Too Much Data or Too Little Cooperation?

Tom Plasterer, PhD.
Research & Development Information (RDI)
Director, US Cross-Science

BigDiP USA 2015
Analytics & Insights for Pharma

AstraZeneca

Approaching a Pharma Big Data Problem: Requirements of the CI Informatics Landscape

Must span the entire drug development lifecycle

- and back (post-market surveillance to discovery)

Must support large and very heterogeneous da

- single nucleotide polymorphisms to countries

Will change with new science & new regulation

- Medline just under 1M articles/year

Must work with multiple, international regulator

- Emerging markets

Partners, customers and collaborators will cha

- and will have divergent technical aptitudes

Must be work with precompetitive consortia

- Can they perform common tasks for the community

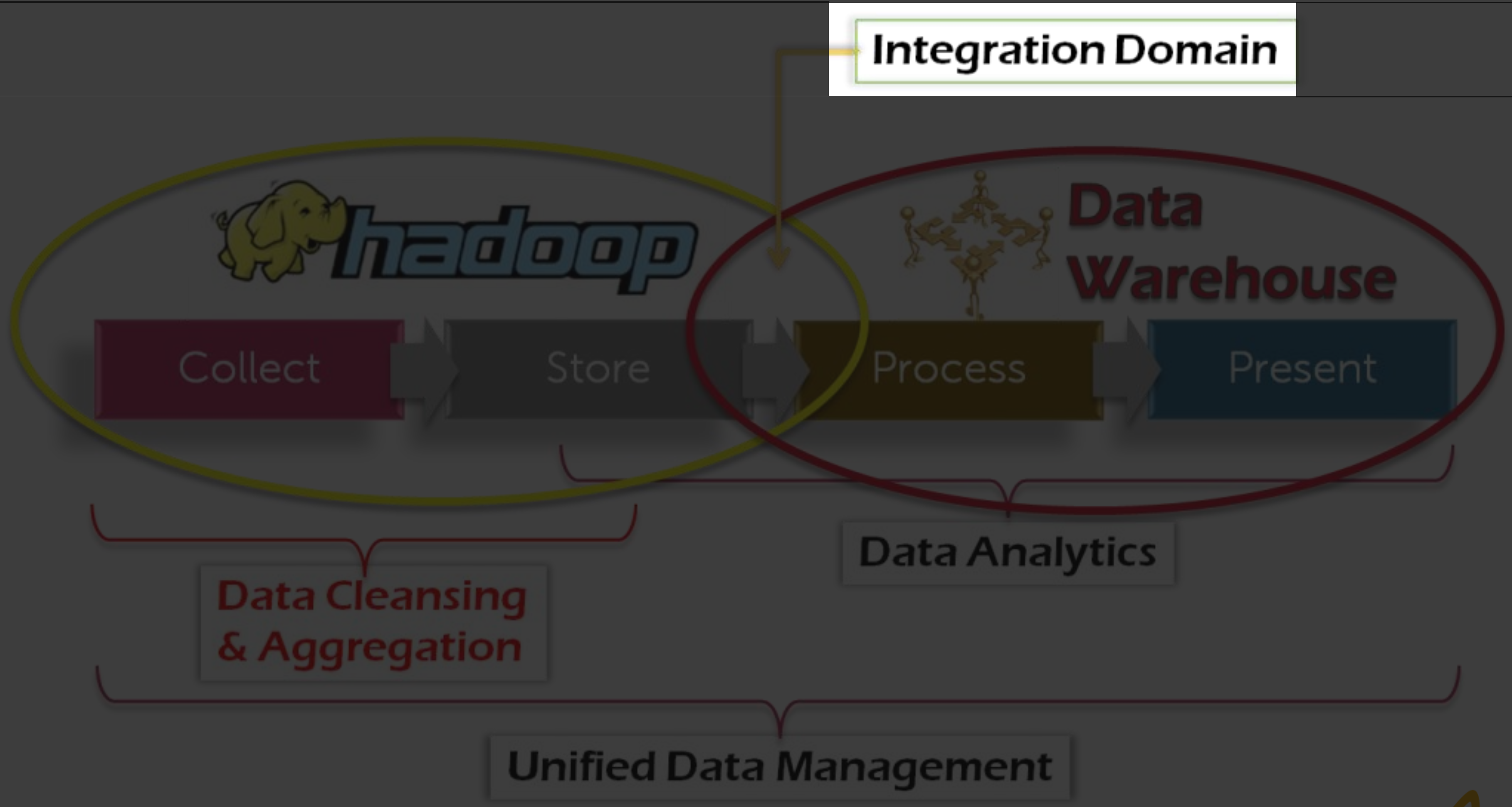
Must be able to work with legacy data

- Lots of unmined gems here!

These are
Big Data
Variety
and
Veracity
Challenges

Typical Big Data Integration Process

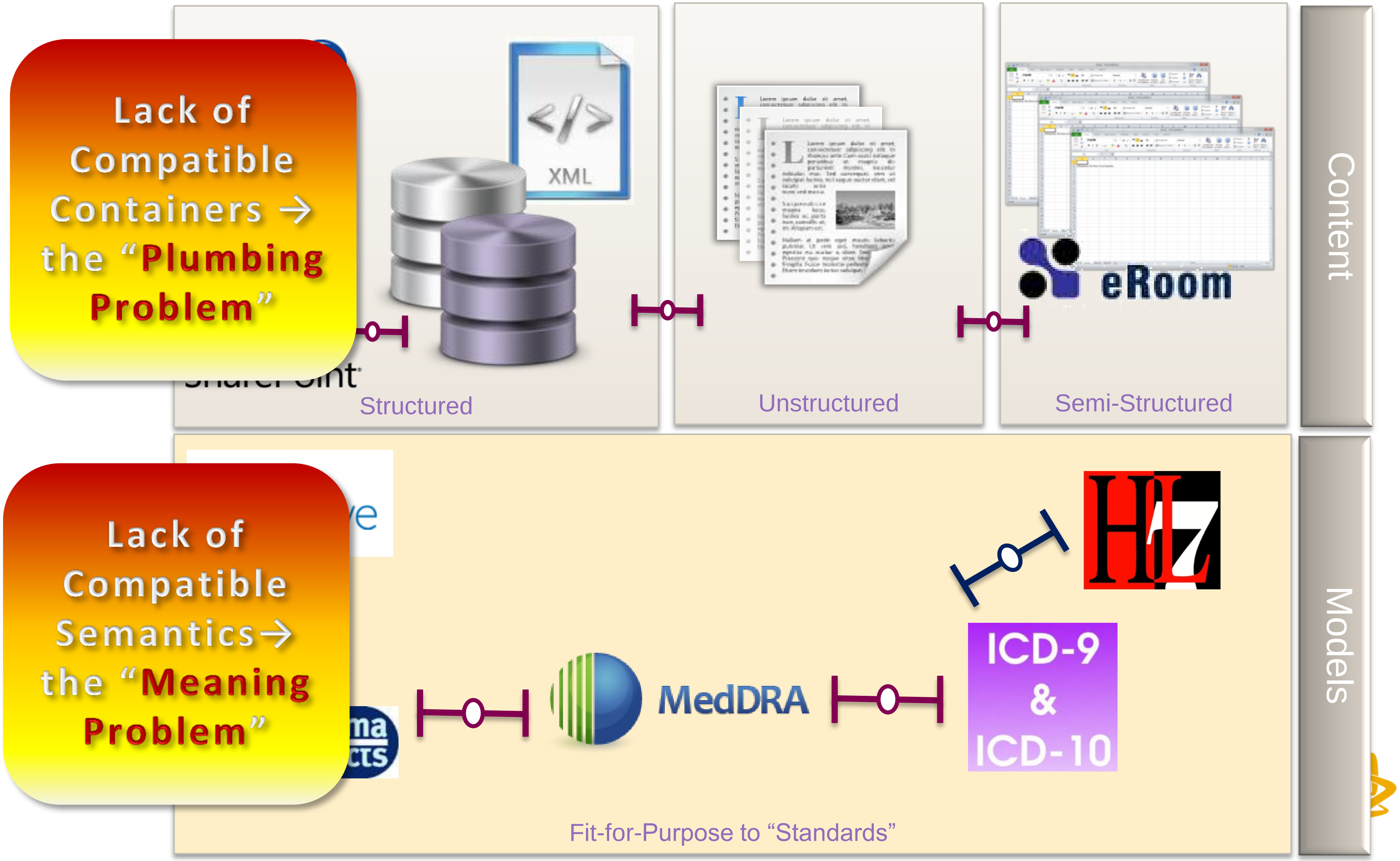
Document-Centric model



From: Hadoop/EDW Integration session at OSCON 2011
http://rationalintelligence.com/wp_log/?p=44

Integration Quandary:

Content Does Not Combine Easily



Lack of Compatible Containers → the "Plumbing Problem"

Lack of Compatible Semantics→ the "Meaning Problem"

