The Translational Medicine Ontology

A Small Compass for Navigating a Large Sea of Biomedical Data

Elgar Pichler

W3C HCLSIG TMO Team Members

CSHALS 2010, Cambridge, MA

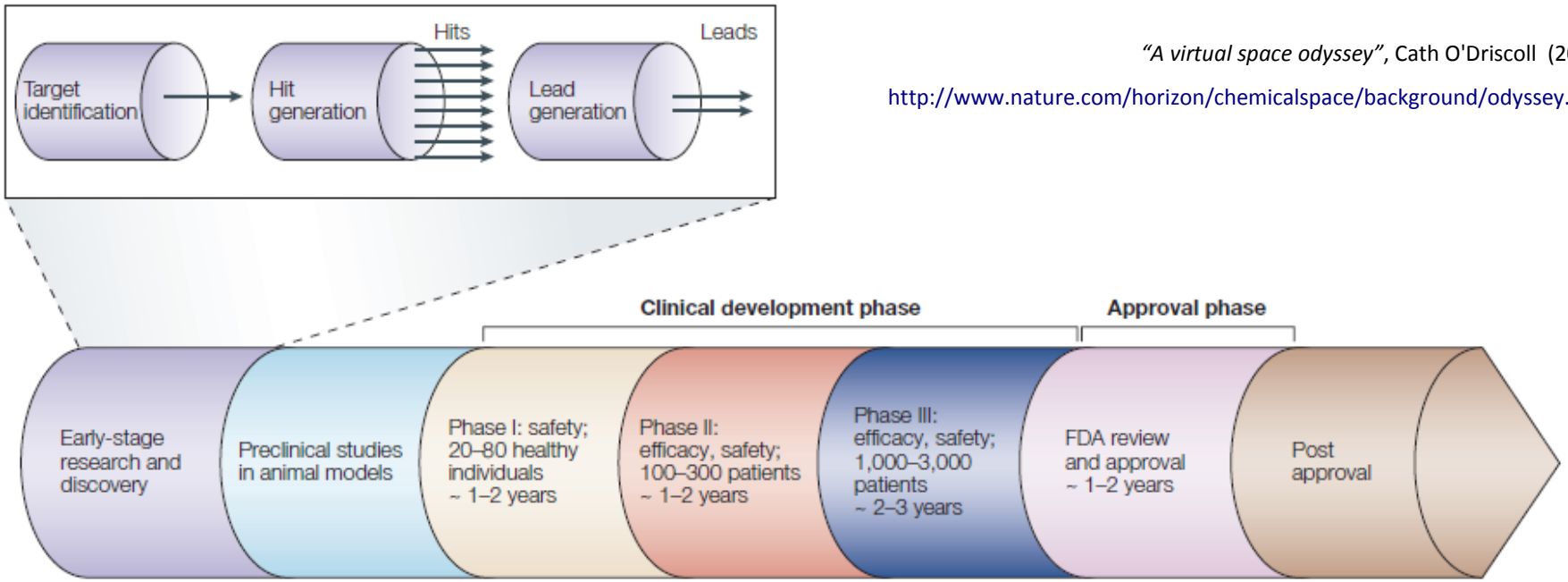
February 25, 2010

Outline

- Questions & Problems
- Translational Medicine Ontology (TMO)
  - Ontology
  - Data
  - Examples
- General Comments

Questions & Problems

The Drug Development Pipeline



"A virtual space odyssey", Cath O'Driscoll (2004)

http://www.nature.com/horizon/chemicalspace/background/odyssey.html

- The road is long, and costly.
- How do we contain costs and develop better drugs?

Questions & Problems

Aspirin – nothing new, right?

THE WALL STREET JOURNAL.
WSJ.com

HEART BEAT | FEBRUARY 23, 2010

The Danger of Daily Aspirin

By ANNA WILDE MATHEWS

If you're taking a daily aspirin for your heart, you may want to reconsider.

For years, many middle-aged people have taken the drug in hopes of reducing the chance of a heart attack or stroke. Americans bought more than 44 million packages of low-dose aspirin marketed for heart protection in the year ended September, up about 12% from 2005, according to research firm IMS Health.

Now, medical experts say some people who are taking aspirin on a regular basis should think about stopping. Public-health officials are scaling back official recommendations for the painkiller to target a narrower group of patients who are at risk of a heart attack or stroke. The concern is that aspirin's side effects, which can include bleeding ulcers, might outweigh the potential benefits when taken by many healthy or older people.

