Open PHACTS

Open Pharmacological Space

Incorporating Commercial and Private Data into an Open Linked Data Platform for Drug Discovery

Carole Goble, Alasdair J G Gray, Lee Harland,
Karen Karapetyan, Antonis Loizou, Ivan Mikhailov,
Yrjänä Rankka, Stefan Senger, Valery Tkachenko,
Antony J Williams, and Egon L Willighagen

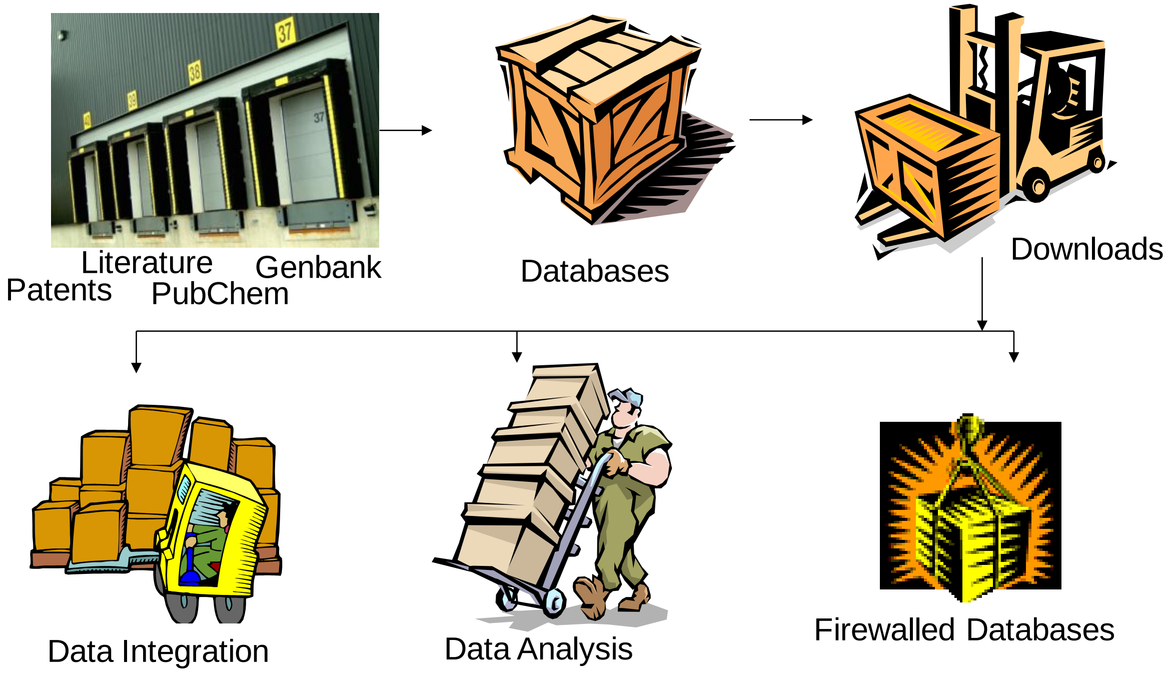
www.openphacts.org A.J.G.Gray@hw.ac.uk

@open_phacts @gray_alasdair

Pre-competitive Informatics

Open PHACTS
Open Pharmacological Space

Pharmaceutical companies are all accessing, processing, storing & re-processing external research data



Literature
Genbank
Patents
PubChem
Databases
Downloads
Data Integration
Data Analysis
Firewalled Databases

x Repeat @ each company

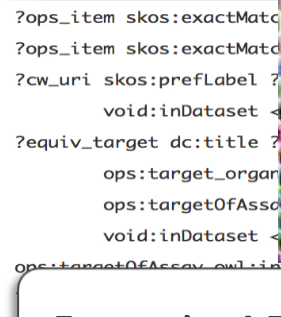
Lowering industry firewalls: pre-competitive informatics in drug discovery
Nature Reviews Drug Discovery (2009) 8, 701-708 doi:10.1038/nrd2944

Open PHACTS objective

Open Standards

Apps

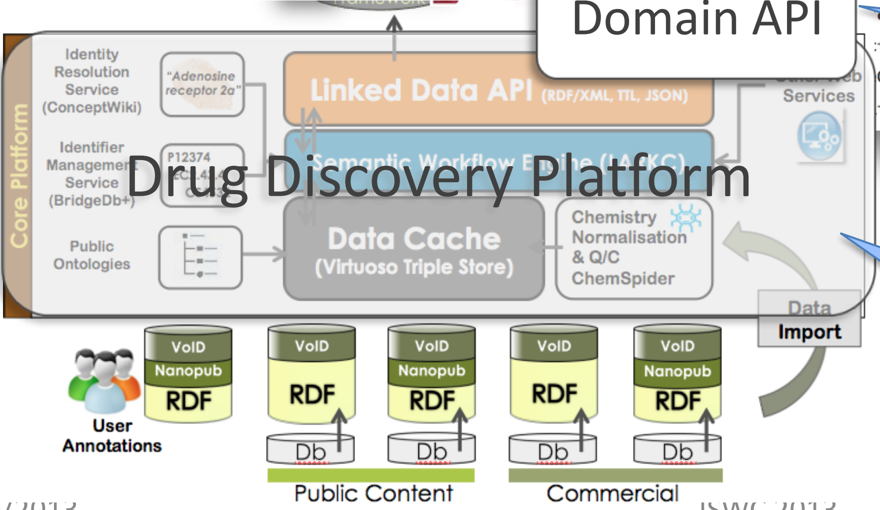
Interactive responses



```
?ops_item skos:exactMatc
?ops_item skos:exactMatc
?cw_uri  skos:prefLabel ?
         void:inDataset <
?equiv_target dc:title ?
         ops:target_organ
         ops:targetOfAsso
         void:inDataset <
ops:targetOfAssay owl:inverseOf chembl:hasTarge
                               ?assay_organism ;
                               as    descriptio
                               :tivity_uri
                               Of chembl:onAssa
                               type ;
```