Structured

Unstructured

Semi-Structured

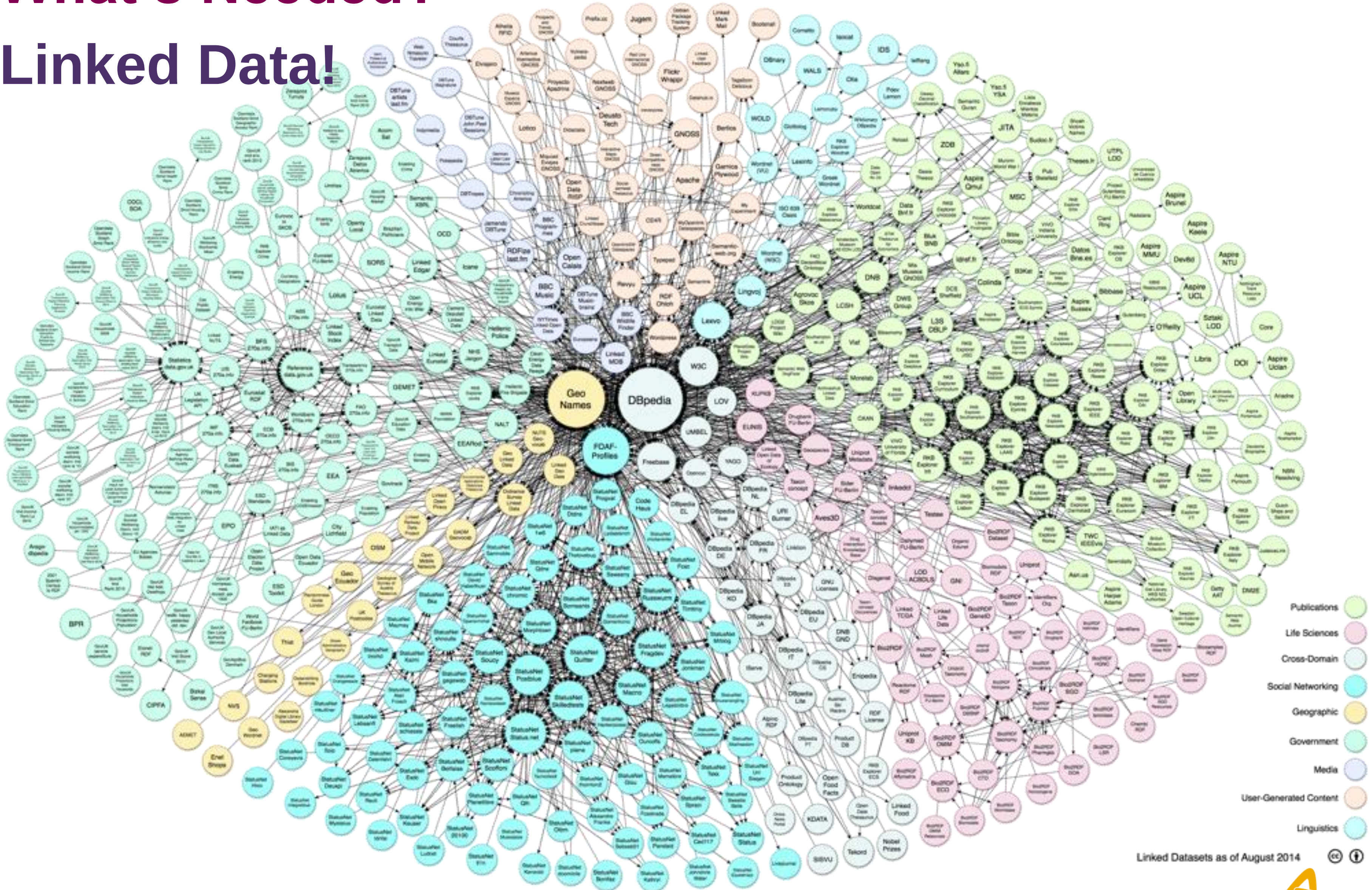
Content

Fit-for-Purpose to "Standards"

Models

What’s Needed?

Linked Data!



Schmachtenberg, Bizer, Jentzsch and Cyganiak.
http://lod-cloud.net/

LOD Cloud 2014

Linked Data Demystified

Addresses *Plumbing* and *Meaning* Challenges

A Graph is the fundamental data model

- Not a table or a hierarchy or a document

This model uses RDF* and is the web data model

- The underlying content need not be RDF only its published interface

Web uniform resource identifiers (URIs) name things

- Resolving the URI gives a useful description

URIs link data and integrate with other Linked Data

- Two things sharing the same URI are the same thing

Fit-for-Purpose yet Scalable Applications are enabled

- Easy to mash-up and scales as the web scales

Features Flexible & Adaptable Information Models

- You can change the data model without breaking downstream applications

Encourages Shared Understanding via shared vocabularies

- Communities build out as needed to support their business questions

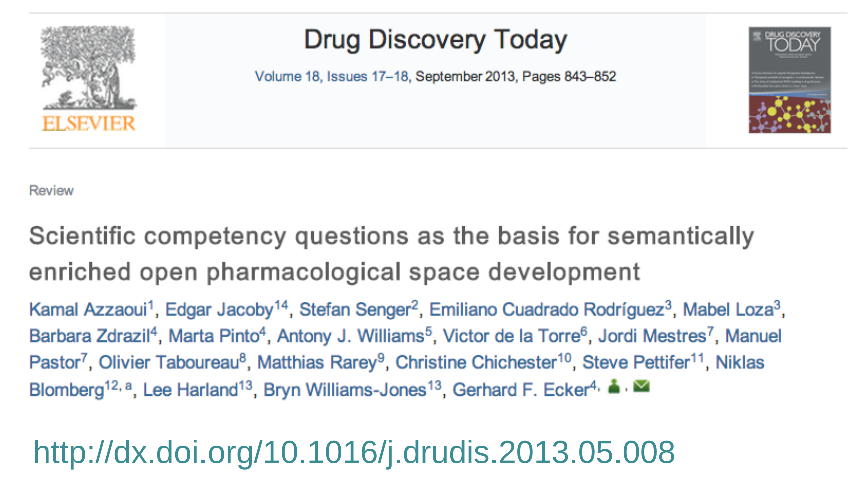
Solution: IMI Open PHACTS Project

Open PHACTS Mission:

Integrate Multiple Research Biomedical Data Resources Into A Single Open, Sustainable & Free Access Point

The Open PHACTS Discovery Platform

- **Cloud-Based "Production" Level System. Secure & Private**
- **Guided By Business Questions**
- **Uses Semantic Web Technology And provides a simple REST-ful API for the everyone else**



Drug Discovery Today

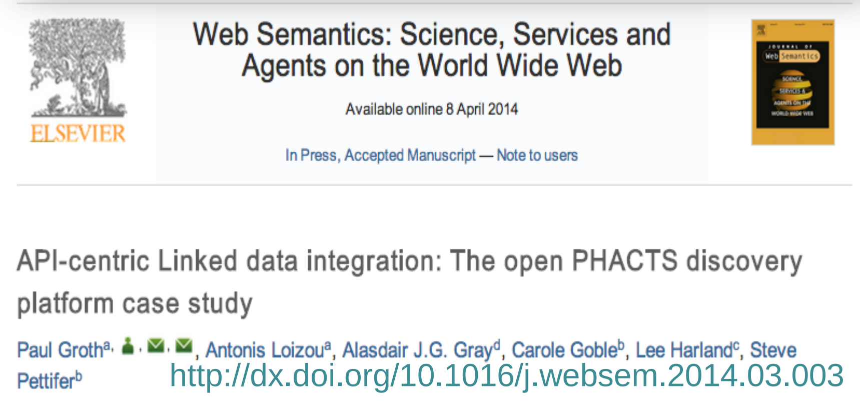
Volume 18, Issues 17–18, September 2013, Pages 843–852

Review

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

Kamal Azzaoui[1], Edgar Jacoby[14], Stefan Senger[2], Emiliano Cuadrado Rodríguez[3], Mabel Loza[3], Barbara Zdrazil[4], Marta Pinto[4], Antony J. Williams[5], Victor de la Torre[6], Jordi Mestres[7], Manuel Pastor[7], Olivier Taboureau[8], Matthias Rarey[9], Christine Chichester[10], Steve Pettifer[11], Niklas Blomberg[12, a], Lee Harland[13], Bryn Williams-Jones[13], Gerhard F. Ecker[4]

http://dx.doi.org/10.1016/j.drudis.2013.05.008



Web Semantics: Science, Services and Agents on the World Wide Web

Available online 8 April 2014

In Press, Accepted Manuscript — Note to users

API-centric Linked data integration: The open PHACTS discovery platform case study

