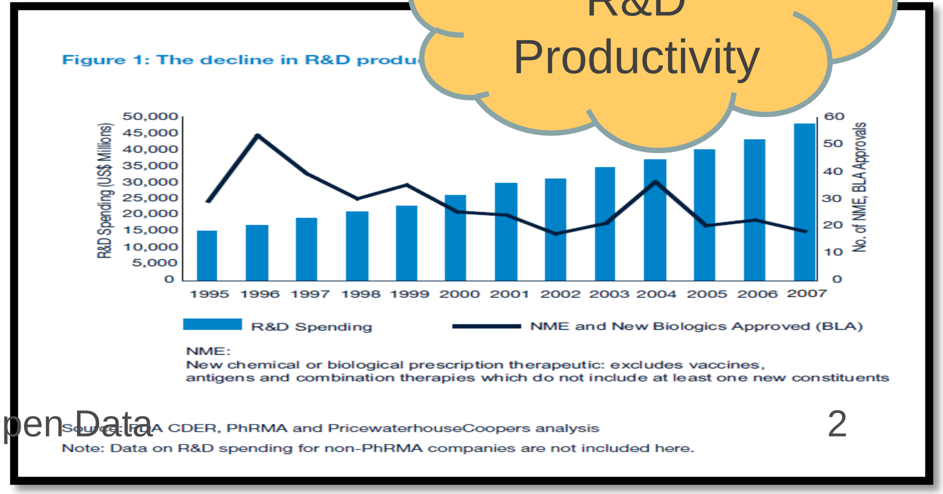


Generic Competition

Cost Containment



Improve R&D Productivity

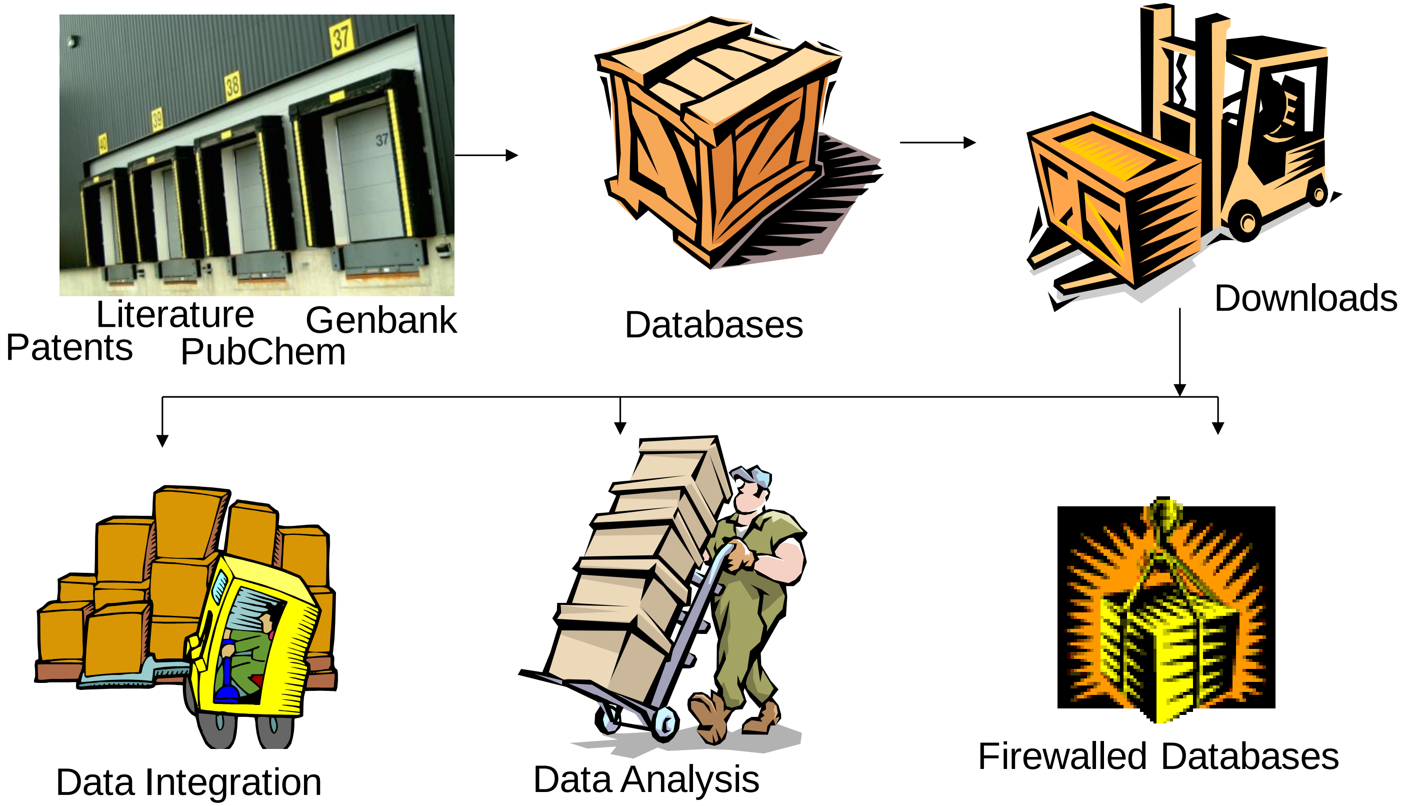


NME:
New chemical or biological prescription therapeutic: excludes vaccines, antigens and combination therapies which do not include at least one new constituents

Source: FDA CDER, PhRMA and PricewaterhouseCoopers analysis
Note: Data on R&D spending for non-PhRMA companies are not included here.

Public Domain Drug Discovery Data:

Pharma are accessing, processing, storing & re-processing



Literature

Genbank

Patents

PubChem

Databases

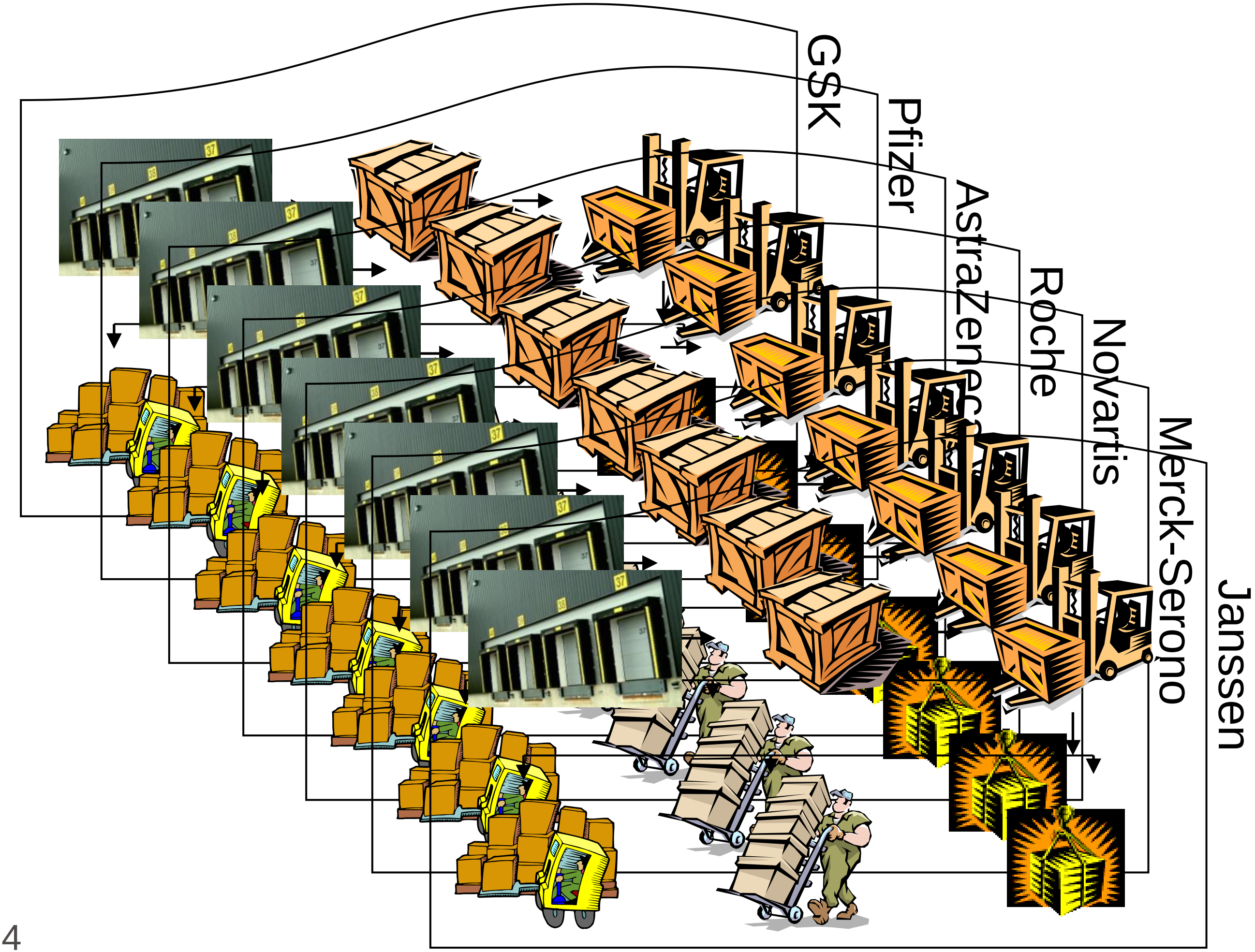
Downloads

Data Integration

Data Analysis

Firewalled Databases

We are all doing this many times……



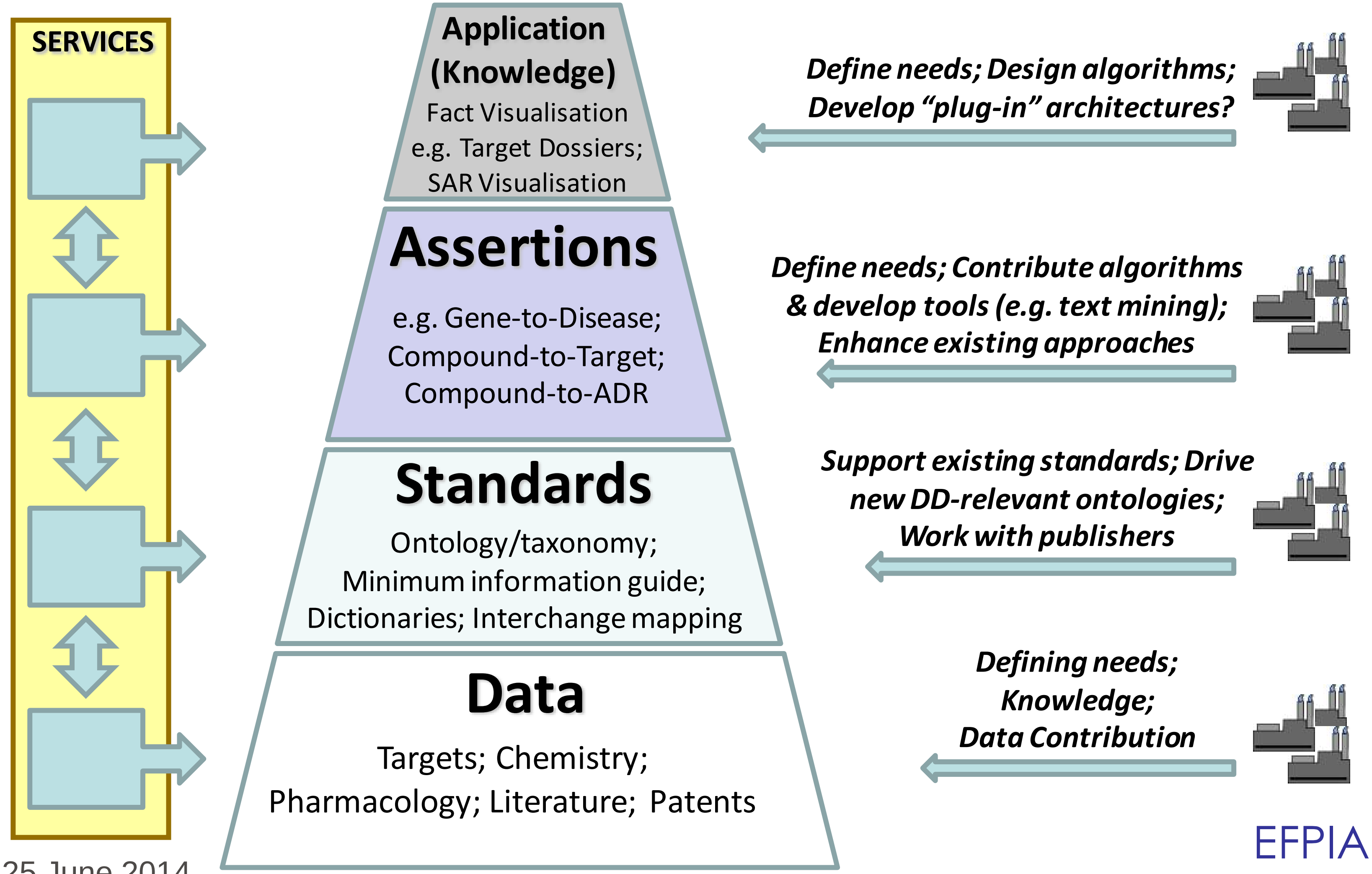
The Open PHACTS Project

Open PHACTS
Open Pharmacological Space

- Create a *semantic integration hub (“Open Pharmacological Space”)*
- Delivering services to support on-going drug discovery programs in pharma and public domain
- *Not just another project*; Leading academics in semantics, pharmacology and informatics, driven by solid industry business requirements
- 16 academic partners, 9 pharmaceutical companies, 4 biotechs
- Work split into clusters:
  - Technical Build: Create the technology
  - Scientific Drive: Develop use cases and exemplar applications
  - Community & Sustainability: Engage community and build the future

OPS Components

Open PHACTS
Open Pharmacological Space



SERVICES

Application (Knowledge)
Fact Visualisation
e.g. Target Dossiers;
SAR Visualisation

Assertions
e.g. Gene-to-Disease;
Compound-to-Target;
Compound-to-ADR

Standards
Ontology/taxonomy;
Minimum information guide;
Dictionaries; Interchange mapping

Data
Targets; Chemistry;
Pharmacology; Literature; Patents

Define needs; Design algorithms;
Develop "plug-in" architectures?

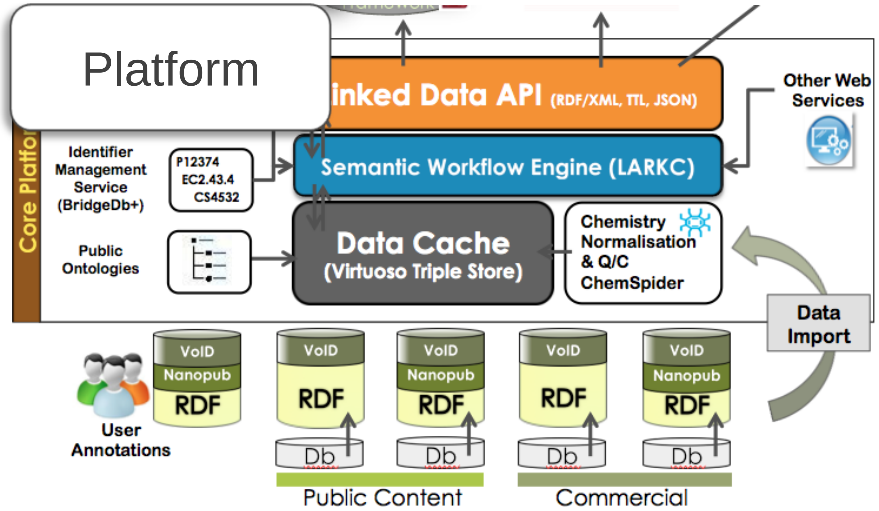
Define needs; Contribute algorithms
& develop tools (e.g. text mining);
Enhance existing approaches

Support existing standards; Drive
new DD-relevant ontologies;
Work with publishers

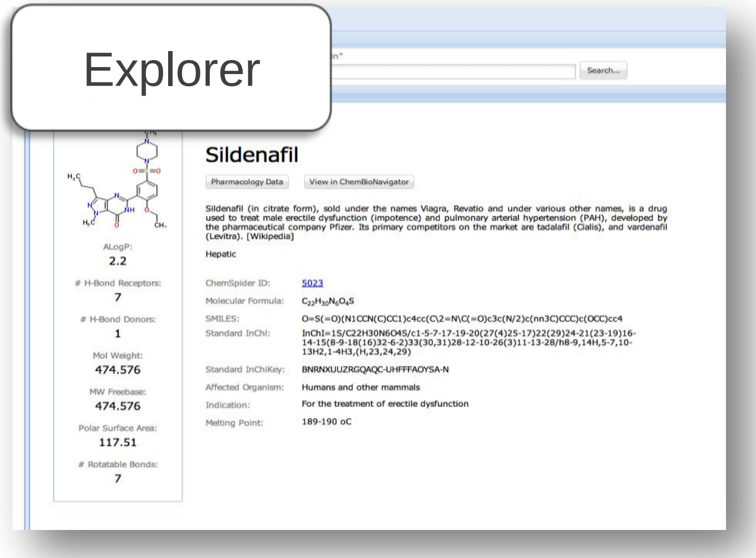
Defining needs;
Knowledge;
Data Contribution

EFPIA

25 June 2014


Platform
Linked Data API (RDF/XML, TTL, JSON)
Other Web Services
Core Platform
Identifier Management Service (BridgeDb+)
P12374
EC2.43.4
CS4532
Semantic Workflow Engine (LARKC)
Public Ontologies
Data Cache (Virtuoso Triple Store)
Chemistry Normalisation & Q/C ChemSpider
Data Import
User Annotations
VoID
Nanopub
RDF
Db
Public Content
Commercial

Explorer



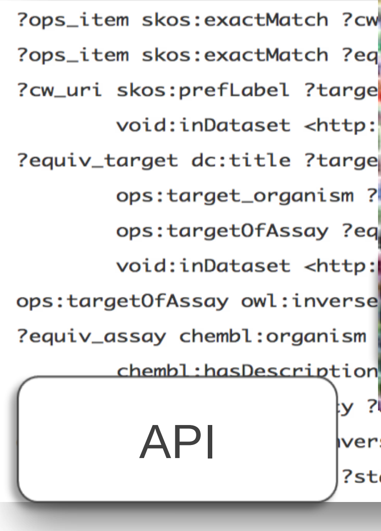
Sildenafil

Sildenafil (in citrate form), sold under the names Viagra, Revatio and under various other names, is a drug used to treat male erectile dysfunction (impotence) and pulmonary arterial hypertension (PAH), developed by the pharmaceutical company Pfizer. Its primary competitors on the market are tadalafil (Cialis), and vardenafil (Levitra). [Wikipedia]

Hepatic

ChemSpider ID:	5023
Molecular Formula:	$C_{22}H_{30}N_6O_4S$
SMILES:	O=S(=O)(N1CCN(C)CC1)c4cc(C\2=N\C(=O)c3c(N/2)c(nn3C)CCC)c(OCC)cc4
Standard InChI:	InChI=1S/C22H30N6O4S/c1-5-7-17-19-20(27(4)25-17)22(29)24-21(23-19)16-14-15(8-9-18(16)32-6-2)33(30,31)28-12-10-26(3)11-13-28/h8-9,14H,5-7,10-13H2,1-4H3,(H,23,24,29)
Standard InChiKey:	BNRNXUUZRGQAQC-UHFFFAOYSA-N
Affected Organism:	Humans and other mammals
Indication:	For the treatment of erectile dysfunction
Melting Point:	189-190 oC