Domain API

Provenance of data

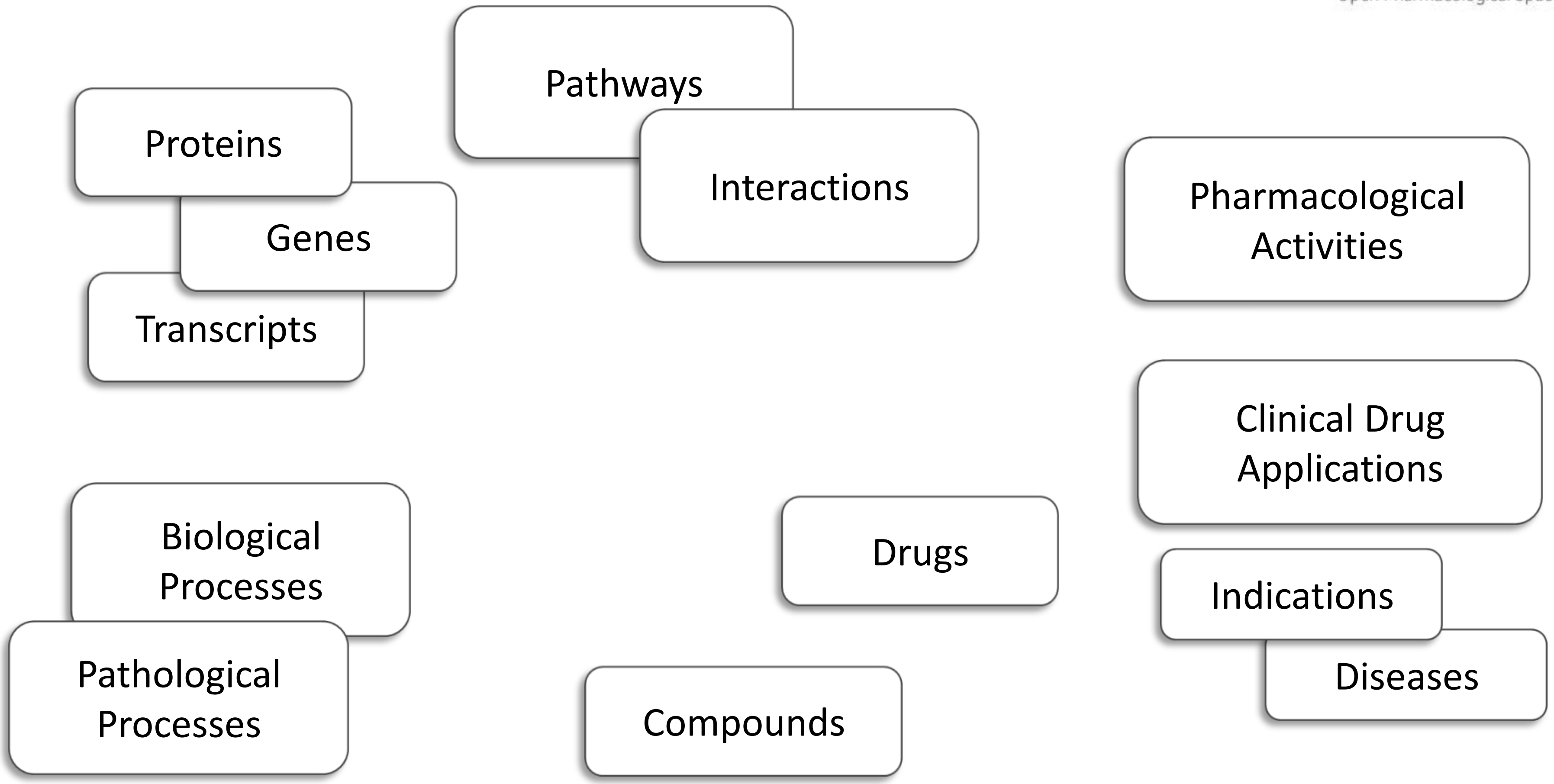


Core Platform
Identity Resolution Service (ConceptWiki)
"Adenosine receptor 2a"
Linked Data API (RDF/XML, TTL, JSON)
Semantic Workflow Engine (LArKC)
Identifier Management Service (BridgeDb+)
P12374
Public Ontologies
Data Cache (Virtuoso Triple Store)
Chemistry Normalisation & Q/C ChemSpider
Other Web Services
Data Import
VoID
Nanopub
RDF
Db
User Annotations
Public Content
Commercial