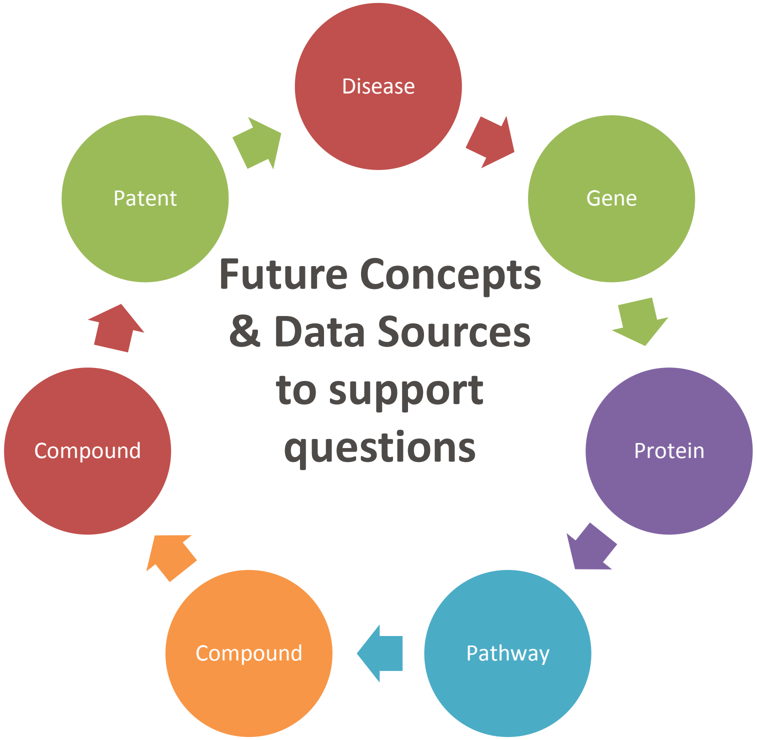
Paul Groth[a], Antonis Loizou[a], Alasdair J.G. Gray[d], Carole Goble[b], Lee Harland[c], Steve Pettifer[b]

http://dx.doi.org/10.1016/j.websem.2014.03.003

Open PHACTS
Open Pharmacological Space

Open PHACTS has built a cutting edge, unique, flexible and powerful infrastructure for semantic data in life sciences We have the platform, data, services, experience and capabilities to tackle big data challenges

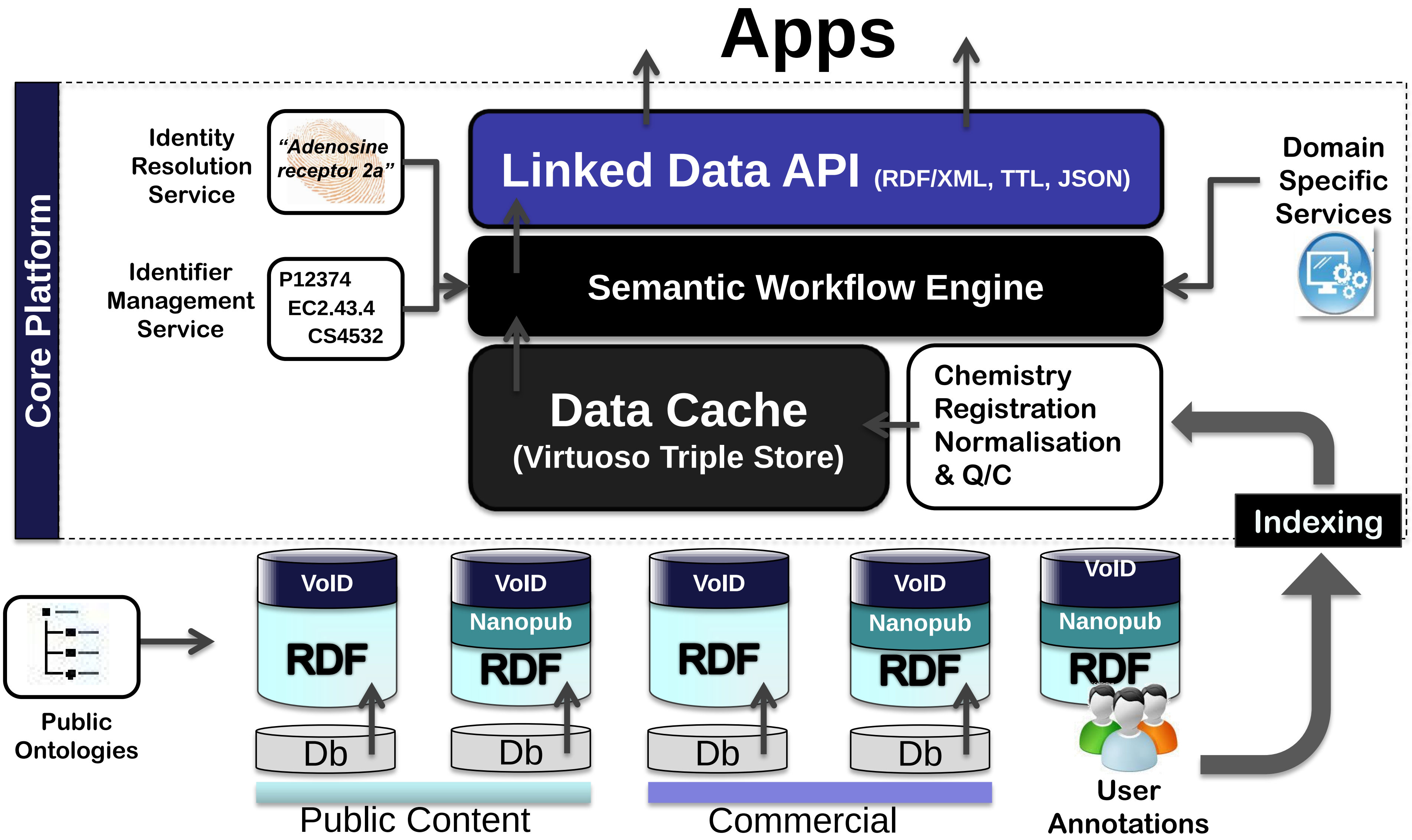
- Cloud-based ‘**production**’ level secure and private system
- We are expanding into new data areas
- Pathways, patents, disease
- Comprehensive workflow components for advanced use cases
- Open PHACTS is big data



Disease
Gene
Protein
Pathway
Compound
Compound
Patent

Future Concepts & Data Sources to support questions

http://www.openphactsfoundation.org/



Apps

Core Platform

Identity Resolution Service

“Adenosine receptor 2a”

Identifier Management Service

P12374
EC2.43.4
CS4532

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

Semantic Workflow Engine

Domain Specific Services

Data Cache
(Virtuoso Triple Store)

Chemistry Registration Normalisation & Q/C

Indexing

Public Ontologies

VoID

RDF

Db

VoID

Nanopub

RDF

Db

VoID

RDF

Db

VoID

Nanopub

RDF

Db

VoID

Nanopub

RDF

User Annotations

Public Content

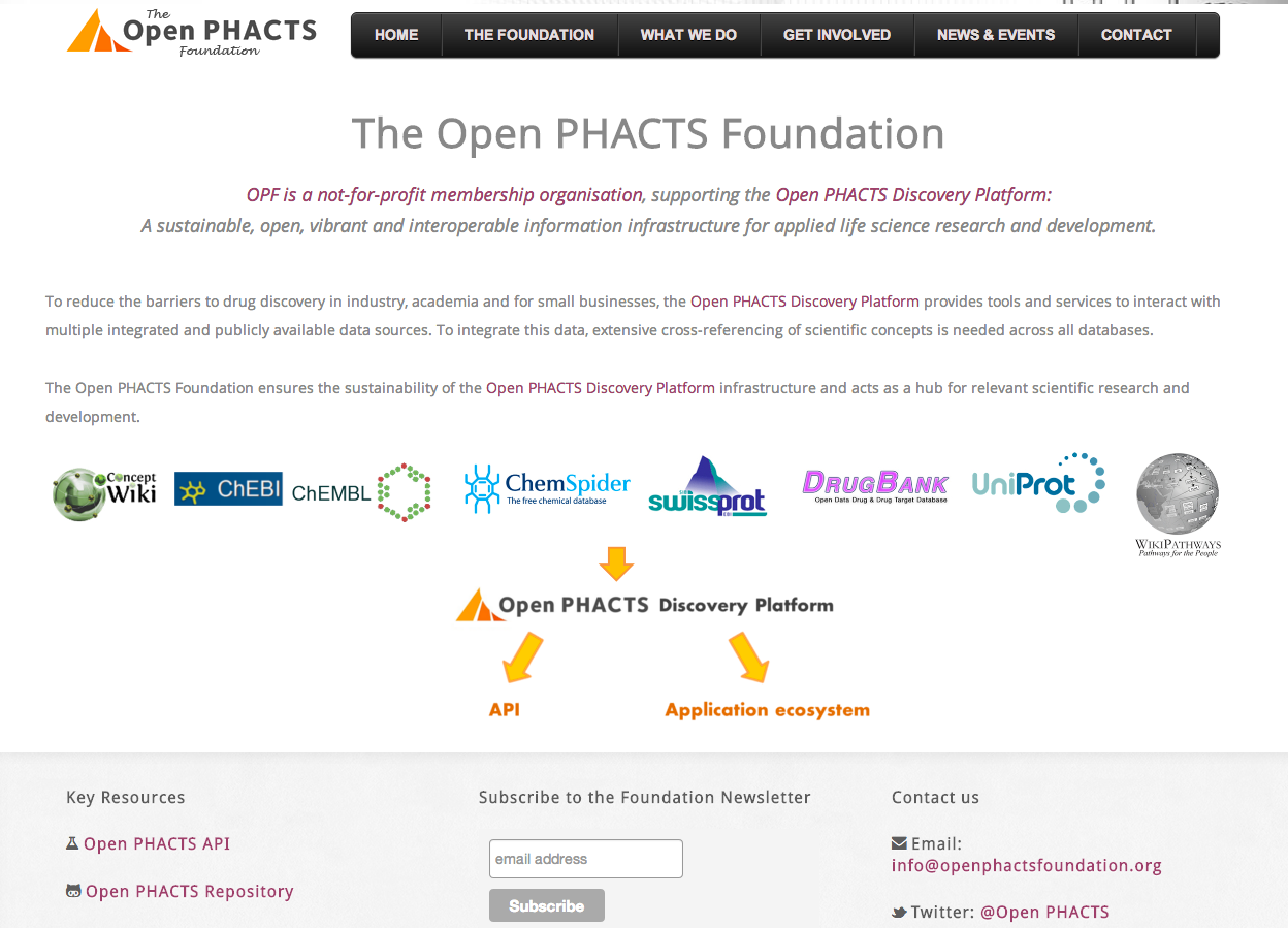
Commercial

- Delivered on all of the IMI project deliverables on or ahead of time
- Open PHACTS Foundation is a UK-based member-owned non-profit company founded to sustain and develop the Open PHACTS infrastructure
- OPF is now a registered charity with aims to further the public understanding of science through research
- OPF operations are distributed across funded partners to maintain and further develop public:private partnership
- A sought-after partner for Horizon 2020 projects
- Bring industry perspective to new academic partners
- Leverage the heritage of open data services and public:private partnership
- Significant academic research and professional network of subject matter experts