ALogP:	2.2
# H-Bond Receptors:	7
# H-Bond Donors:	1
Mol Weight:	474.576
MW Freebase:	474.576
Polar Surface Area:	117.51
# Rotatable Bonds:	7



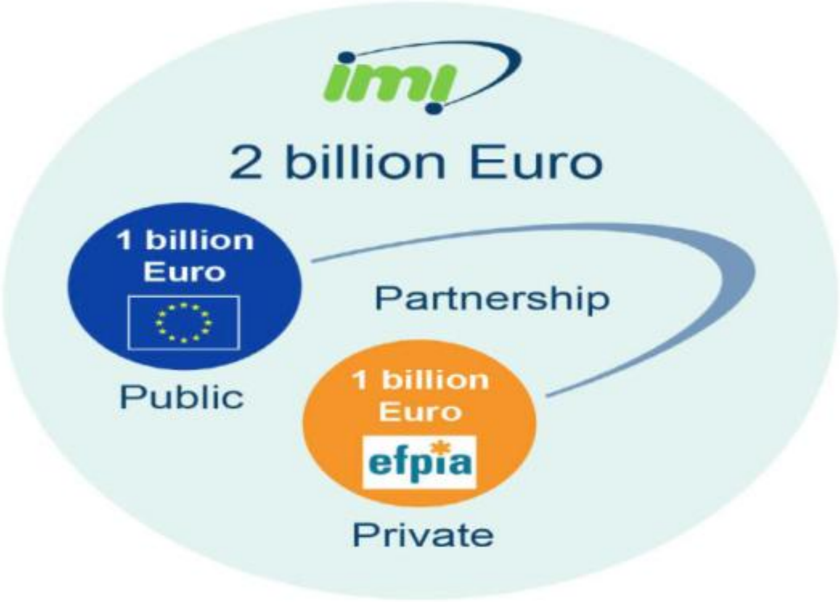
```
?ops_item skos:exactMatch ?cw
?ops_item skos:exactMatch ?eq
?cw_uri skos:prefLabel ?targe
        void:inDataset <http:
?equiv_target dc:title ?targe
        ops:target_organism ?
        ops:targetOfAssay ?eq
        void:inDataset <http:
ops:targetOfAssay owl:inverse
?equiv_assay chembl:organism
        chembl:hasDescription
                            y ?
                     verseOf chembl:onAssa
                     ?std_type ;
```

API

Apps

Standards

IMI: The Innovative Medicines Initiative



- Biggest public-private partnership in area of medicine
- Collaboration between European Commission and European Federation of Pharmaceutical Industries and Associations (EFPIA)
- Promotion of medical innovation in Europe
- Tackle key bottlenecks
- Recognises “in kind” contributions
- Focus on key problems
  - Efficacy, Safety, Education & Training, Knowledge Management

Project Partners

Universität Wien

Technical University of Denmark

University of Hamburg, Center for Bioinformatics

BioSolveIT GmBH

Consorci Mar Parc de Salut de Barcelona

Leiden University Medical Centre

Royal Society of Chemistry

Vrije Universiteit Amsterdam

Spanish National Cancer Research Centre

University of Manchester

Maastricht University

Aqnowledge

University of Santiago de Compostela

Rheinische Friedrich-Wilhelms-Universität Bonn

Netherlands Bioinformatics Centre

Swiss Institute of Bioinformatics

ConnectedDiscovery

EMBL-European Bioinformatics Institute

OpenLink Software

Open PHACTS Foundation

Pfizer

Novartis

Merck Serono

H. Lundbeck A/S

Eli Lilly

Janssen

AstraZeneca

GlaxoSmithKline

Esteve

Associate Partners

Open PHACTS
Open Pharmacological Space

orphanet

Global Health Equity Foundation
research · advocacy · capacity building

VIVO connect share discover

Concept Web Alliance

Numedicus

thehyve

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

INDIANA UNIVERSITY

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

Ian Harrow Consulting

Erasmus MC

Parc Científic de Barcelona

eagle genomics

Sage BIONETWORKS

mobileflow

ACD/Labs

MGH 1811

codeN

Genetta SOFT

CERTARA
Implementing Translational Science

AUREUS SCIENCES

GVK BIO
Accelerating Research

Sequenomics

accelrys

IUPHAR DATABASE
International Union of Basic and Clinical Pharmacology

THOMSON REUTERS

A use-case driven approach, focussed on delivery for the real world

- Main architecture, technical implementation and primary capabilities driven by a set of **prioritised research questions**
- Based on the main research questions define **prioritised data sources**
- Develop **three Exemplars** to demonstrate the capabilites of the Open PHACTS System and to define interfaces and input/output standards

What do we need?

"All oxidoreductase inhibitors active <100nM in both human and mouse"

"Compounds that agonize targets in pathway X assayed in functional assays with a potency <1 mM"

"Find me compounds that inhibit targets in NFkB pathway assayed in only functional assays with a potency <1 µM"

ChEMBL

DrugBank

Gene Ontology

Wikipathways

GeneGo

ChEBI

Uniprot

UMLS

GVKBio

ConceptWiki

ChemSpider

TrialTrove

TR Integrity

The Open PHACTS infrastructure can support many different domains & questions

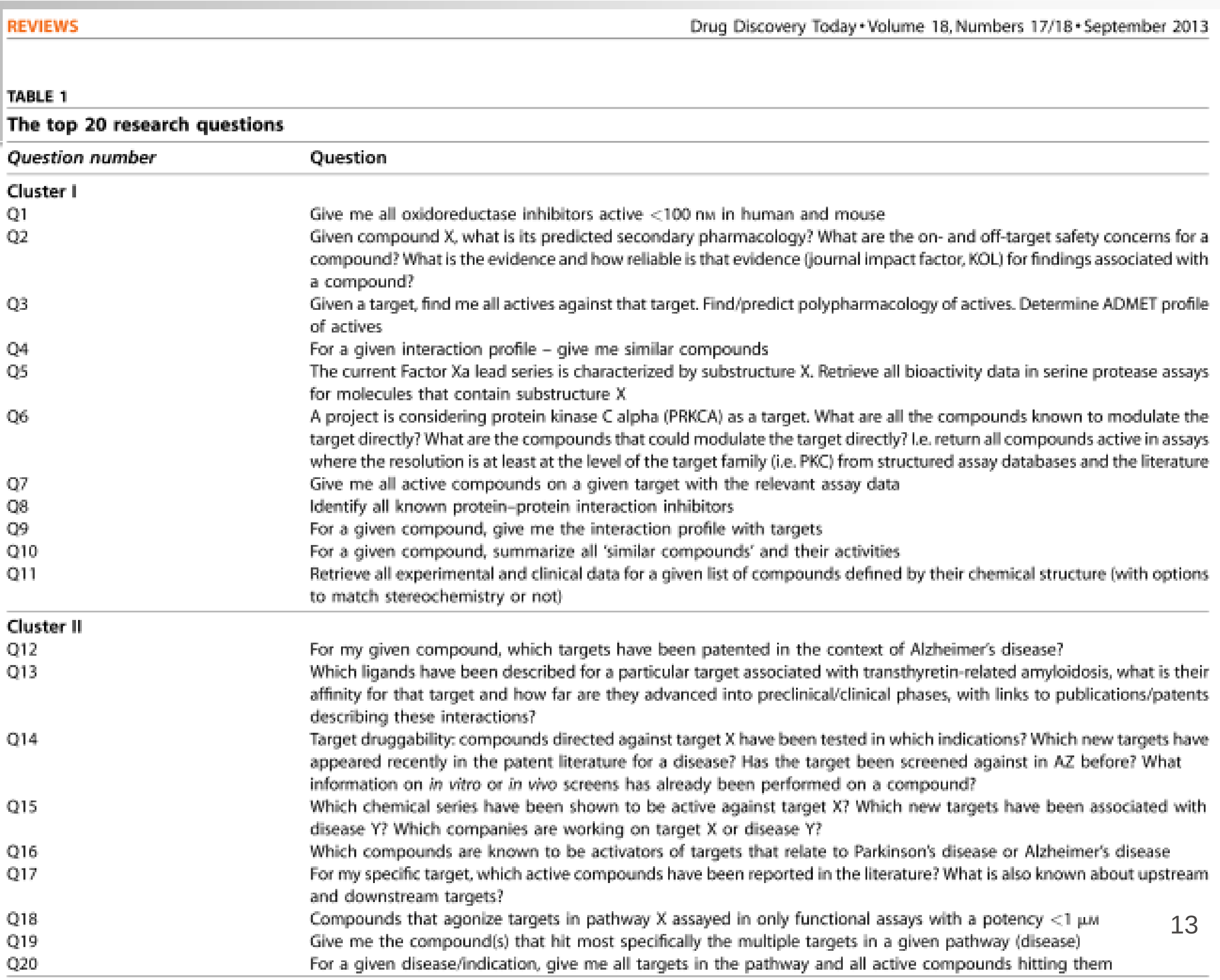


TABLE 1

The top 20 research questions

Question number	Question
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 information on *in vitro* or *in vivo* screens has already been performed on a compound?
Q15	Which chemical series have been shown to be active against target X? Which new targets have been associated with disease Y? Which companies are working on target X or disease Y?
Q16	Which compounds are known to be activators of targets that relate to Parkinson's disease or Alzheimer's disease
Q17	For my specific target, which active compounds have been reported in the literature? What is also known about upstream and downstream targets?
Q18	Compounds that agonize targets in pathway X assayed in only functional assays with a potency <1 μM
Q19	Give me the compound(s) that hit most specifically the multiple targets in a given pathway (disease)
Q20	For a given disease/indication, give me all targets in the pathway and all active compounds hitting them