Drug Discovery Platform

Production quality

Drug Discovery Data



Pathways

Proteins

Interactions

Genes

Pharmacological Activities

Transcripts

Clinical Drug Applications

Biological Processes

Drugs

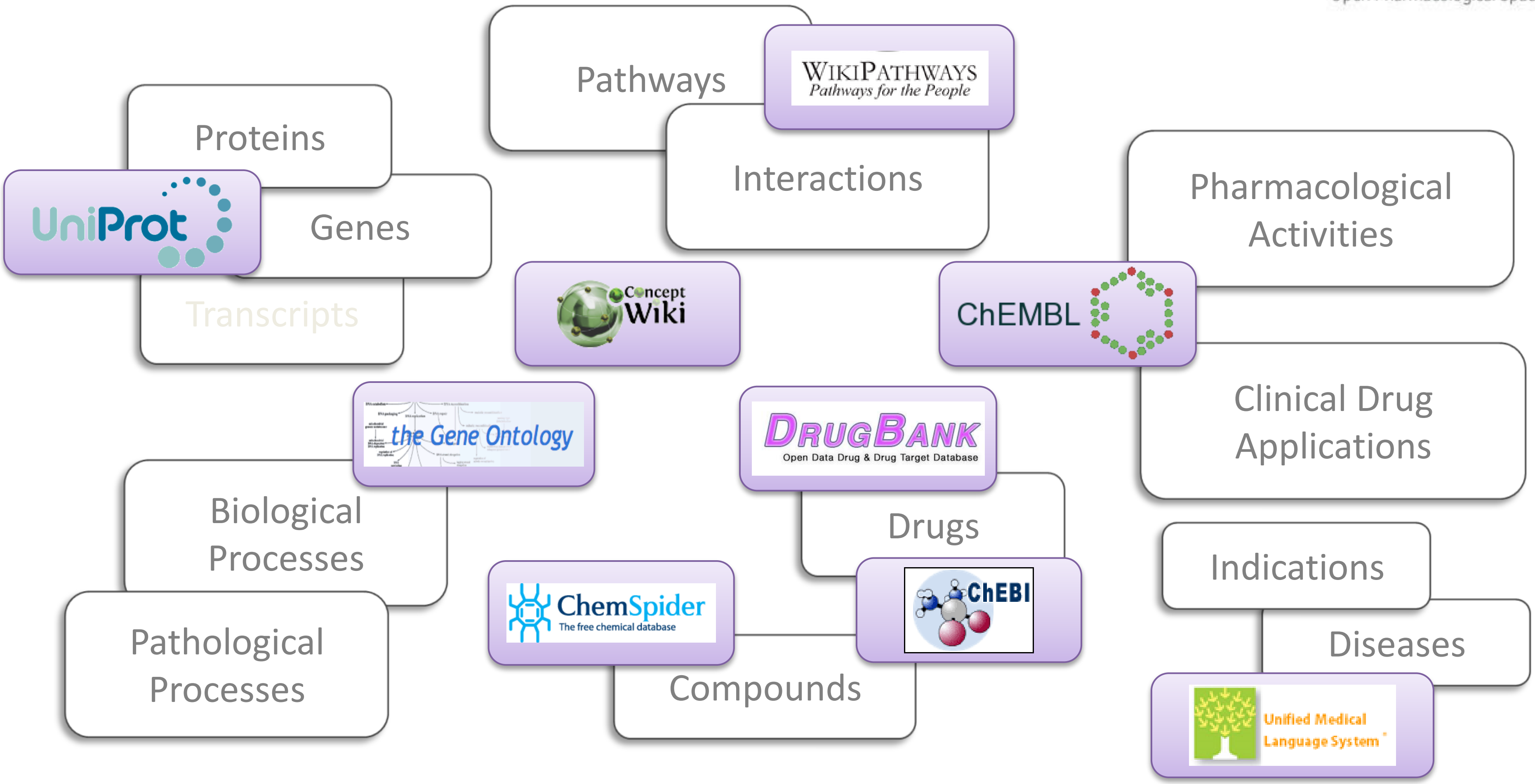
Indications

Pathological Processes

Diseases

Compounds

Public Data



Pathways

Interactions

Proteins

Genes

Transcripts

Pharmacological Activities

Clinical Drug Applications

Biological Processes

Pathological Processes

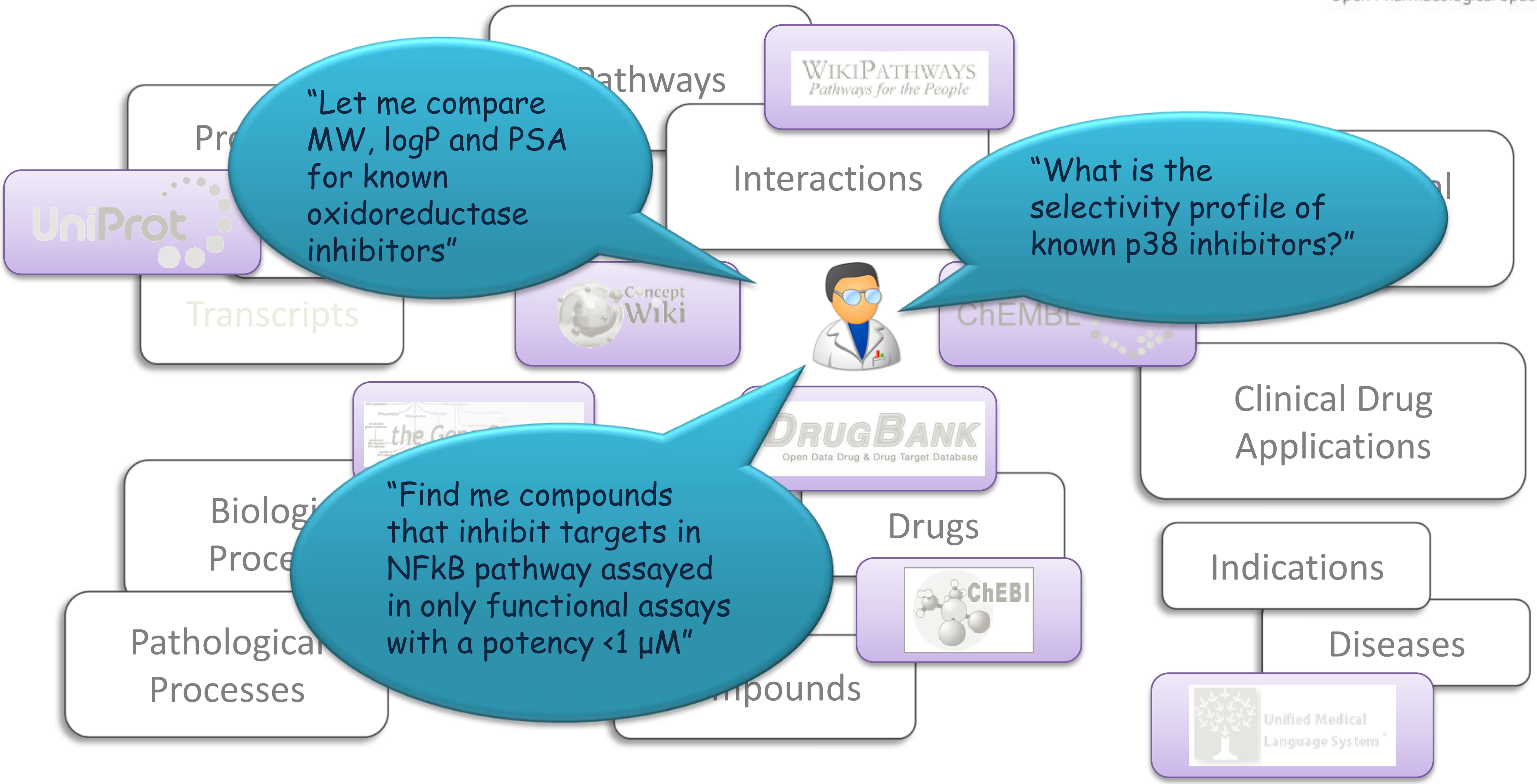
Drugs

Compounds

Indications

Diseases

Real Business Questions

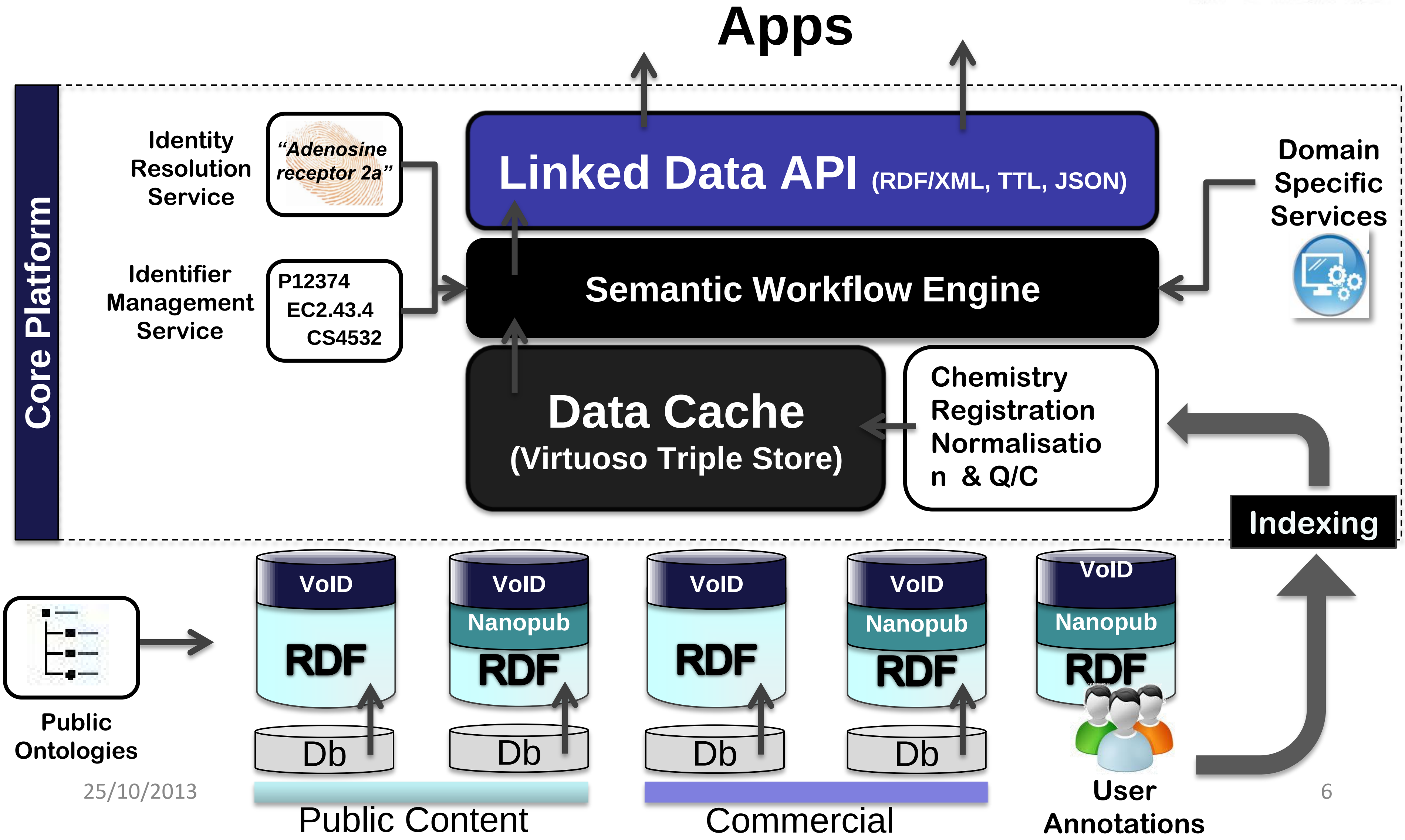