BIG DATA EUROPE
Empowering Communities
with Data Technologies

OPF is a funded partner in the BDE project, another project approved and completing grant agreement, and 4 other pending submissions

http://www.openphactsfoundation.org/



The Open PHACTS Foundation

OPF is a not-for-profit membership organisation, supporting the Open PHACTS Discovery Platform:
A sustainable, open, vibrant and interoperable information infrastructure for applied life science research and development.

To reduce the barriers to drug discovery in industry, academia and for small businesses, the Open PHACTS Discovery Platform provides tools and services to interact with multiple integrated and publicly available data sources. To integrate this data, extensive cross-referencing of scientific concepts is needed across all databases.

The Open PHACTS Foundation ensures the sustainability of the Open PHACTS Discovery Platform infrastructure and acts as a hub for relevant scientific research and development.

Key Resources

Open PHACTS API

Open PHACTS Repository

Subscribe to the Foundation Newsletter

email address

Subscribe

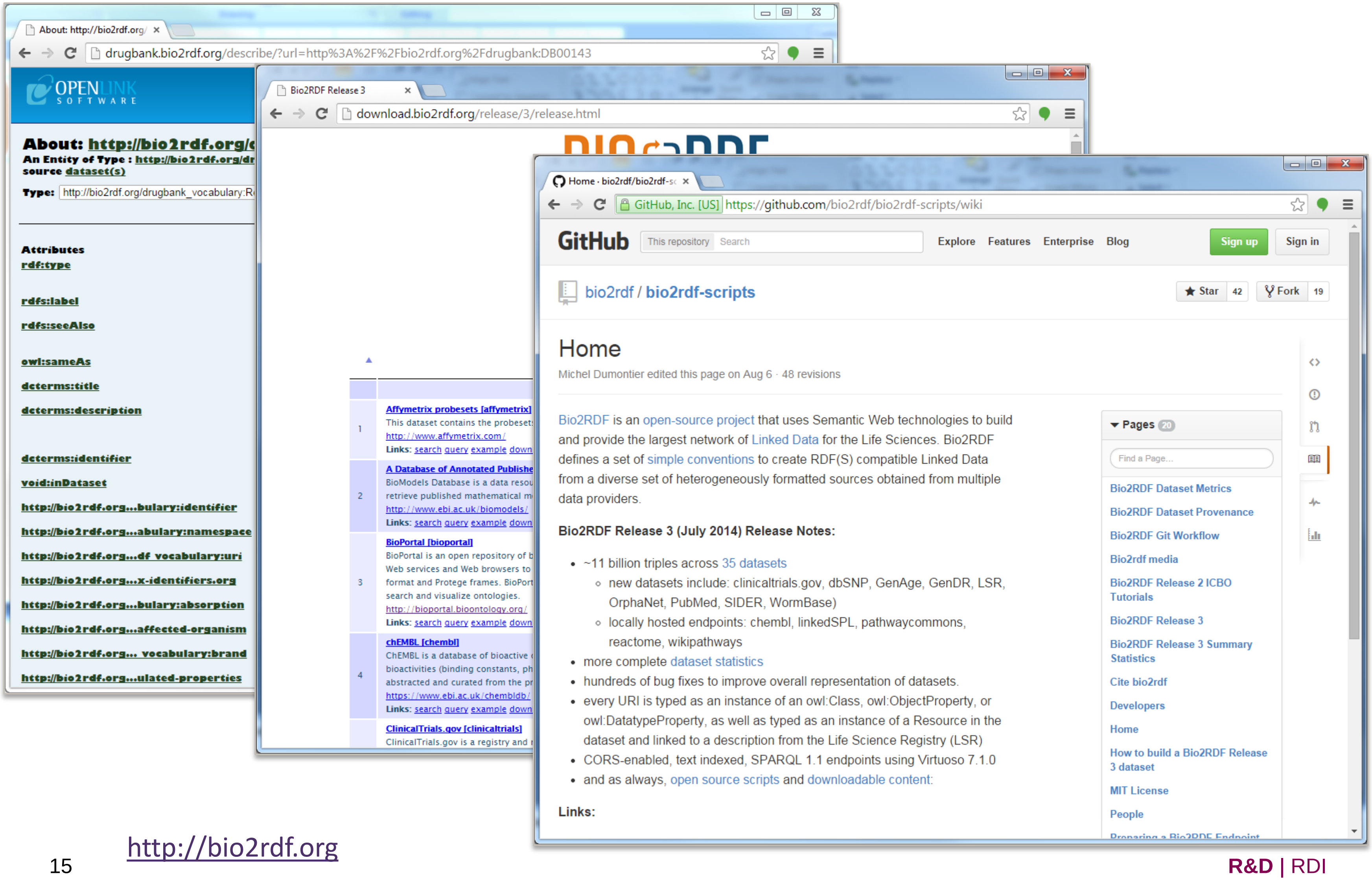
Contact us

Email:
info@openphactsfoundation.org

Twitter: @Open PHACTS

Solution: Emerging Public Solutions

BIO2RDF



About: http://bio2rdf.org/

drugbank.bio2rdf.org/describe/?url=http%3A%2F%2Fbio2rdf.org%2Fdrugbank:DB00143

OPENLINK SOFTWARE

About: http://bio2rdf.org/

An Entity of Type : http://bio2rdf.org/dr

source dataset(s)

Type: http://bio2rdf.org/drugbank_vocabulary:R

Attributes

rdf:type

rdfs:label

rdfs:seeAlso

owl:sameAs

dcterms:title

dcterms:description

dcterms:identifier

void:inDataset

http://bio2rdf.org...bulary:identifier

http://bio2rdf.org...abulary:namespace

http://bio2rdf.org...df_vocabulary:uri

http://bio2rdf.org...x-identifiers.org

http://bio2rdf.org...bulary:absorption

http://bio2rdf.org...affected-organism

http://bio2rdf.org..._vocabulary:brand

http://bio2rdf.org...ulated-properties

Bio2RDF Release 3

download.bio2rdf.org/release/3/release.html

1	**Affymetrix probesets [affymetrix]** This dataset contains the probesets http://www.affymetrix.com/ **Links:** search query example down
2	**A Database of Annotated Publishe** BioModels Database is a data resou retrieve published mathematical m http://www.ebi.ac.uk/biomodels/ **Links:** search query example down
3	**BioPortal [bioportal]** BioPortal is an open repository of b Web services and Web browsers to format and Protege frames. BioPort search and visualize ontologies. http://bioportal.bioontology.org/ **Links:** search query example down
4	**chEMBL [chembl]** ChEMBL is a database of bioactive bioactivities (binding constants, ph abstracted and curated from the pr https://www.ebi.ac.uk/chembldb/ **Links:** search query example down
	ClinicalTrials.gov [clinicaltrials] ClinicalTrials.gov is a registry and

Home · bio2rdf/bio2rdf-sc

GitHub, Inc. [US] https://github.com/bio2rdf/bio2rdf-scripts/wiki

Home

Michel Dumontier edited this page on Aug 6 · 48 revisions

Bio2RDF is an open-source project that uses Semantic Web technologies to build and provide the largest network of Linked Data for the Life Sciences. Bio2RDF defines a set of simple conventions to create RDF(S) compatible Linked Data from a diverse set of heterogeneously formatted sources obtained from multiple data providers.