Data Associations

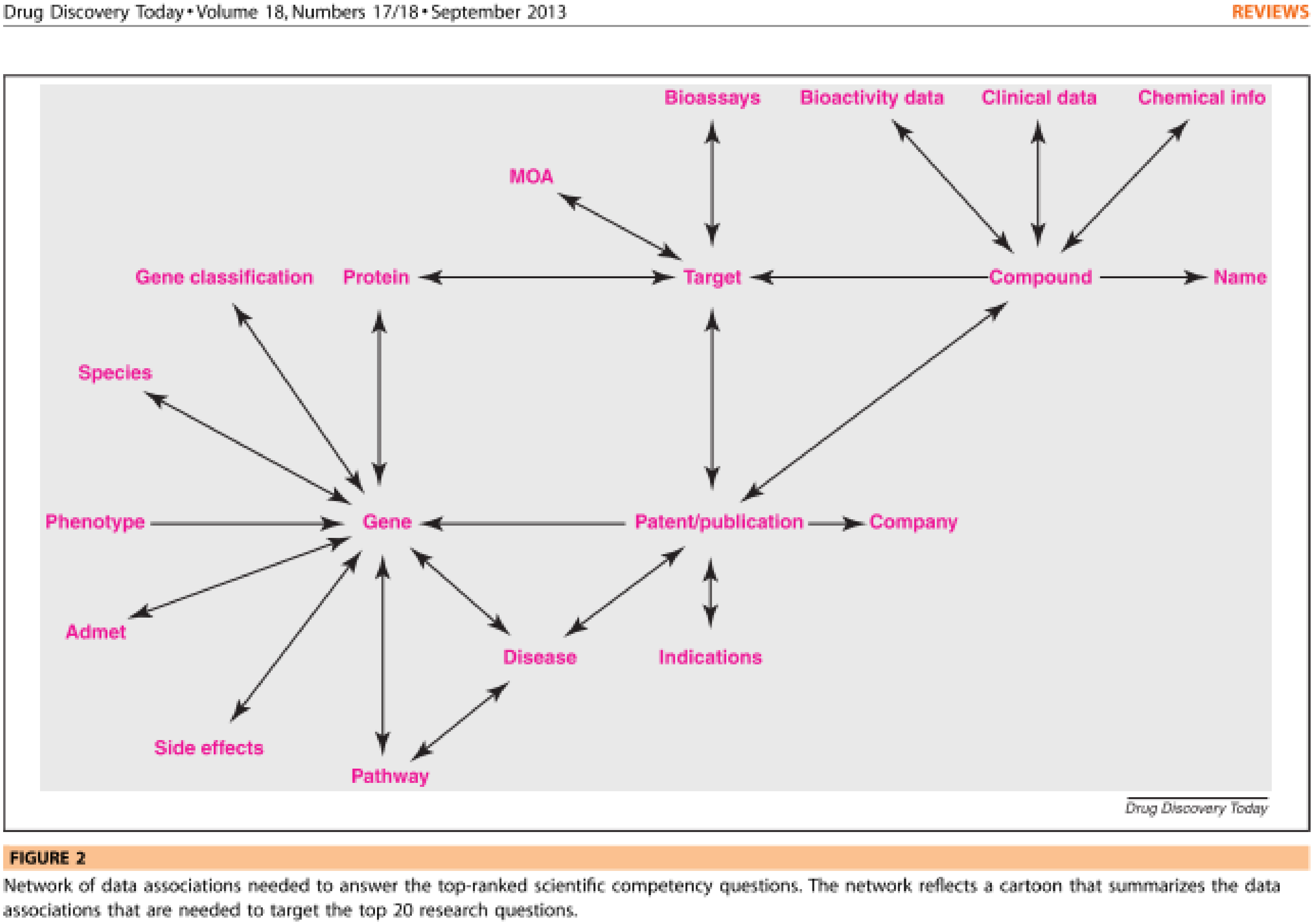


FIGURE 2

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.