"Let me compare MW, logP and PSA for known oxidoreductase inhibitors"

"What is the selectivity profile of known p38 inhibitors?"

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

OPS Discovery Platform



Present Content: Public Data

Open PHACTS
Open Pharmacological Space

Source	Initial Records	Triples	Properties
ChEMBL	1,247,403	305,419,649	77
DrugBank	19,628	517,584	74
UniProt	?	533,394,147	82
ENZYME	[illegible]	73,838	2
ChEBI	[illegible]	40,575	2
GeneOntology	38,137	[illegible]	26
GOA	?	[illegible]	15
ChemSpider	1,194,437	161,336,857	26
ConceptWiki	2,828,966	3,739,884	1
WikiPathways	946	1,449,981	34

Over a billion triples

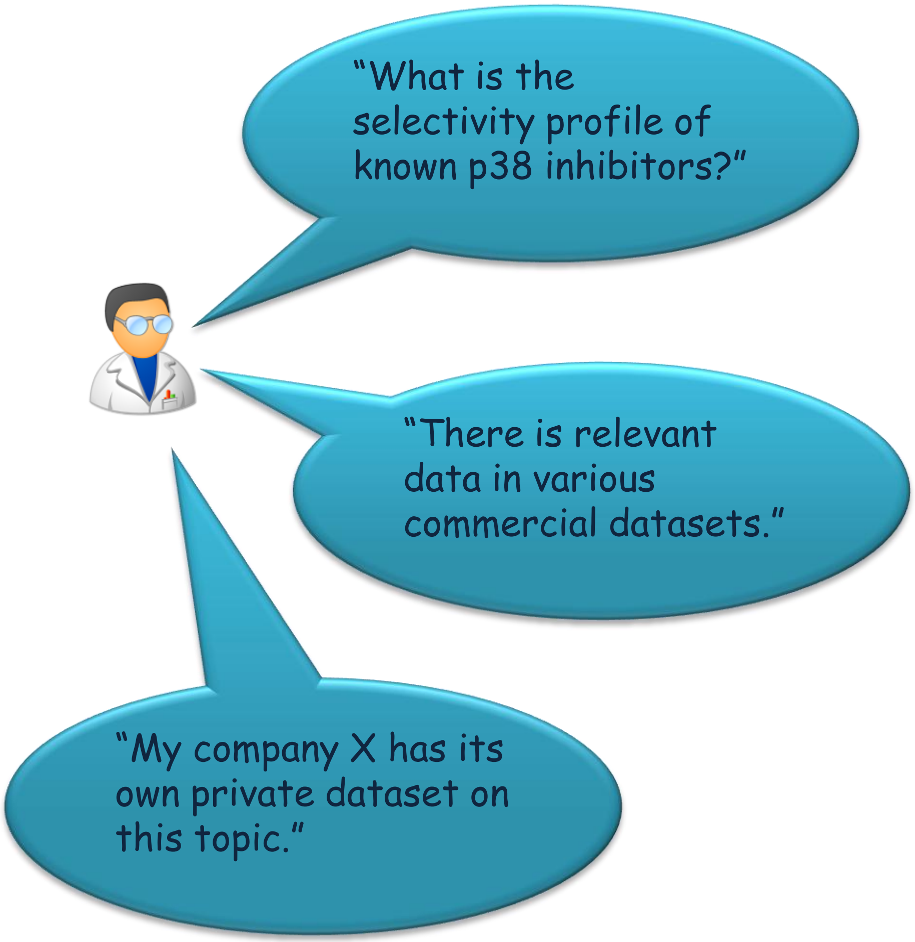
Semantic Integration Methodology

1. Define use cases
2. Identify Data
   - Create RDF
   - VoID dataset descriptions
3. Create mappings
   - between data set and known data sets (instance level)
   - index for text to URL conversion

Semantic Integration Methodology

4. Ingest RDF into data cache
   (i.e. triple store)
5. Define access paths to core concepts in data
6. Extend or create SPARQL queries for API calls
7. Publish API calls

Commercial Data Use Case



"What is the selectivity profile of known p38 inhibitors?"