Bio2RDF Release 3 (July 2014) Release Notes:

- ~11 billion triples across 35 datasets
  - new datasets include: clinicaltrials.gov, dbSNP, GenAge, GenDR, LSR, OrphaNet, PubMed, SIDER, WormBase)
  - locally hosted endpoints: chembl, linkedSPL, pathwaycommons, reactome, wikipathways
- more complete dataset statistics
- hundreds of bug fixes to improve overall representation of datasets.
- every URI is typed as an instance of an owl:Class, owl:ObjectProperty, or owl:DatatypeProperty, as well as typed as an instance of a Resource in the dataset and linked to a description from the Life Science Registry (LSR)
- CORS-enabled, text indexed, SPARQL 1.1 endpoints using Virtuoso 7.1.0
- and as always, open source scripts and downloadable content:

Links:

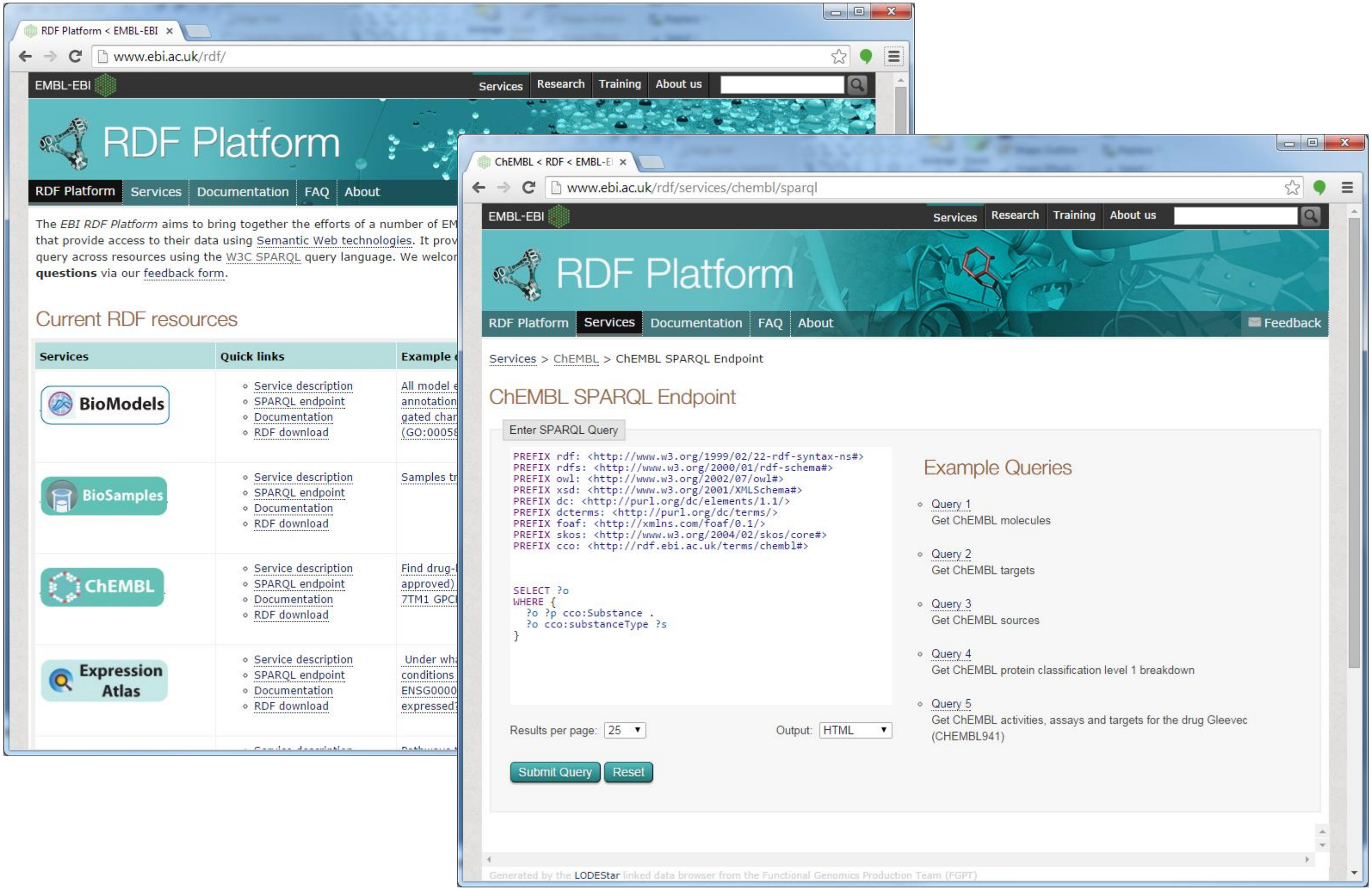
Pages 20

Find a Page...

- Bio2RDF Dataset Metrics
- Bio2RDF Dataset Provenance
- Bio2RDF Git Workflow
- Bio2rdf media
- Bio2RDF Release 2 ICBO Tutorials
- Bio2RDF Release 3
- Bio2RDF Release 3 Summary Statistics
- Cite bio2rdf
- Developers
- Home
- How to build a Bio2RDF Release 3 dataset
- MIT License
- People
- Preparing a Bio2RDF Endpoint

http://bio2rdf.org

EMBL-EBI RDF Platform



The *EBI RDF Platform* aims to bring together the efforts of a number of EM
that provide access to their data using Semantic Web technologies. It prov
query across resources using the W3C SPARQL query language. We welcor
questions via our feedback form.

Current RDF resources

Services	Quick links	Example
BioModels	○ Service description ○ SPARQL endpoint ○ Documentation ○ RDF download	All model e annotation gated char (GO:00058
BioSamples	○ Service description ○ SPARQL endpoint ○ Documentation ○ RDF download	Samples tr
ChEMBL	○ Service description ○ SPARQL endpoint ○ Documentation ○ RDF download	Find drug- approved) 7TM1 GPC
Expression Atlas	○ Service description ○ SPARQL endpoint ○ Documentation ○ RDF download	Under wh conditions ENSG0000 expressed?

Services > ChEMBL > ChEMBL SPARQL Endpoint

ChEMBL SPARQL Endpoint

Enter SPARQL Query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX cco: <http://rdf.ebi.ac.uk/terms/chembl#>