Public Data Sources

ChEMBL

UniProt

IntAct

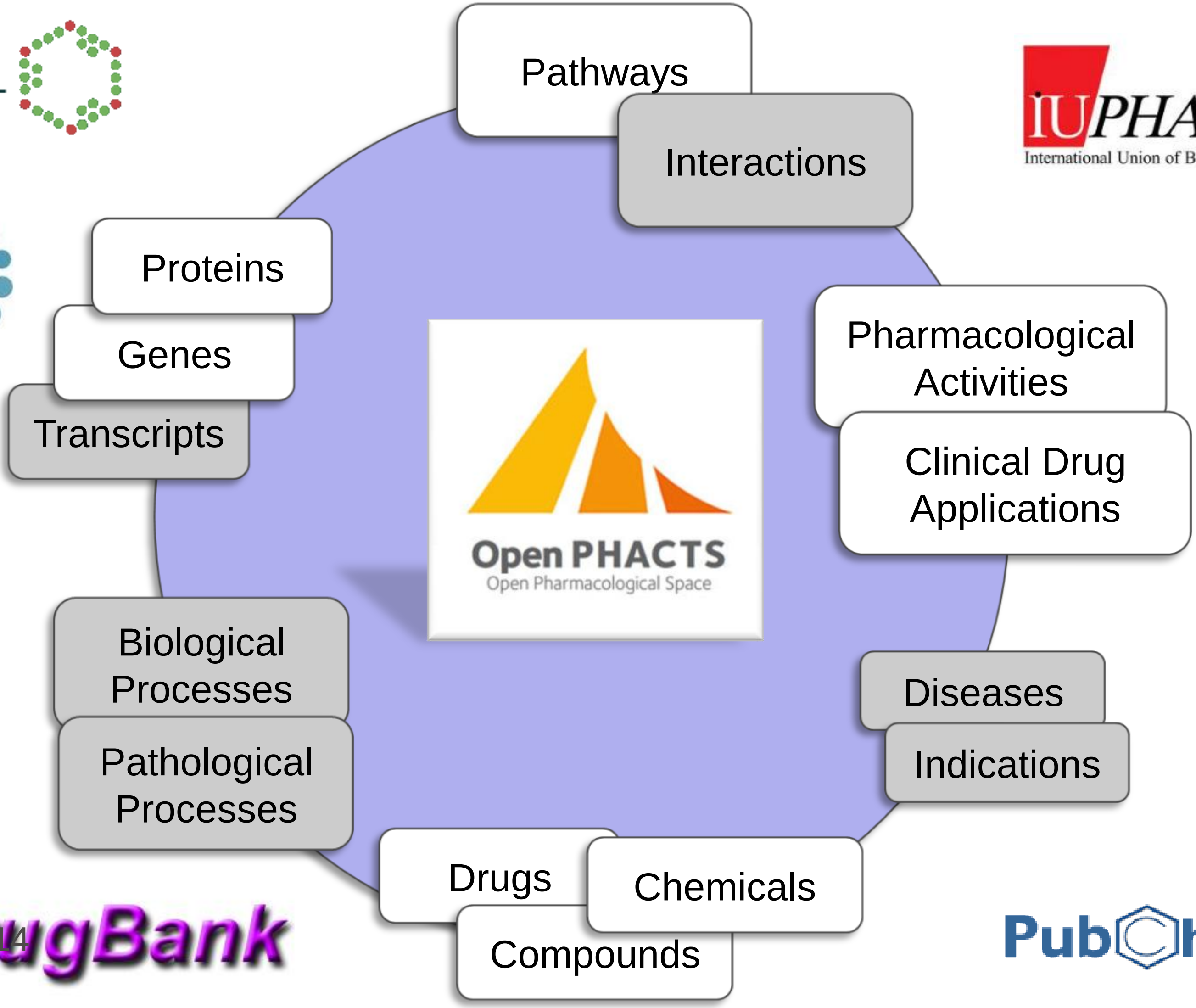
WikiPathways
Pathways for the People

DrugBank

IUPHAR DATABASE
International Union of Basic and Clinical Pharmacology

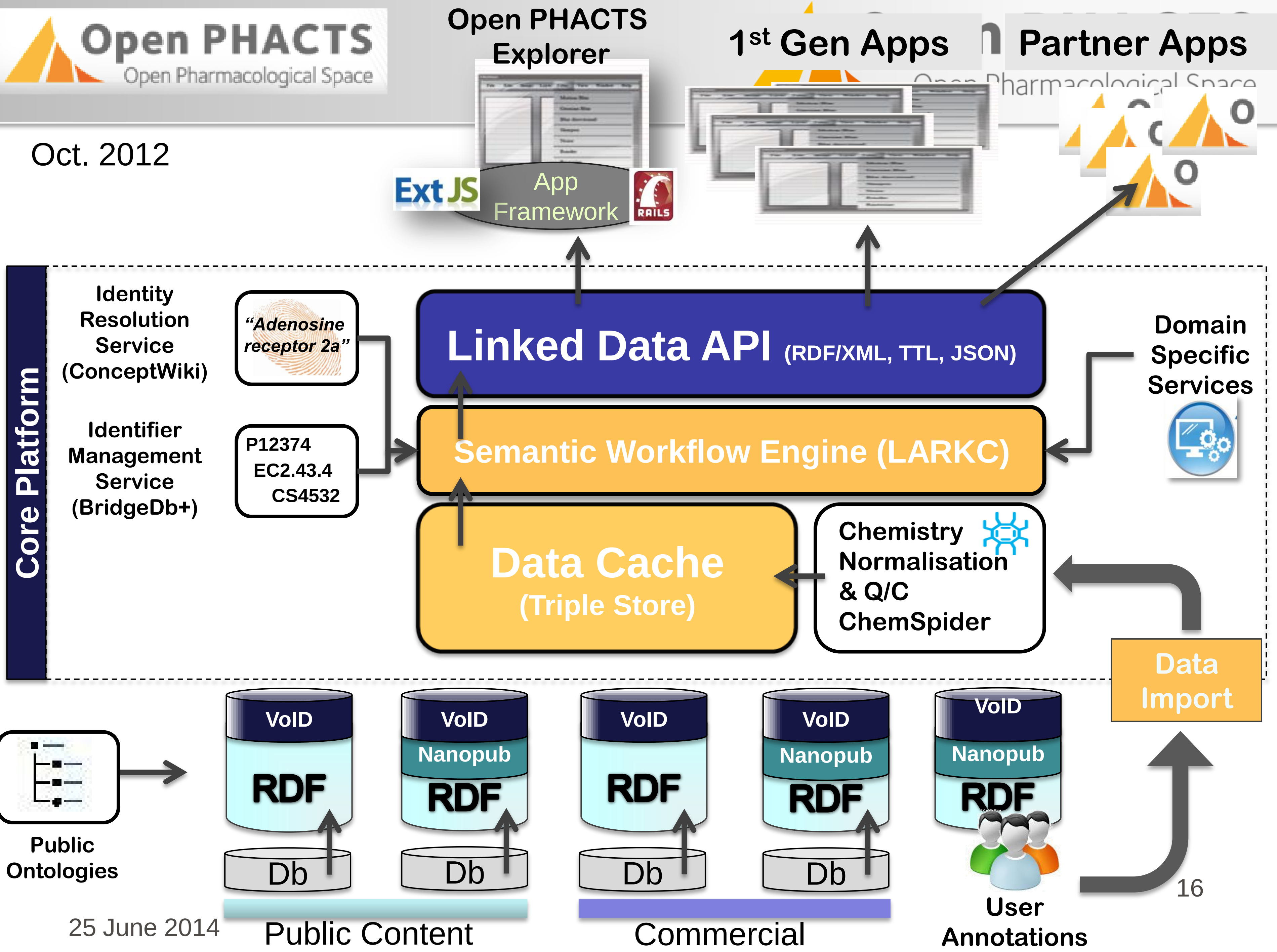
NCBI
National Center for Biotechnology Information

PubChem



Open PHACTS
Open Pharmacological Space

Oct. 2012



Open PHACTS Explorer

1st Gen Apps

Partner Apps

Ext JS

App Framework

RAILS

Core Platform

Identity Resolution Service (ConceptWiki)

“Adenosine receptor 2a”

Identifier Management Service (BridgeDb+)

P12374
EC2.43.4
CS4532

Linked Data API (RDF/XML, TTL, JSON)

Semantic Workflow Engine (LARKC)

Domain Specific Services

Data Cache
(Triple Store)