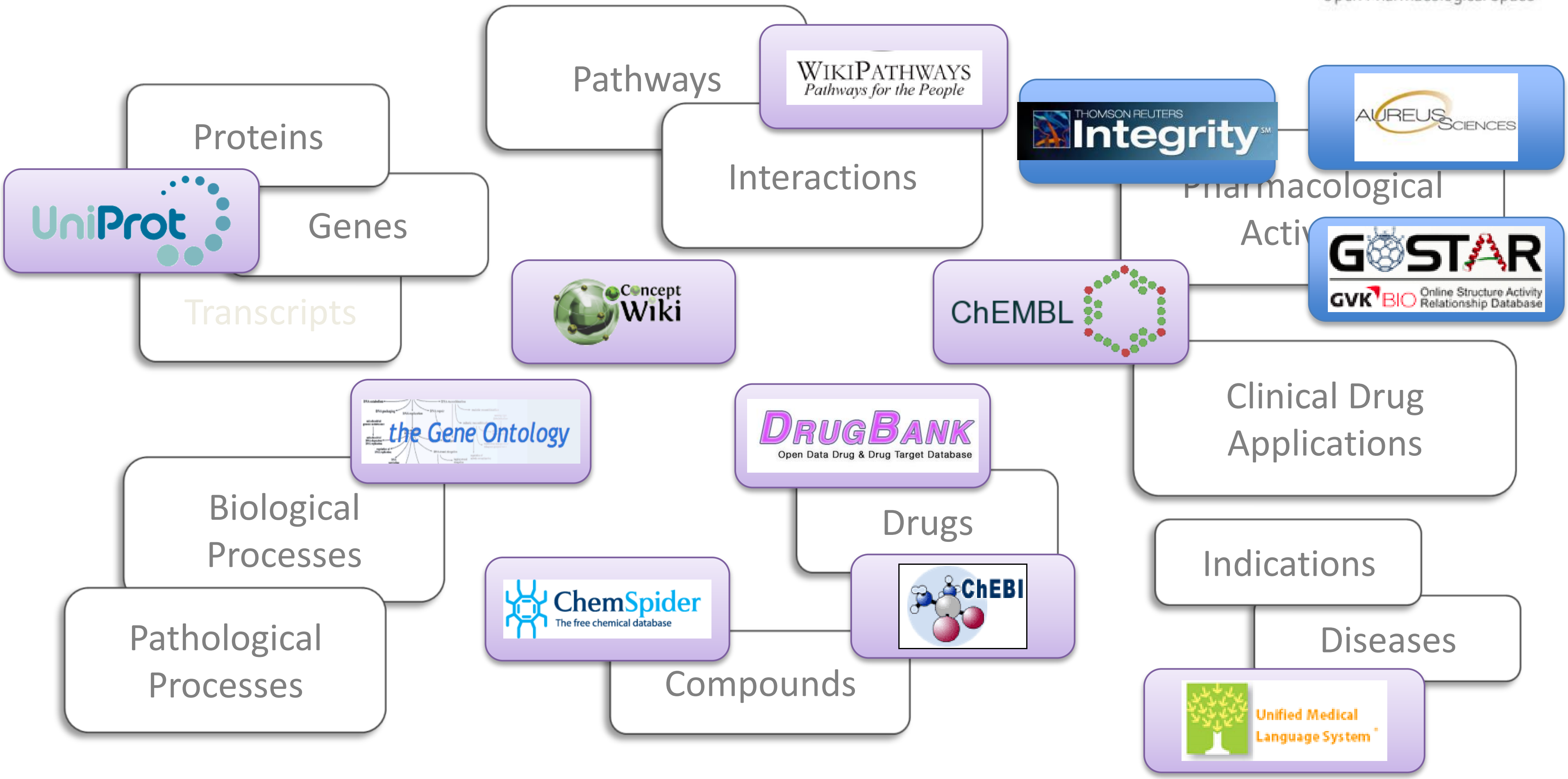
"There is relevant data in various commercial datasets."

"My company X has its own private dataset on this topic."

- Comprehensive data coverage
  - Commercial data collections
  - Extensive private collections
- Control data responses
  - Only authorised data

Commercial Data Sets Pilot

Open PHACTS
Open Pharmacological Space



Pathways

Proteins

Genes

Transcripts

Interactions

Pharmacological

Activ

Clinical Drug Applications

Biological Processes

Pathological Processes

Drugs

Compounds

Indications

Diseases

Linked Open Data

Open PHACTS
Open Pharmacological Space

★	make your stuff available on the Web (whatever format) under an open license
★★	make it available as structured data (e.g. Excel instead of image scan of a table)
★★★	use non-proprietary formats (e.g. CSV instead of Excel)
★★★★	use URIs to denote things, so that people can point at your stuff
★★★★★	link your data to other data to provide context

http://5stardata.info/

Commercial Linked Data

Open PHACTS
Open Pharmacological Space

- Same conversion challenges as Open Data!
  - Goal to have 5★ linked data
  - www.openphacts.org/specs/rdfguide/
- Pilot (sample) data provided as data dumps
  - XML
  - CSV
  - RDF
- Structurally similar to ChEMBL
- Converted to interoperable RDF