SELECT ?o
WHERE {
  ?o ?p cco:Substance .
  ?o cco:substanceType ?s
}
```

Results per page: 25 ▾

Output: HTML ▾

Submit Query Reset

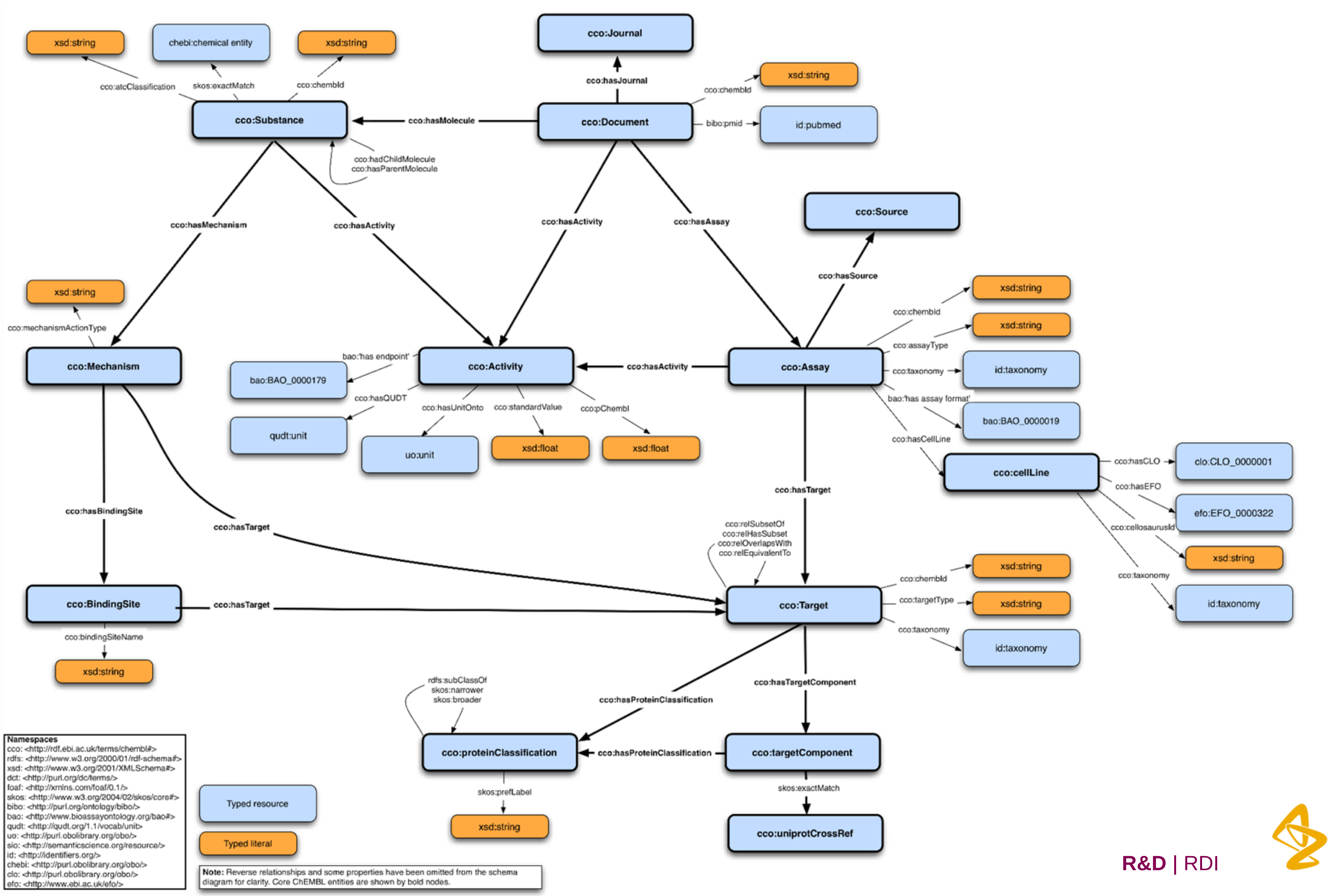
Example Queries

- Query 1
  Get ChEMBL molecules
- Query 2
  Get ChEMBL targets
- Query 3
  Get ChEMBL sources
- Query 4
  Get ChEMBL protein classification level 1 breakdown
- Query 5
  Get ChEMBL activities, assays and targets for the drug Gleevec (CHEMBL941)

Generated by the LODEStar linked data browser from the Functional Genomics Production Team (FGPT)

http://www.ebi.ac.uk/rdf/

ChEMBL 18 RDF Model

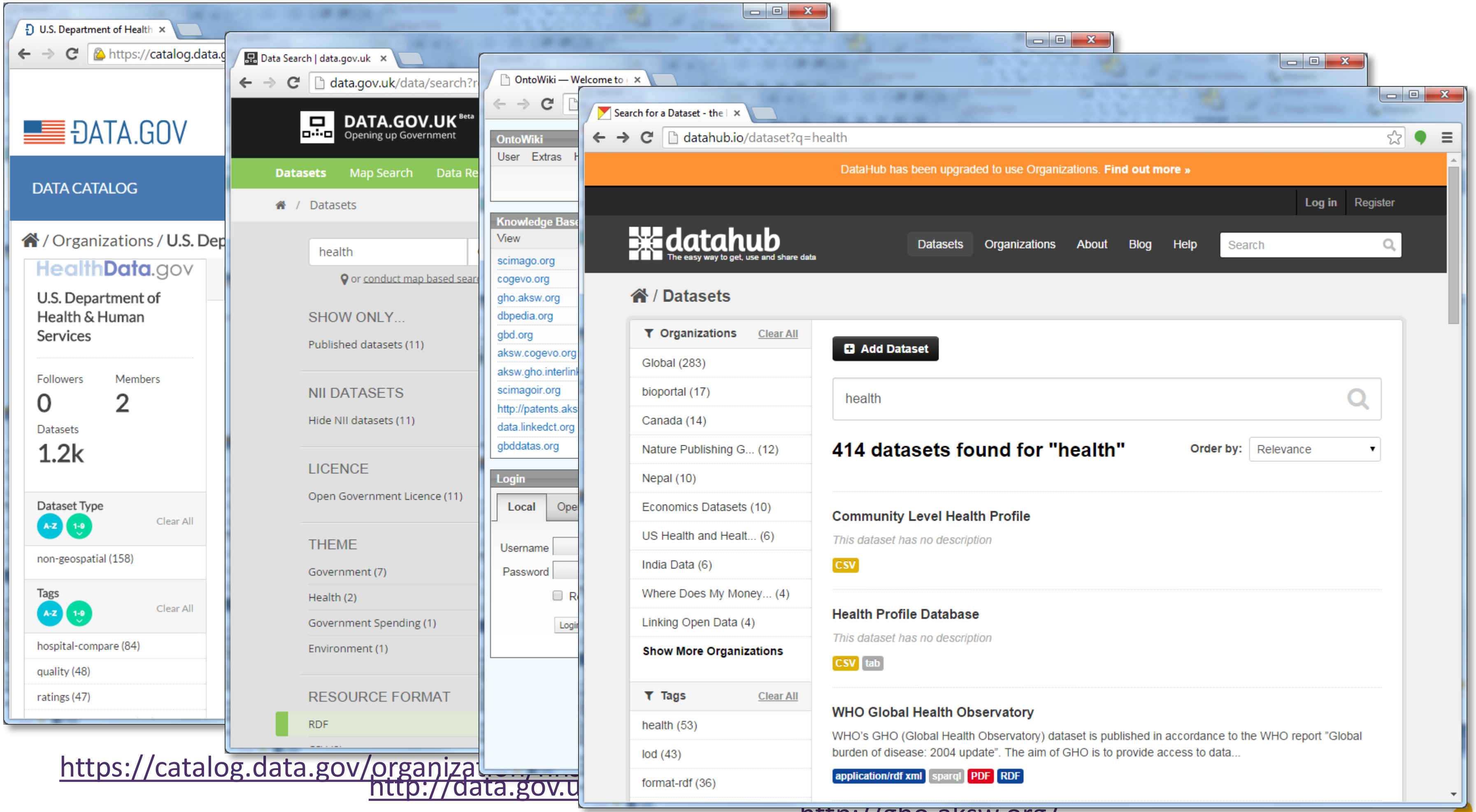


Namespaces
cco: <http://rdf.ebi.ac.uk/terms/chembl#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
xsd: <http://www.w3.org/2001/XMLSchema#>
dct: <http://purl.org/dc/terms/>
foaf: <http://xmlns.com/foaf/0.1/>
skos: <http://www.w3.org/2004/02/skos/core#>
bibo: <http://purl.org/ontology/bibo/>
bao: <http://www.bioassayontology.org/bao#>
qudt: <http://qudt.org/1.1/vocab/unit>
uo: <http://purl.obolibrary.org/obo/>
sio: <http://semanticscience.org/resource/>
id: <http://identifiers.org/>
chebi: <http://purl.obolibrary.org/obo/>
clo: <http://purl.obolibrary.org/obo/>
efo: <http://www.ebi.ac.uk/efo/>

Note: Reverse relationships and some properties have been omitted from the schema diagram for clarity. Core ChEMBL entities are shown by bold nodes.

National, International Health Systems

Data.gov, Data.gov.uk, WHO, datahub.io



U.S. Department of Health ×

https://catalog.data.g

DATA.GOV

DATA CATALOG

/ Organizations / U.S. Dep

HealthData.gov

U.S. Department of Health & Human Services

Followers 0

Members 2

Datasets 1.2k

Dataset Type A-Z 1-9 Clear All

non-geospatial (158)

Tags A-Z 1-9 Clear All

hospital-compare (84)

quality (48)

ratings (47)

Data Search | data.gov.uk ×

data.gov.uk/data/search?r

DATA.GOV.UK Beta
Opening up Government

Datasets Map Search Data Re

/ Datasets

health

or conduct map based sear

SHOW ONLY...

Published datasets (11)

NII DATASETS

Hide NII datasets (11)

LICENCE

Open Government Licence (11)

THEME

Government (7)

Health (2)

Government Spending (1)

Environment (1)

RESOURCE FORMAT

RDF

OntoWiki — Welcome to ×

OntoWiki

User Extras H

Knowledge Base

View

scimago.org

cogevo.org

gho.aksw.org

dbpedia.org

gbd.org

aksw.cogevo.org

aksw.gho.interlink

scimagoir.org

http://patents.aks

data.linkedct.org