Chemistry Normalisation & Q/C ChemSpider

Data Import

Public Ontologies

VoID
RDF
Db

VoID
Nanopub
RDF
Db

VoID
RDF
Db

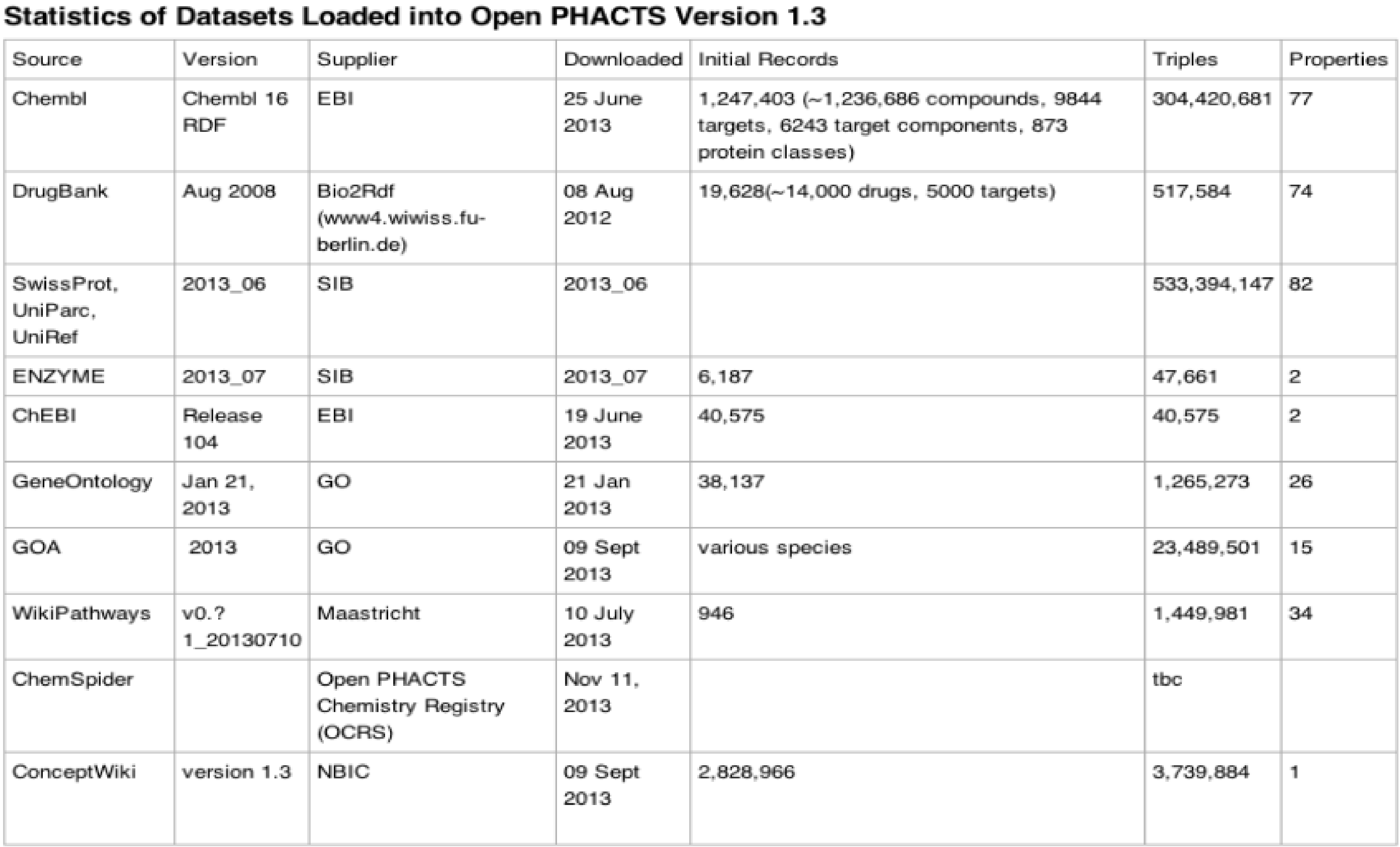
VoID
Nanopub
RDF
Db

VoID
Nanopub
RDF

Public Content

Commercial

User Annotations

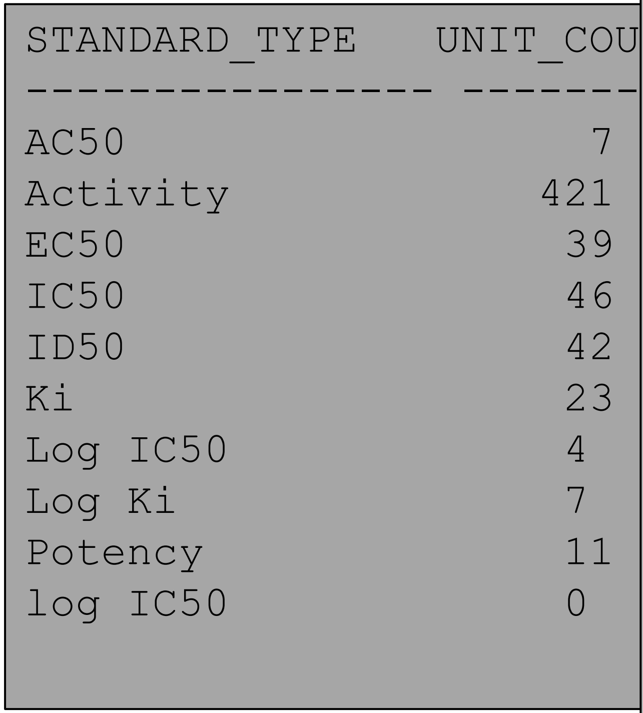


Statistics of Datasets Loaded into Open PHACTS Version 1.3

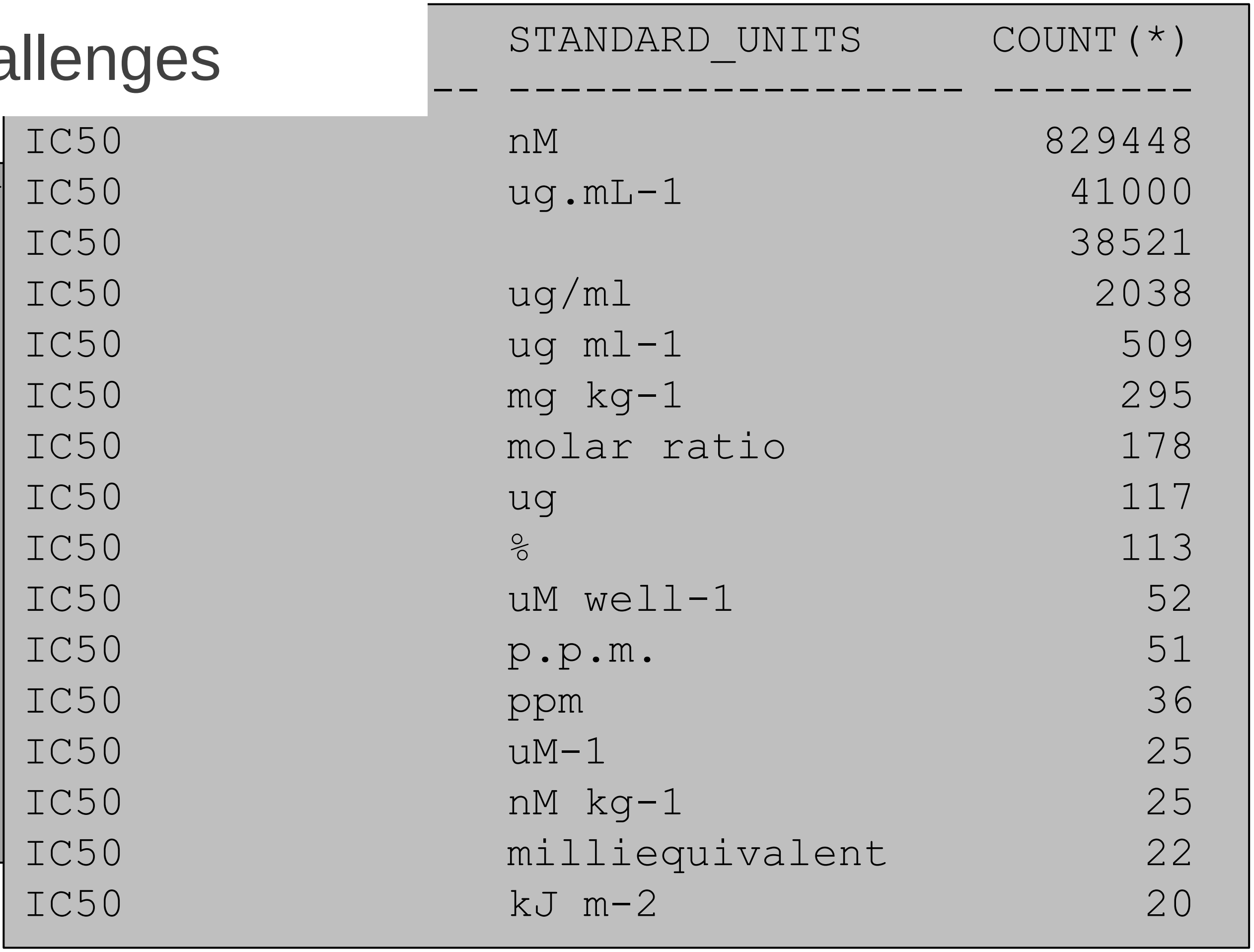
Source	Version	Supplier	Downloaded	Initial Records	Triples	Properties
Chembl	Chembl 16 RDF	EBI	25 June 2013	1,247,403 (~1,236,686 compounds, 9844 targets, 6243 target components, 873 protein classes)	304,420,681	77
DrugBank	Aug 2008	Bio2Rdf (www4.wiwiss.fu-berlin.de)	08 Aug 2012	19,628(~14,000 drugs, 5000 targets)	517,584	74
SwissProt, UniParc, UniRef	2013_06	SIB	2013_06		533,394,147	82
ENZYME	2013_07	SIB	2013_07	6,187	47,661	2
ChEBI	Release 104	EBI	19 June 2013	40,575	40,575	2
GeneOntology	Jan 21, 2013	GO	21 Jan 2013	38,137	1,265,273	26
GOA	2013	GO	09 Sept 2013	various species	23,489,501	15
WikiPathways	v0.? 1_20130710	Maastricht	10 July 2013	946	1,449,981	34
ChemSpider		Open PHACTS Chemistry Registry (OCRS)	Nov 11, 2013		tbc	
ConceptWiki	version 1.3	NBIC	09 Sept 2013	2,828,966	3,739,884	1