"Not everybody needs to take aspirin," says Sidney Smith, a professor at the University of North Carolina who is chairing a new National Institutes of Health effort to compile treatment recommendations on cardiovascular-disease prevention. Physicians are beginning to tailor aspirin recommendations to "groups where the benefits are especially well established," he says.

New recommendations for cardiovascular disease prevention with Aspirin:

- slightly lower daily dose than baby aspirin
- yes for person with risk factors but no history of bleeding and ulcers; for men >45y, women >55y
- no for men <45y, women <55y, or >80y

- New findings every day.
- How does this affect the use of a drug? How does it affect me?

TMO
Mission

Focuses on the development of a **high level patient-centric ontology for the pharmaceutical industry**. The ontology should enable silos in **discovery research**, **hypothesis management**, **experimental studies**, **compounds**, **formulation**, **drug development**, **market size**, **competitive data**, **population data**, etc. to be brought together. This would enable scientists to answer new questions, and to answer existing scientific questions more quickly. This will help pharmaceutical companies to model patient-centric information, which is essential for the tailoring of drugs, and for early detection of compounds that may have sub-optimal safety profiles. The ontology should **link to existing publicly available domain ontologies**.

TMO Development
Concept Identification via Use Cases

Process:

- describe roles, work out use cases
- identify used concepts
- map concepts to other ontologies/vocabularies
- align with Basic Formal Ontology (BFO)
- identification of candidate domain ontologies
- refine and start over again

[bottom-up approach; compare tutorial by John Madden]

Roles

Role	Primary Interests
Cellular and molecular biologists	Assessing target viability
Cheminformatician	Analyzing chemical data and making predictions
Clinical decision support	Analyzing response to therapies
Clinical trial formulator	Designing clinical trials
Health plan provider	Providing insurance coverage to individuals
Immunologist	Developing large molecules for therapeutic purposes
In vitro biologist	Predicting success of compounds to be tested in vivo
In vivo biologist	Performing toxicology and efficacy studies in animals
Medicinal chemist	Exploring structural patterns and properties of compounds
Primary care clinician	Treating broad range of patients
Project manager	Prioritizing activities and resources
Sales and marketing	Driving sales
Specialty medical provider	Treating patients with specific diseases
Statistician	Testing scientific hypotheses using statistical approaches
Strategic/portfolio manager	Assessing market opportunities
Systems physiologist	Understanding the biological system

TMO Development

Concept Identification via Use Cases

Example

(see http://esw.w3.org/topic/HCLSIG/PharmaOntology/UseCases):

1. Patient [OBI:0000093, patient role] (and family members [NCI:Patient_Family_Member_or_Friend]) report symptoms [IDO:0000048, Symptom] to physician/clinician [NCIt:Physician]. Physician/clinician enters reported symptoms into eHR.

2. Physician [NCIt: Physician] makes a list of differential diagnoses, with a working diagnosis [OBI:0000075] of Alzheimer Disease [DOID:10652]. (Data Source: Physician's head).

