Too Much Data or Too Little Cooperation?
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 data single nucleotide polymorphisms to countries
- Will change with new science & new regulation Medline just under 1M articles/year
- Must work with multiple, international regulatory bodies Emerging markets
- Partners, customers and collaborators will change 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"

Fit-for-Purpose to "Standards"
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 everyone else

http://dx.doi.org/10.1016/j.drudis.2013.05.008
http://dx.doi.org/10.1016/j.websem.2014.03.003
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

http://www.openphactsfoundation.org/
Open PHACTS
Open Pharmacological Space

Core Platform

Identity Resolution Service

Identifier Management Service
  “Adenosine receptor 2a”

P12374 EC2.43.4 CS4532

Apps

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

Semantic Workflow Engine

Data Cache (Virtuoso Triple Store)

Domain Specific Services

Indexing

Public Ontologies

User Annotations

Public Content

Commercial

VoID

Db

RDF

Nanopub

Apps
• 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

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.
Solution: Emerging Public Solutions
EMBL-EBI RDF Platform

The EBI RDF Platform aims to bring together the efforts of a number of EBI services that provide access to their data using Semantic Web technologies. It provides query across resources using the Web-Scale Query Language. We welcome questions via our feedback form.

Current RDF resources

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
<th>Quick links</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioModels</td>
<td>Service description</td>
<td>SPARQL endpoint</td>
<td>All model annotations</td>
</tr>
<tr>
<td>BioSamples</td>
<td>Service description</td>
<td>SPARQL endpoint</td>
<td>Samples downloaded</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>Service description</td>
<td>SPARQL endpoint</td>
<td>Compound Search</td>
</tr>
<tr>
<td>Expression Atlas</td>
<td>Service description</td>
<td>SPARQL endpoint</td>
<td>Expression data</td>
</tr>
</tbody>
</table>

ChEMBL SPARQL Endpoint

Enter SPARQL Query

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)

http://www.ebi.ac.uk/rdf/
ChEMBL 18 RDF Model
National, International Health Systems
Data.gov, Data.gov.uk, WHO, datahub.io

https://catalog.data.gov/organization/hhs-
gov
http://data.gov.uk/data/search?res_format=RDF&q=health
http://gho.aksw.org/
http://datahub.io/dataset?q=health
Best Practices
Node and Edge Informatics

Interfaces within the Drug Development Process

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

Lead Discovery
- RNAi
- Assay Development
- HTS
- Exploratory PK

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

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

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

Registration
- NDA/BLA
- MAA
- PMR
- REMS
- PSUR

Marketing & Sales
- Observational Research

Seamless information connectivity (an EDGE) needed across domain NODEs
It’s the Data…
…not the app, not the container
Disposable Applications
Questions, Answers, Insights Persist
Cooperation…
…without Coordination

Cooperation without Coordination:
Managing Distributed Clinical Trial 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