Example of vocabulary/ontology challenge

Quantitative Data Challenges



```
STANDARD_TYPE    UNIT_COU
---------------- --------
AC50                    7
Activity              421
EC50                   39
IC50                   46
ID50                   42
Ki                     23
Log IC50               4
Log Ki                 7
Potency                11
log IC50               0
```



```
                    STANDARD_UNITS      COUNT(*)
                 -- ------------------- --------
IC50                nM                    829448
IC50                ug.mL-1                41000
IC50                                       38521
IC50                ug/ml                   2038
IC50                ug ml-1                  509
IC50                mg kg-1                  295
IC50                molar ratio              178
IC50                ug                       117
IC50                %                        113
IC50                uM well-1                 52
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

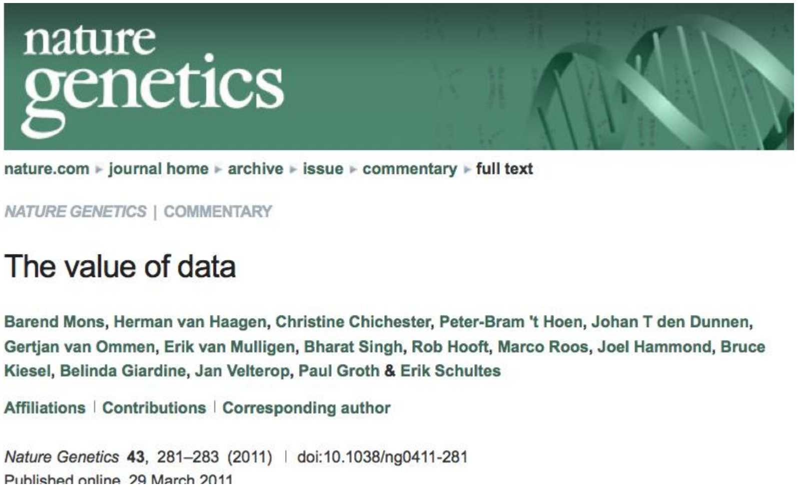
Implemented using the Quantities, Dimension, Units, Types Ontology (http://www.qudt.org/)

~ 100 units

Concept: nanopublications for provenance

Open PHACTS
Open Pharmacological Space

Nanopublications – Capturing scientific information in the Triple Store



nature genetics

nature.com ▸ journal home ▸ archive ▸ issue ▸ commentary ▸ full text

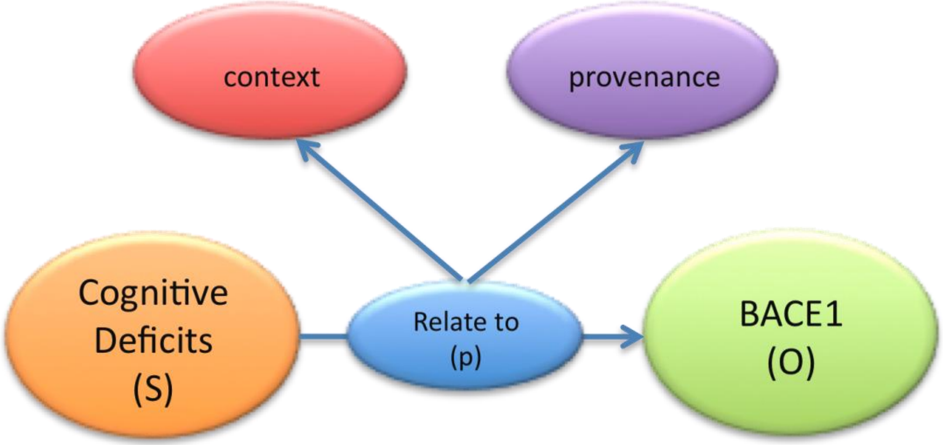
NATURE GENETICS | COMMENTARY

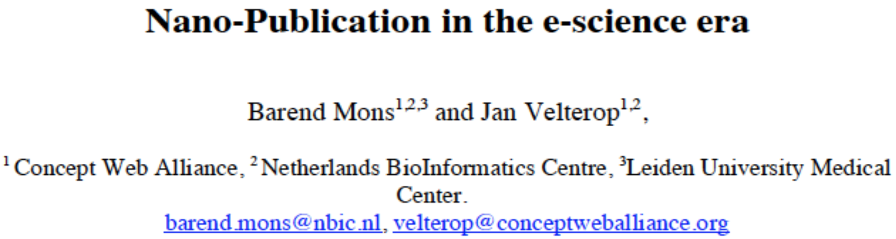
The value of data

Barend Mons, Herman van Haagen, Christine Chichester, Peter-Bram 't Hoen, Johan T den Dunnen, Gertjan van Ommen, Erik van Mulligen, Bharat Singh, Rob Hooft, Marco Roos, Joel Hammond, Bruce Kiesel, Belinda Giardine, Jan Velterop, Paul Groth & Erik Schultes

Affiliations | Contributions | Corresponding author

Nature Genetics **43**, 281–283 (2011) | doi:10.1038/ng0411-281
Published online 29 March 2011

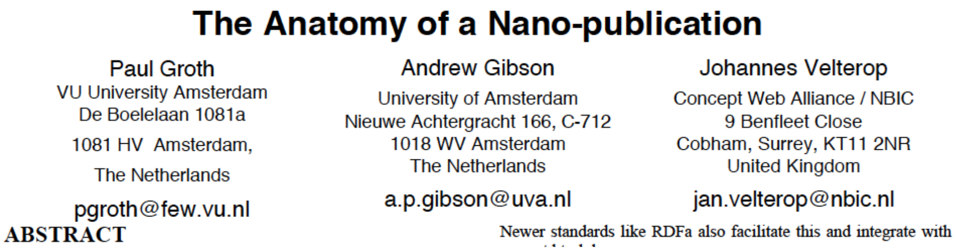




Nano-Publication in the e-science era

Barend Mons[1,2,3] and Jan Velterop[1,2],

[1] Concept Web Alliance, [2] Netherlands BioInformatics Centre, [3] Leiden University Medical Center.
barend.mons@nbic.nl, velterop@conceptweballiance.org



The Anatomy of a Nano-publication

Paul Groth
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam,
The Netherlands
pgroth@few.vu.nl

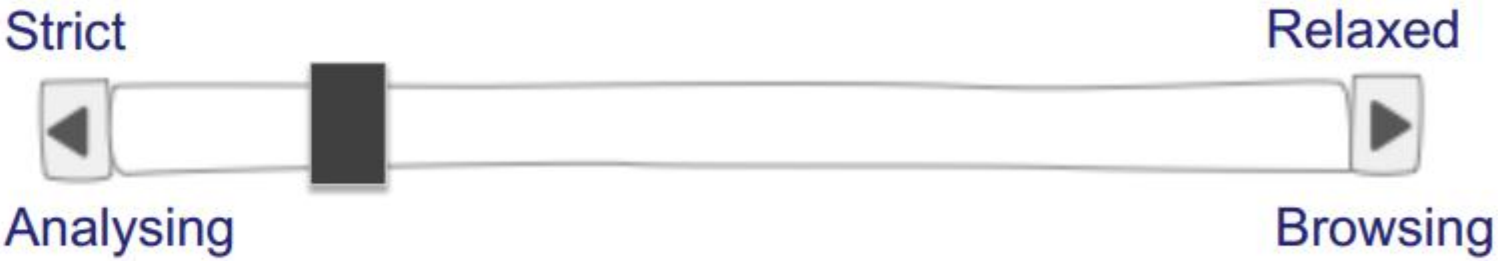
Andrew Gibson
University of Amsterdam
Nieuwe Achtergracht 166, C-712
1018 WV Amsterdam
The Netherlands
a.p.gibson@uva.nl

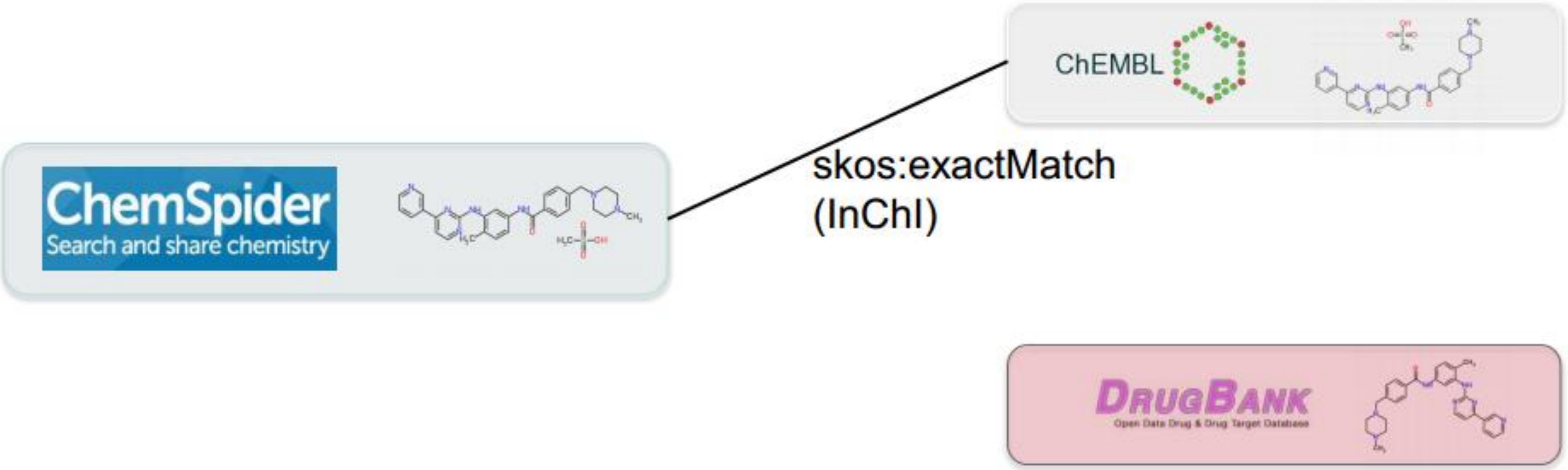
Johannes Velterop
Concept Web Alliance / NBIC
9 Benfleet Close
Cobham, Surrey, KT11 2NR
United Kingdom
jan.velterop@nbic.nl

ABSTRACT

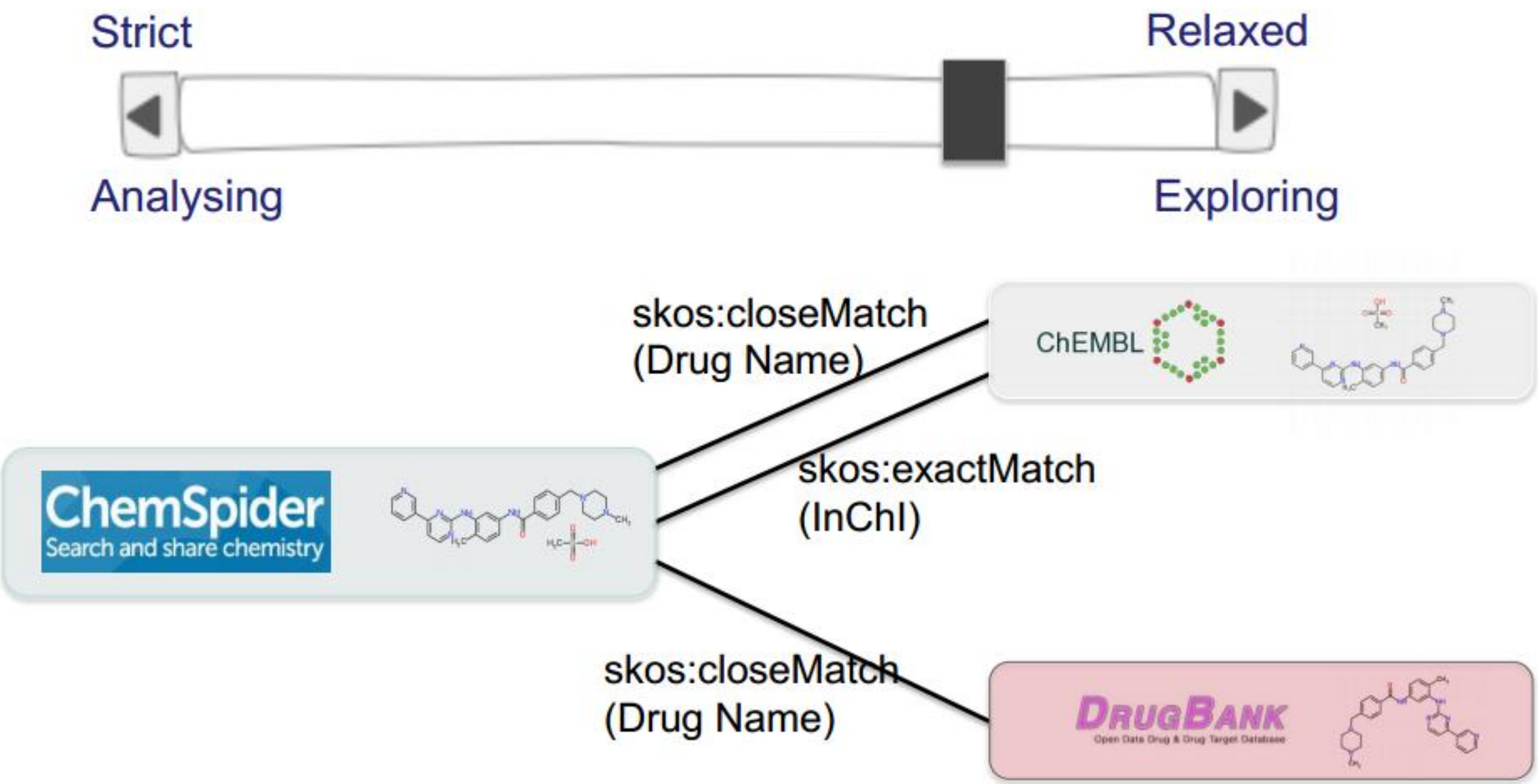
Newer standards like RDFa also facilitate this and integrate with

Concept: Scientific lenses




ChEMBL
ChemSpider
Search and share chemistry
skos:exactMatch
(InChI)
DrugBank
Open Data Drug & Drug Target Database

Concept: Scientific lenses



Strict

Relaxed

Analysing

Exploring

skos:closeMatch
(Drug Name)

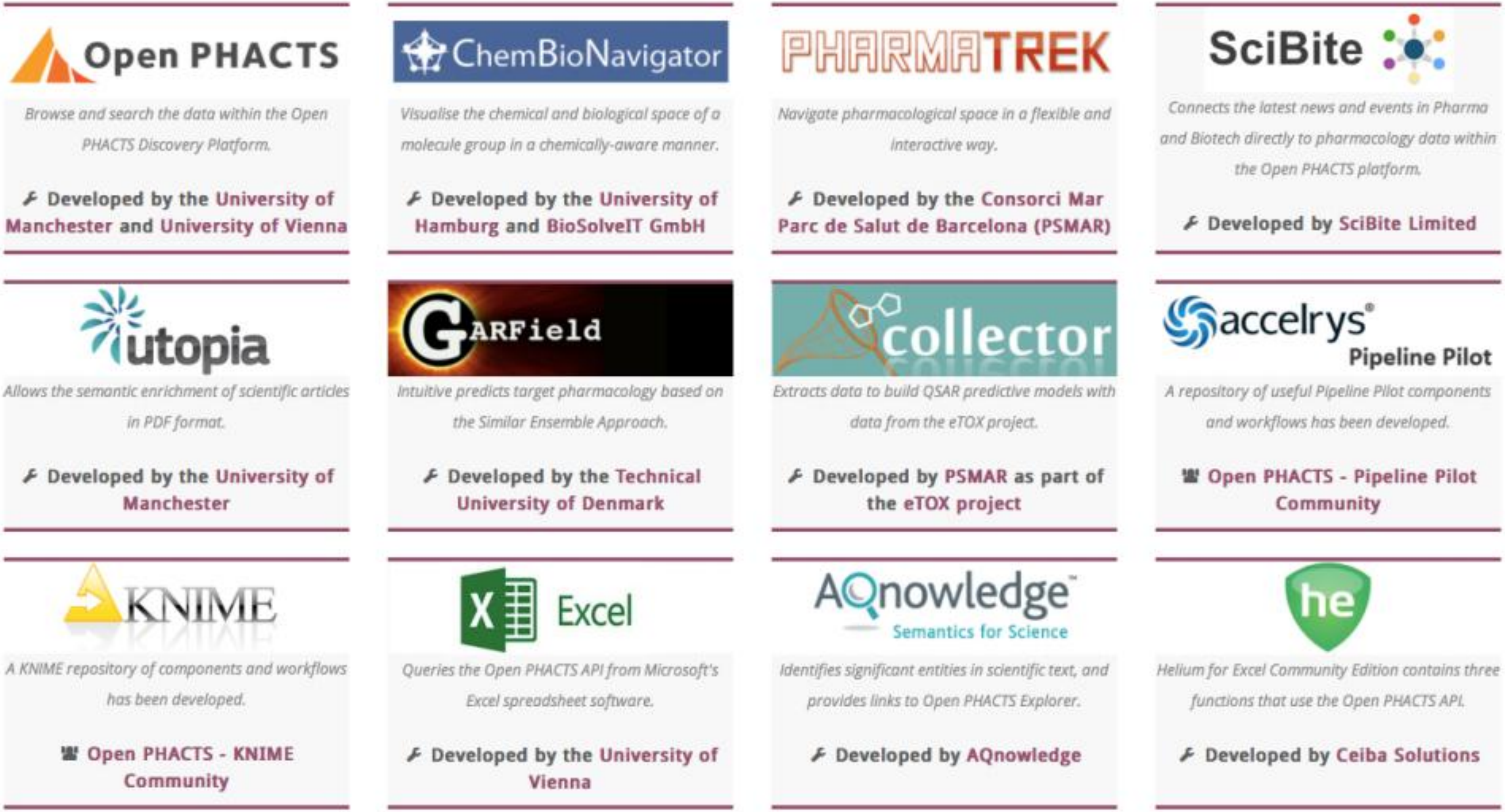
skos:exactMatch
(InChI)

skos:closeMatch
(Drug Name)

ChEMBL
ChemSpider
Search and share chemistry
DrugBank
Open Data Drug & Drug Target Database

Example applications

Open PHACTS
Open Pharmacological Space