Data Modelling Challenges

- Contain private terminologies
  - – Mapped to public equivalents
  - – On going work
- Units represented as strings
  - – Not always consistent, e.g. IC50, IC_50, IC-50
  - – QUDT extended, e.g. IC_{50}
  - – www.openphacts.org/specs/units/

Dataset Descriptions

www.openphacts.org/specs/datadesc/

Open PHACTS
Open Pharmacological Space

Enable

- Discovery
  - Name
  - Description
  - Coverage
- Access control
  - License
  - File locations
- Answer Provenance
  - Returned data links to description

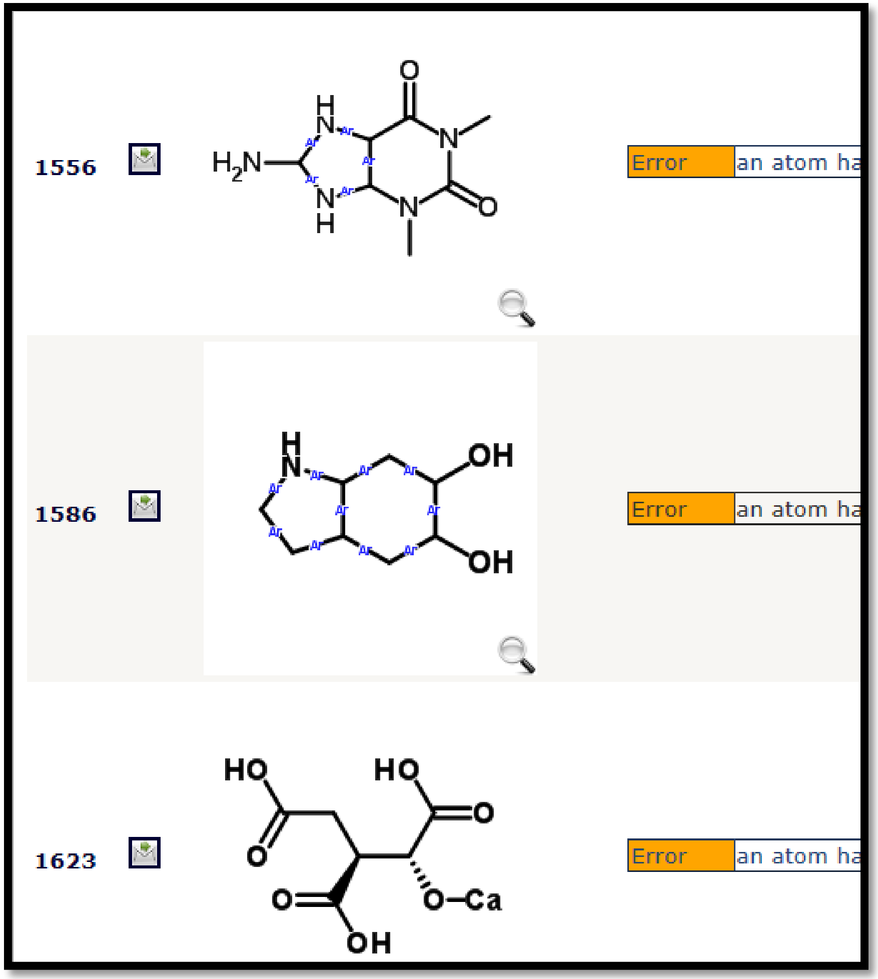
Commercial Data Description

- Publicly discoverable
  - Advertisement for data
  - Bring in more customers
- Restricted access by license

Private Data Description

- Hidden to all but authorised
- Restricted access

Chemical mappings



1556	Error	an atom ha
1586	Error	an atom ha
1623	Error	an atom ha

- Data is messy!
- Identify common problems:
  - Charge imbalance
  - Stereochemistry
- Link based on structure