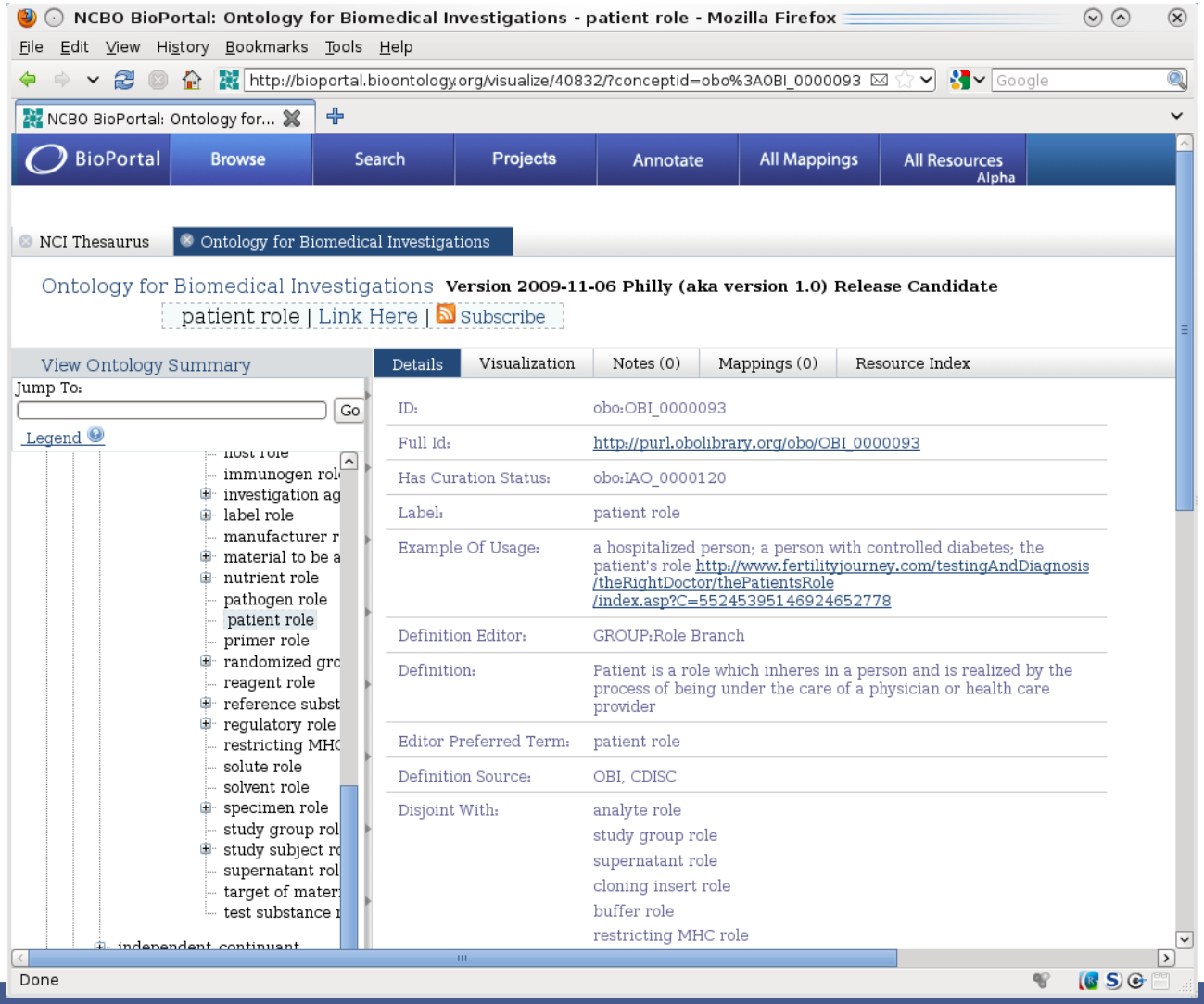
3. Physician [NCIt:Physician] arranges for patient [OBI:0000093, patient role] to have a basic biochemical/haematological, and SNP [SO:0000694, SNP] profile undertaken. Biochemistry, Haematology, and SNP requests are input by respective departments directly into patient's eHR [HL7:EHR, UMLS:C1555708, HID:20081] from laboratory (Data Source: eRecord). Preliminary SNP and genetic data will be submitted directly to the NIH Pharmacogenetics Research Network (PGRN).

[...]