gbddatas.org

Login

Local Ope

Username

Password

□ R

Login

Search for a Dataset - the ×

datahub.io/dataset?q=health

DataHub has been upgraded to use Organizations. **Find out more »**

Log in Register

datahub
The easy way to get, use and share data

Datasets Organizations About Blog Help Search

/ Datasets

Organizations Clear All

Global (283)

bioportal (17)

Canada (14)

Nature Publishing G... (12)

Nepal (10)

Economics Datasets (10)

US Health and Healt... (6)

India Data (6)

Where Does My Money... (4)

Linking Open Data (4)

Show More Organizations

Tags Clear All

health (53)

lod (43)

format-rdf (36)

Add Dataset

health

414 datasets found for "health"

Order by: Relevance

Community Level Health Profile

This dataset has no description

CSV

Health Profile Database

This dataset has no description

CSV tab

WHO Global Health Observatory

WHO's GHO (Global Health Observatory) dataset is published in accordance to the WHO report "Global burden of disease: 2004 update". The aim of GHO is to provide access to data...

application/rdf xml sparql PDF RDF

https://catalog.data.gov/organizat

http://data.gov.u

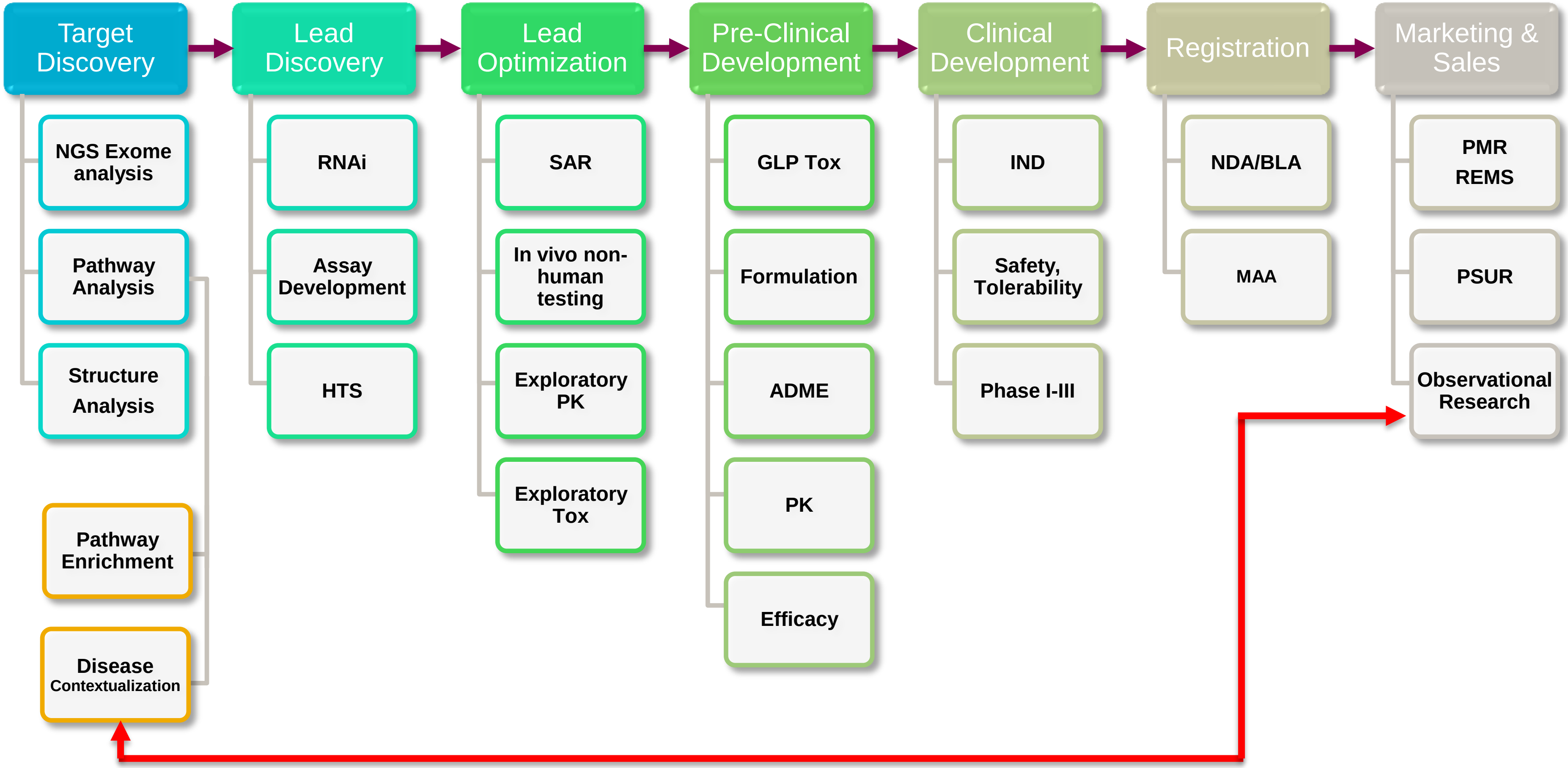
http://gho.aksw.org/

http://datahub.io/dataset?q=health

Best Practices

Node and Edge Informatics

Interfaces within the Drug Development Process



Target Discovery → Lead Discovery → Lead Optimization → Pre-Clinical Development → Clinical Development → Registration → Marketing & Sales

Target Discovery
- NGS Exome analysis
- Pathway Analysis
- Structure Analysis
- Pathway Enrichment
- Disease Contextualization

Lead Discovery
- RNAi
- Assay Development
- HTS

Lead Optimization
- SAR
- In vivo non-human testing
- Exploratory PK
- Exploratory Tox

Pre-Clinical Development
- GLP Tox
- Formulation
- ADME
- PK
- Efficacy

Clinical Development
- IND
- Safety, Tolerability
- Phase I-III

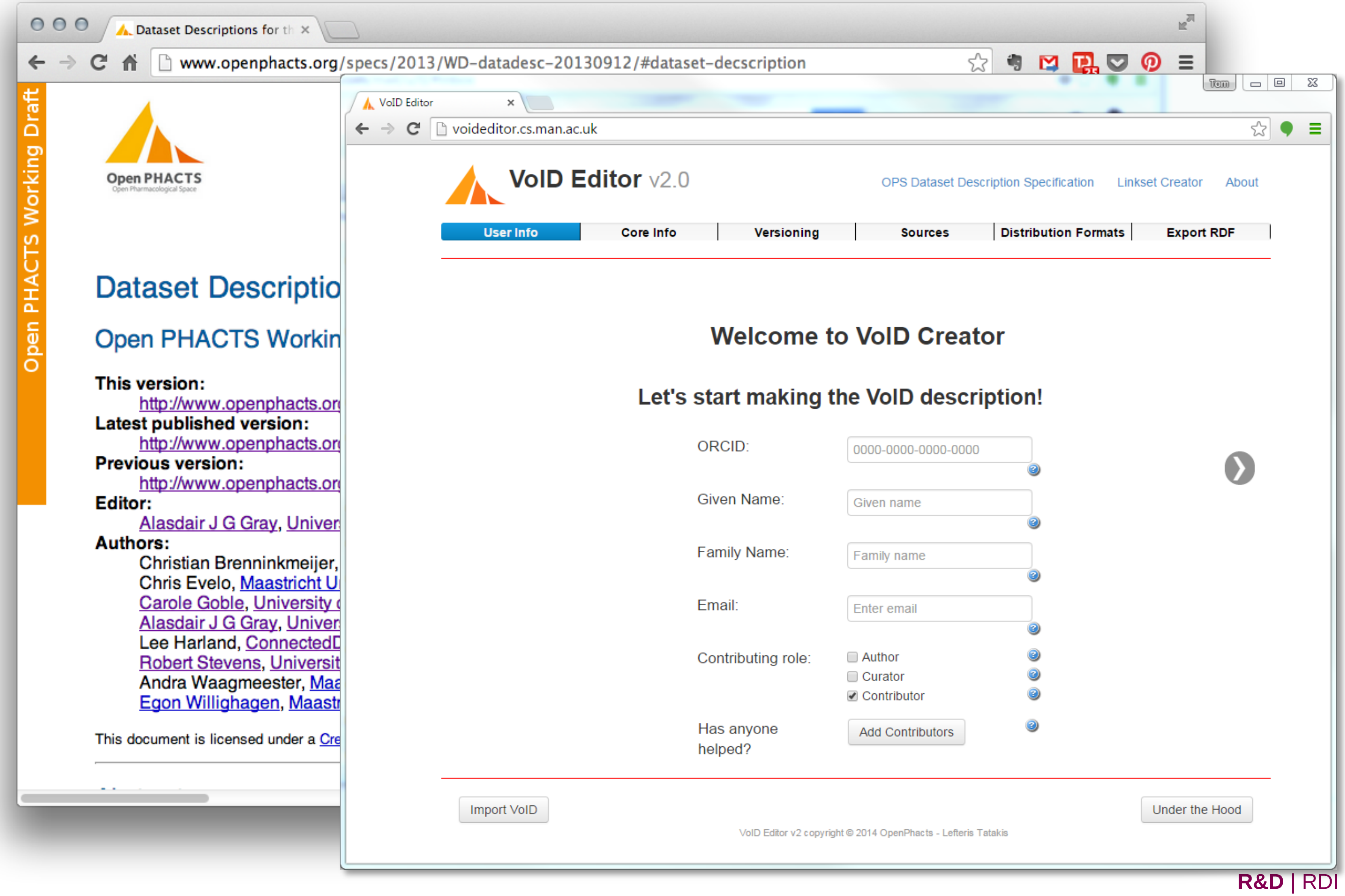
Registration
- NDA/BLA
- MAA

Marketing & Sales
- PMR REMS
- PSUR
- Observational Research

Seamless information connectivity (an EDGE) needed across domain NODEs

It's the Data…

…not the app, not the container



Dataset Descriptions for th ×
www.openphacts.org/specs/2013/WD-datadesc-20130912/#dataset-decscription
Open PHACTS Working Draft
Open PHACTS
Open Pharmacological Space

Dataset Descriptio

Open PHACTS Workin

This version:
http://www.openphacts.or
Latest published version:
http://www.openphacts.or