Open PHACTS

Browse and search the data within the Open PHACTS Discovery Platform.

Developed by the University of Manchester and University of Vienna

ChemBioNavigator

Visualise the chemical and biological space of a molecule group in a chemically-aware manner.

Developed by the University of Hamburg and BioSolveIT GmbH

PHARMATREK

Navigate pharmacological space in a flexible and interactive way.

Developed by the Consorci Mar Parc de Salut de Barcelona (PSMAR)

SciBite

Connects the latest news and events in Pharma and Biotech directly to pharmacology data within the Open PHACTS platform.

Developed by SciBite Limited

utopia

Allows the semantic enrichment of scientific articles in PDF format.

Developed by the University of Manchester

GARField

Intuitive predicts target pharmacology based on the Similar Ensemble Approach.

Developed by the Technical University of Denmark

collector

Extracts data to build QSAR predictive models with data from the eTOX project.

Developed by PSMAR as part of the eTOX project

accelrys® Pipeline Pilot

A repository of useful Pipeline Pilot components and workflows has been developed.

Open PHACTS - Pipeline Pilot Community

KNIME

A KNIME repository of components and workflows has been developed.

Open PHACTS - KNIME Community

Excel

Queries the Open PHACTS API from Microsoft's Excel spreadsheet software.

Developed by the University of Vienna

AQnowledge™ Semantics for Science

Identifies significant entities in scientific text, and provides links to Open PHACTS Explorer.

Developed by AQnowledge

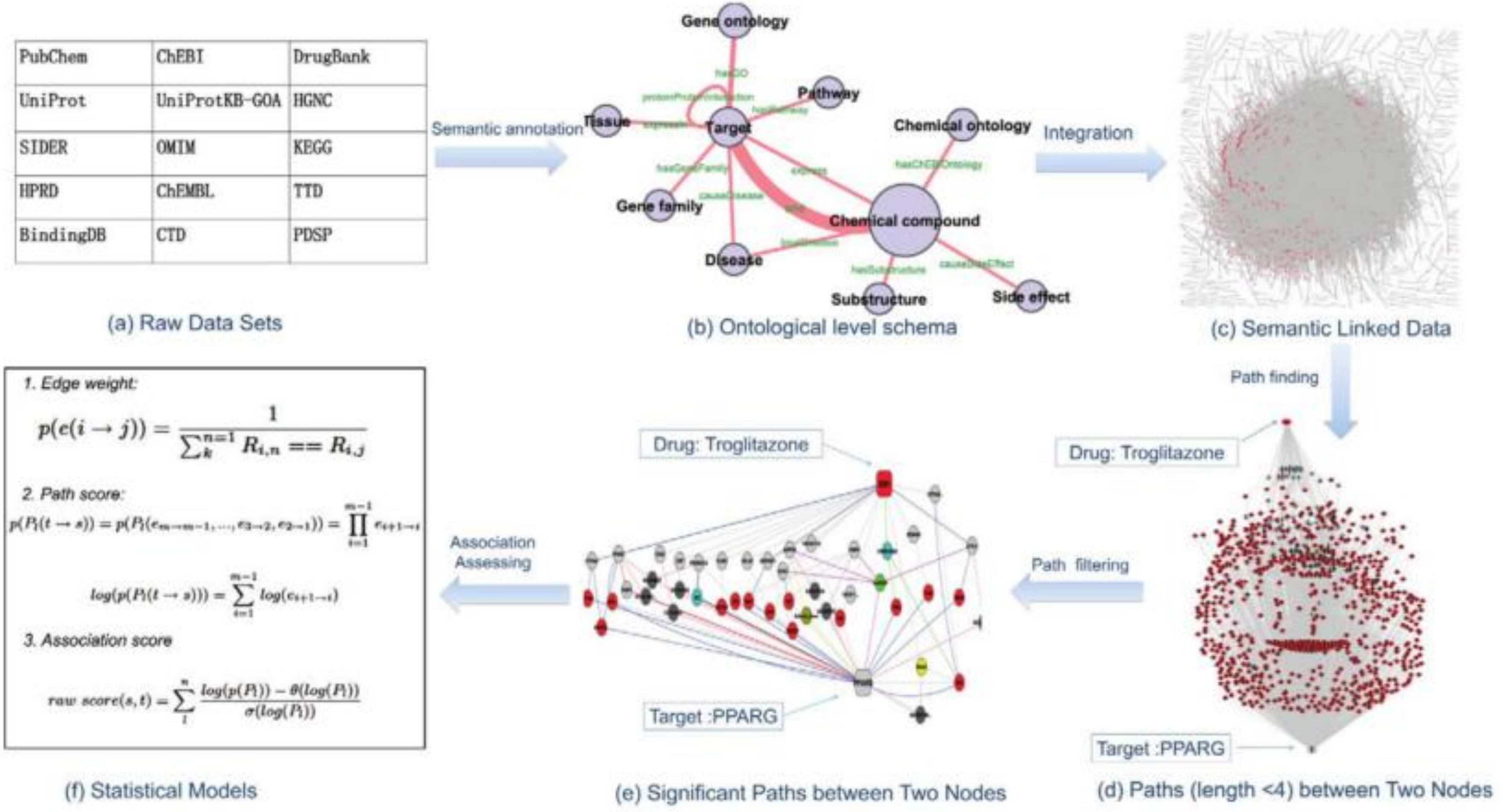
he

Helium for Excel Community Edition contains three functions that use the Open PHACTS API.

Developed by Ceiba Solutions

More complex use cases: Relating two distant concepts

Open PHACTS
Open Pharmacological Space



PubChem	ChEBI	DrugBank
UniProt	UniProtKB-GOA	HGNC
SIDER	OMIM	KEGG
HPRD	ChEMBL	TTD
BindingDB	CTD	PDSP

1. Edge weight:

$$p(e(i \rightarrow j)) = \frac{1}{\sum_{k}^{n=1} R_{i,n} == R_{i,j}}$$

2. Path score:

$$p(P_i(t \rightarrow s)) = p(P_i(e_{m \rightarrow m-1}, \ldots, e_{3 \rightarrow 2}, e_{2 \rightarrow 1})) = \prod_{i=1}^{m-1} e_{i+1 \rightarrow i}$$

$$log(p(P_i(t \rightarrow s))) = \sum_{i=1}^{m-1} log(e_{i+1 \rightarrow i})$$

3. Association score

$$raw\ score(s,t) = \sum_{i}^{n} \frac{log(p(P_i)) - \theta(log(P_i))}{\sigma(log(P_i))}$$

Figure 1. SLAP pipeline. An ontology is used to annotate public data sets and integrate them into a semantic linked network. Two nodes are linked by one or more number of paths, but only a small number of significant paths are kept for association estimation. The path significance and drug target associations are assessed by statistical models derived from random samples.

doi:10.1371/journal.pcbi.1002574.g001

Chen et at. PLoS Comp Biol 2012

Developments

- Continue improving the system: features, performance, API calls, etc
- Expand implementation of data sources based on new set of scientific use cases – Project received 2 years additional funding
- Development and improvement of new and existing applications that use the Open PHACTS API
- Set up organizational model to continue maintenance and development after IMI funding

Acknowledgments

The Open PHACTS consortium

Play!
API: https://dev.openphacts.org/