TMO Development

Mapping to Other Ontologies/Vocabularies

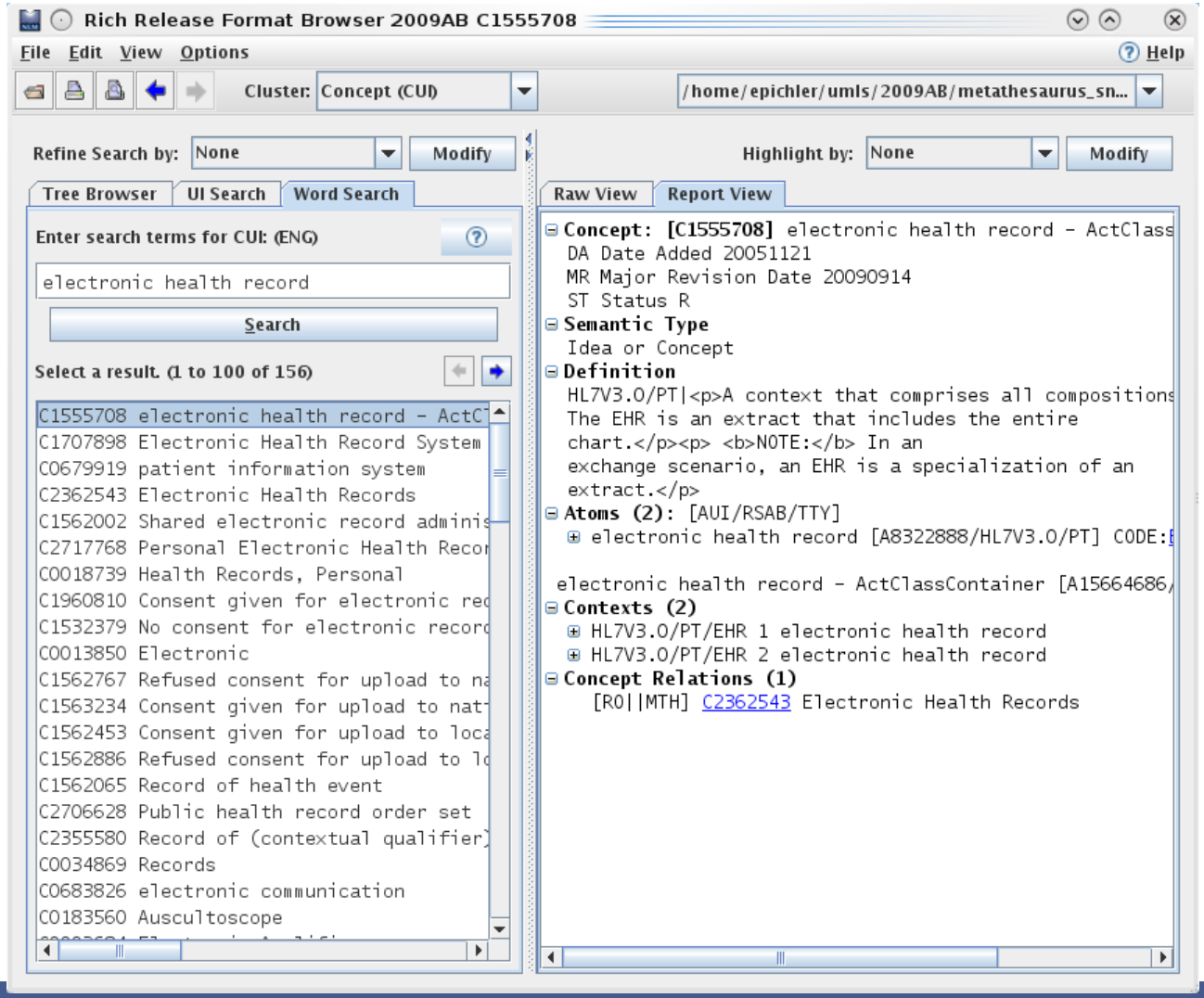
NCBO



TMO Development

Mapping to Other Ontologies/Vocabularies

UMLS



TMO Development

Mapping to Other Ontologies/Vocabularies

Mapping examples:

TMO class	Classes in other ontologies
pharmaceutical product (TMO_0002)	NCIt:Finished_Pharmaceutical_Product, UMLS:C1708062
target (TMO_0006)	NCIt:Target, OCRe:research2:target, UMLS:C1521840
institution (TMO_0025)	ACGT:Institution, BIRNLex:2085, LNC:LP76237-4, NCIt:Institution, SNOMEDCT:385437003, UMLS:C1272753
intervention (TMO_0030)	ClinicalTrialOntology:prtont:PeriodType_5, NCIt:Intervention, OCRe:research2:Intervention
clinical trial (TMO_0032)	HL7V3.0:CLNTRL, MSH:D016430, NCIt:Clinical_Trial, SNOMEDCT:110465008
disease (TMO_0047)	ACGT:Disease, BIRNLex:11013, DOID:4, GRO:Disease, LNC:LP21006-9, MSH:D004194, NCIt:Disease_or_Disorder, NDFRT:C2140, OBI:0000155