Previous version:
http://www.openphacts.or
Editor:
Alasdair J G Gray, Univer
Authors:
Christian Brenninkmeijer,
Chris Evelo, Maastricht U
Carole Goble, University
Alasdair J G Gray, Univer
Lee Harland, ConnectedD
Robert Stevens, Universit
Andra Waagmeester, Maa
Egon Willighagen, Maastr

This document is licensed under a Cre

Welcome to VoID Creator

Let's start making the VoID description!

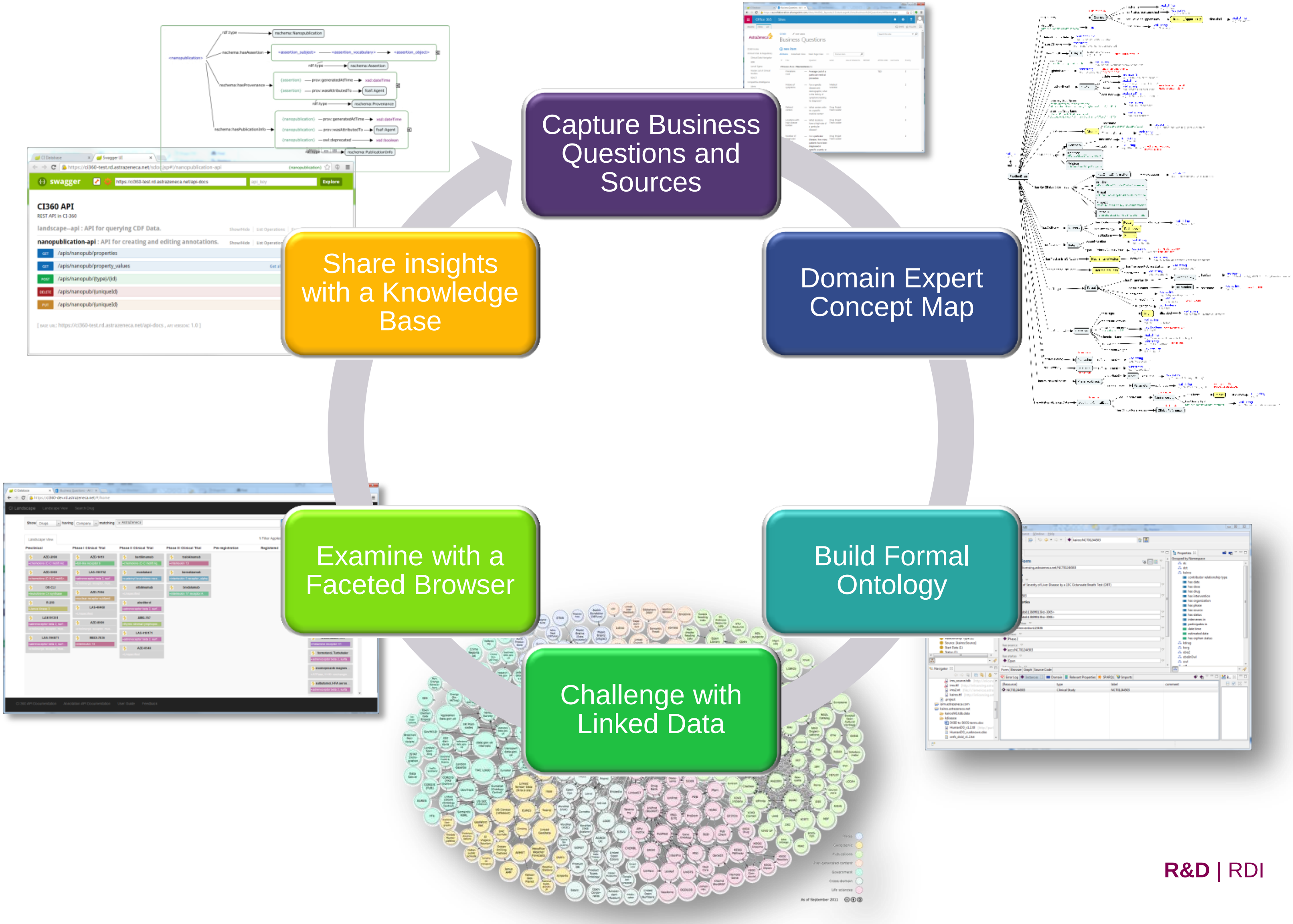
ORCID:	0000-0000-0000-0000
Given Name:	Given name
Family Name:	Family name
Email:	Enter email
Contributing role:	□ Author □ Curator ☑ Contributor
Has anyone helped?	Add Contributors

VoID Editor v2 copyright © 2014 OpenPhacts - Lefteris Tatakis

R&D | RDI

Disposable Applications

Questions, Answers, Insights Persist



Capture Business Questions and Sources

Domain Expert Concept Map

Build Formal Ontology

Challenge with Linked Data

Examine with a Faceted Browser

Share insights with a Knowledge Base

CI Database
Swagger UI
https://ci360-test.rd.astrazeneca.net/sdoc.jsp#!/nanopublication-api
swagger
https://ci360-test.rd.astrazeneca.net/api-docs
api_key
Explore
Show/Hide
List Operations

CI360 API

REST API in CI-360

landscape--api : API for querying CDF Data.

nanopublication-api : API for creating and editing annotations.

GET /apis/nanopub/properties

GET /apis/nanopub/property_values

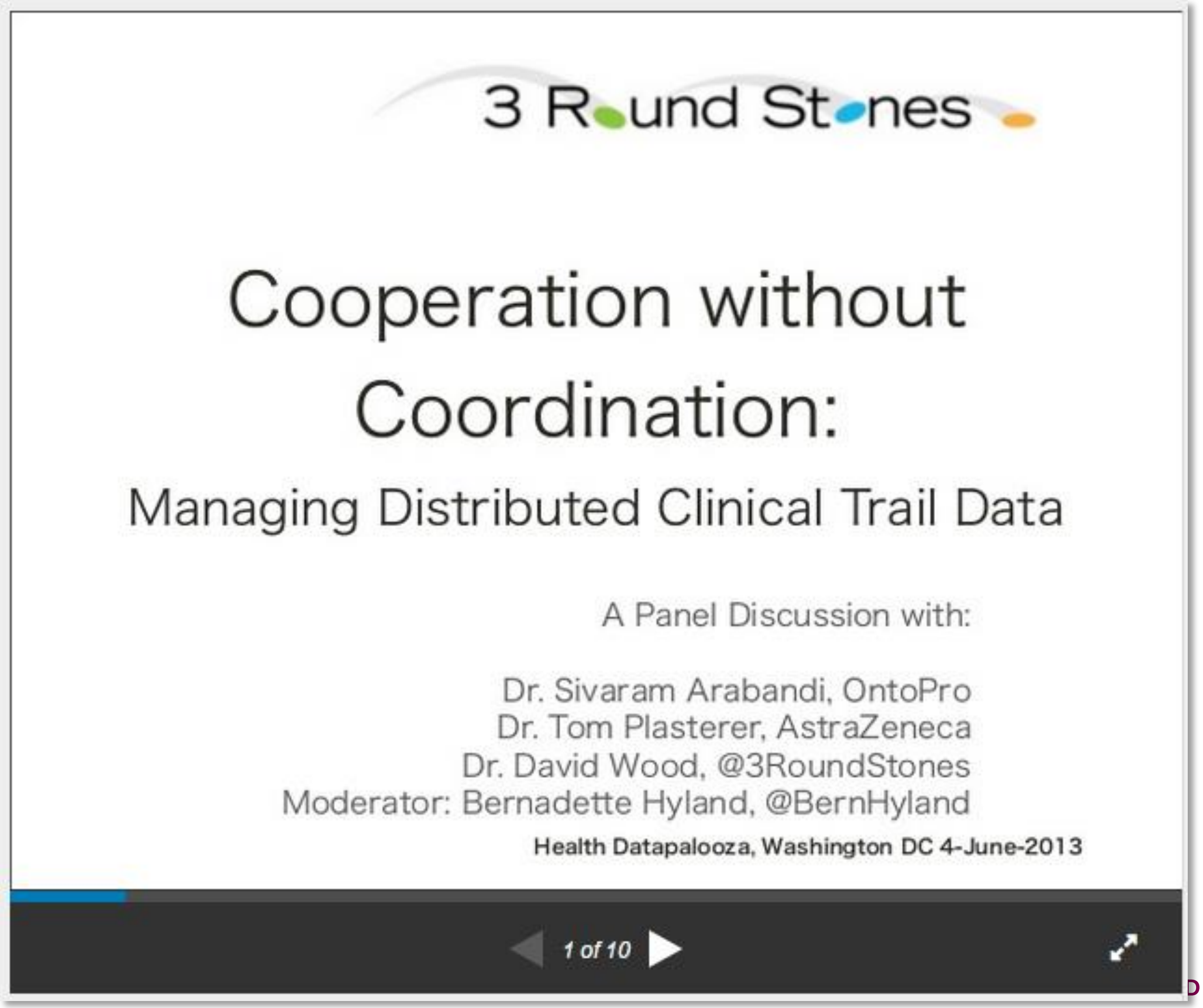
POST /apis/nanopub/{type}/{id}

DELETE /apis/nanopub/{uniqueId}

PUT /apis/nanopub/{uniqueId}

[BASE URL: https://ci360-test.rd.astrazeneca.net/api-docs , API VERSION: 1.0]

Cooperation…
…without Coordination



3 Round Stones

Cooperation without Coordination:

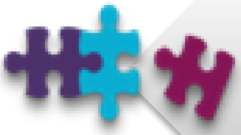
Managing Distributed Clinical Trail Data

A Panel Discussion with:

Dr. Sivaram Arabandi, OntoPro
Dr. Tom Plasterer, AstraZeneca
Dr. David Wood, @3RoundStones
Moderator: Bernadette Hyland, @BernHyland

Health Datapalooza, Washington DC 4-June-2013

Take-Aways



Get your plumbing right

- And you won't be stuck in a silo

Leverage working public solutions

- No need to reinvent the wheel

Use Edge Informatics

- Consider handoffs—you don't know how your data will be used in the future

Thanks

Big Data in Pharma 2015
Conference Organizers

AZ Linked Data Community

Key Influencers
David Wood
Toby Segaran
Tim Berners-Lee
Lee Harland
Bryn Williams-Jones
Eric Neumann
Dean Allemang
Barend Mons
Bernadette Hyland
Bob Stanley