Chemistry Registration

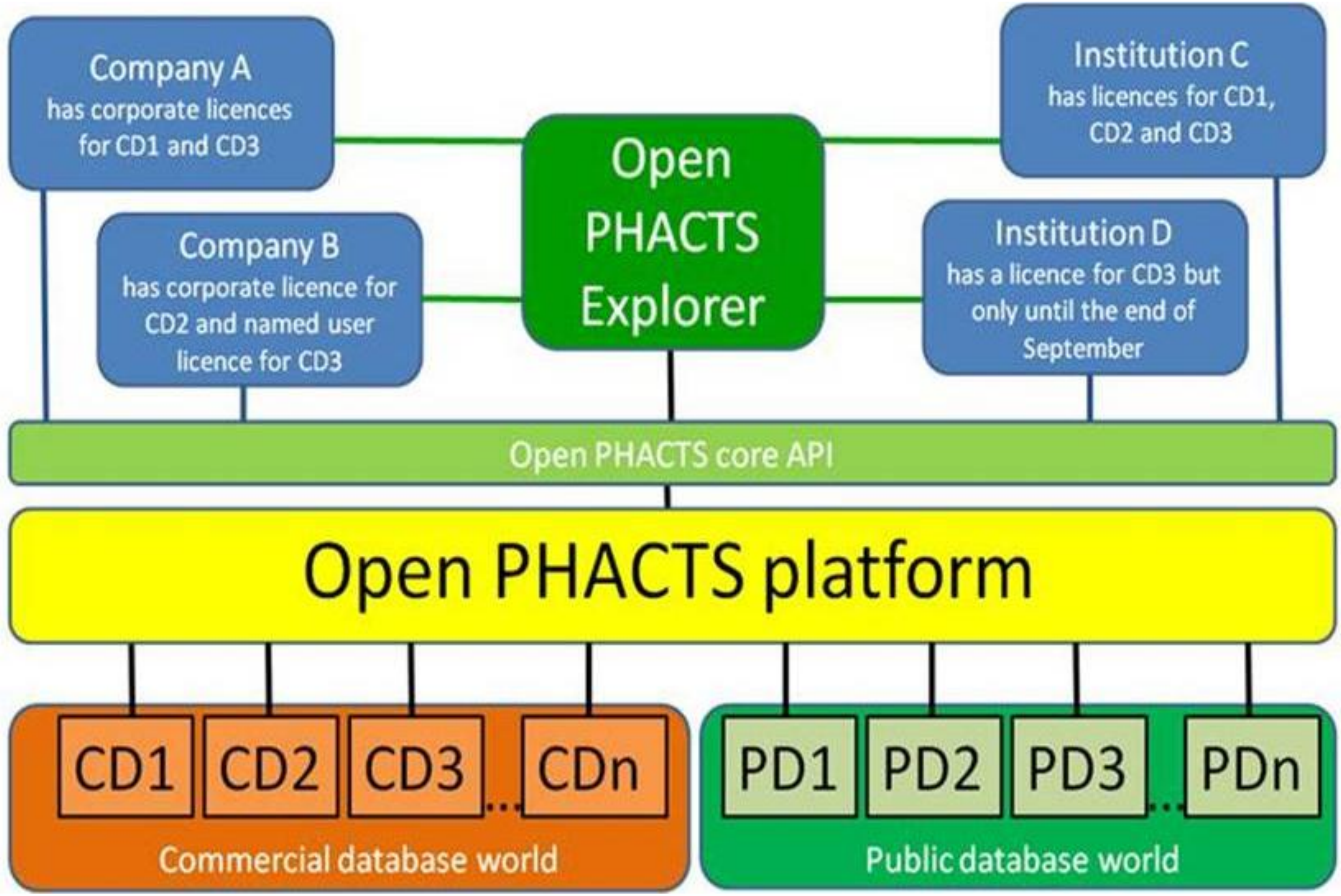
ChemSpider Service

- Validates and standardizes chemical representations
- Manual curation by RSC staff
- Data loaded in ChemSpider
- Open data: unsuitable for
  - Commercial data
  - Private data

Chemical Registration Service

- Utilizes ChemSpider Validation and Standardization platform
- Utilizes FDA rule set as basis for standardization
- Generates OPSID for chemicals
- Computes properties

Access Requirements



Company A
has corporate licences for CD1 and CD3

Company B
has corporate licence for CD2 and named user licence for CD3

Open PHACTS Explorer

Institution C
has licences for CD1, CD2 and CD3

Institution D
has a licence for CD3 but only until the end of September

Open PHACTS core API

Open PHACTS platform

Data Access

- Each data set loaded into separate graph in cache
- Pilot data same form as open ChEMBL data
  - Extend queries with sub-queries for each set
- Restricted access
  - Virtuoso offers graph-based access restriction
  - Commercial data sets turned on/off

Conclusions

- Drug discovery requires full data coverage
  - Public/open data
    - Open description
    - Open data
  - Commercial data
    - Open description
    - Restricted data
  - Private data
    - Restricted description
    - Restricted data
- Pilot study with three commercial datasets

Conclusions

Open PHACTS
Open Pharmacological Space

- Data Modelling
  - – Similar challenges as public data
- Access restriction
  - – Provided by standard mechanisms
  - – Graph-based access
- Open PHACTS Discovery Platform
  - – Releasing version 1.3 (late 2013)
  - – Version 1.4 will contain commercial data (2014)

Acknowledgements

- GVK Bio
  GOSTAR
  gostardb.com
- Thomson Reuters
  Integrity
  integrity.thomson-pharma.com
- Aureus Sciences Elsevier
  AurSCOPE
  www.aureus-sciences.com

Questions

Open PHACTS
Open Pharmacological Space

A.J.G.Gray@hw.ac.uk
www.macs.hw.ac.uk/~ajg33
@gray_alasdair

pmu@openphacts.org
www.openphacts.org
@open_phacts

Open PHACTS Project

Pfizer
universität wien
RSC | Advancing the Chemical Sciences
130 jaar VU VRIJE UNIVERSITEIT AMSTERDAM
AstraZeneca
gsk GlaxoSmithKline
DTU
Universität Hamburg
cnio
MANCHESTER 1824
ESTEVE closer to you
NOVARTIS
35th anniversary Maastricht University Leading in Learning!
AQnowledge Semantics for Science
MERCK
Lundbeck
Bio Solve IT
Parc de Salut MAR
USC
universitätbonn
Lilly
LUMC
janssen
EMBL-EBI
connecteddiscovery
nbic
OPENLINK SOFTWARE Making Technology Work For You®
SIB Swiss Institute of Bioinformatics