TMO Development
Use of Other Ontologies/Vocabularies

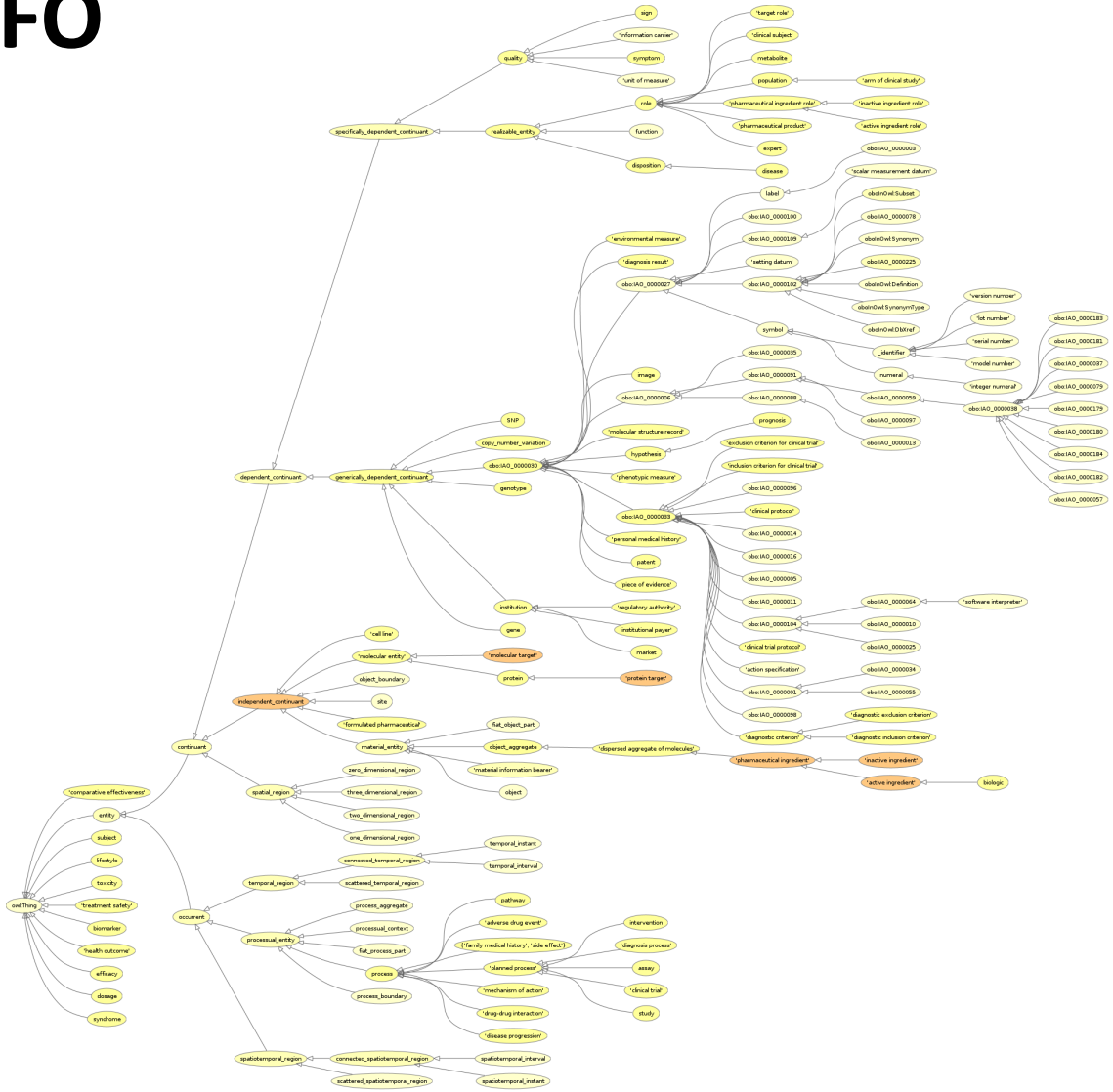
Ontologies used in TMO:

- Experimental Factor Ontology (EFO): cell line
- Information Artifact Ontology (IAO): class annotations
- Ontology for Biomedical Investigations (OBI): planned process, label molecular entity, metabolite
- Protein Ontology (PRO): protein
- Sequence Ontology (SO): SNP, gene, copy number variation, genotype

cell line

TMO Development
Alignment with BFO

<100 main TMO classes
aligned with BFO



TMO
Data Aggregation

Process:

- rdf-ize data
- load data into Virtuoso triple store
- generate mappings (sameAs links) between data sources and TMO via
  - same IDs
  - string & semantic matching (LinQuer, SILK)

TMO

Data Sources

Name	Topic	Short Description	Size	LODD	TMO
DailyMed	Drugs	dailymed.nlm.nih.gov provides information about approved prescription drugs, includes FDA approved labels (package inserts).	164,276 triples; 4,039 drugs	x	x
DBpedia	Drugs / Diseases / Proteins	RDF data about 2.49 million things that has been extracted from Wikipedia.	218M triples; 2,300 drugs; 2,200 proteins	x	
Diagnostic Data	Disease / Diagnosis	AD specific diagnostic data extracted from a paper by DuBois et al (2007).			x
Diseasome	Diseases / Genes	Diseasome describes characteristics of disorders and disease genes linked by known disorder–gene associations.	91,182 triples; 2,600 genes	x	x
DrugBank	Drugs	Drugbank.ca provides drug (i.e., chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e., sequence, structure, and pathway) information.	766,920 triples; 4,800 drugs; 2,500 protein sequences	x	x
LinkedCT	Clinical Trials	Linked data source of trials from ClinicalTrials.gov	7M triples; 62000 trials	x	x
Medicare	Medicare Formulary	List of drugs that recipients of Medicare D are eligible to receive.		x	x
Patient Records	Patient Data	Hand-generated test patient data, assuming data was collected within a PCHR (personally controlled health record).			x
PharmGKB	Genetic Information / Drug Response	Contains information that relates genetic variation to variation in drug response.			x
RDF-TCM	Genes / Diseases / Medicines / Ingredients	Traditional Chinese medicine, gene and disease association dataset and a linkset mapping TCM gene symbols to Extrez Gene IDs created by Neurocommons.	117,643 triples	x	
SIDER	Diseases / Side Effects	SIDER contains information on marketed drugs and their adverse effects.	192,515 triples; 1,737 genes	x	x
STITCH	Chemicals / Proteins	STITCH contains information on chemicals, proteins, and their interactions.	7,500,000 chemicals; 500,000 proteins; 370 organisms	x	

TMO Data

Mapping to TMO

Data Source	Mapping (Data Source to TMO)
ClinicalTrials.gov	- '??' maps to 'textual entity' (IAO_0000300) - 'diagnostic criteria' maps to 'diagnostic criterion' (TMO_0068) - 'diagnostic criteria' maps to 'diagnostic inclusion criterion' (TMO_0069) - 'diagnostic criteria' maps to and 'diagnostic exclusion criterion' (TMO_0070)
Diagnostic Data (DuBois)	- 'drugs' map to 'pharmaceutical product' (TMO_0002) - 'ingredients' map to 'active ingredient' - 'organization' maps to 'institution' (TMO_0025)
DailyMed	- 'disease' maps to 'disease' (TMO_0047) - 'gene' maps to 'gene' (SO:0000704)
Diseasome	- 'drug-drug interactions' maps to 'drug-drug interaction' (TMO_0040) - 'drugs' maps to 'pharmaceutical product' (TMO_0002) - 'targets' map to 'target' (TMO_0006)
DrugBank	- 'drugs' map to 'pharmaceutical product' (TMO_0002)
Patient Records	- 'association' maps to 'study result' (OBI_0000682)
PharmGKB	- 'drugs' map to 'active pharmaceutical ingredient' (TMO_0000) - 'side effects' map to 'adverse drug event' (TMO_0043)
SIDER	- '??' maps to 'textual entity' (IAO_0000300) - 'diagnostic criteria' maps to 'diagnostic criterion' (TMO_0068) - 'diagnostic criteria' maps to 'diagnostic inclusion criterion' (TMO_0069) - 'diagnostic criteria' maps to and 'diagnostic exclusion criterion' (TMO_0070)

TMO
Sample Queries … and Answers

Discovery:

- What genes are associated with or implicated in AD?

  At least 97 genes have some association with AD.

- Which existing marketed drugs might potentially be re-purposed for AD because they are known to modulate genes that are implicated in the disease?

  57 compounds or classes of compounds that are used to treat 45 diseases.

Physician

- What are the diagnostic criteria for AD?

  12 Diagnostic inclusion criteria and 9 exclusion criteria were obtained from the criteria outlined in Dubois et al.

- Is Donepezil covered by Medicare D?

  Yes, Medicare D covers two brand name formulations of Donepezil.

Clinical:

- What active trials are ongoing that would be a good fit for Patient 2?

  58 Alzheimer trials, 2 mild cognitive impairment trials, 1 hypercholesterolaemia trial, 66 myocardial infarction trials, 46 anxiety trials, and 126 depression trials.

TMO

Sample Query

Which existing marketed drugs might potentially be re-purposed for AD because they are known to modulate genes that are implicated in the disease?

drug_name	disease2_name
(s)-rolipram	Schizophrenia
(s)-rolipram	Autistic Disorder
(s)-rolipram	Bipolar Disorder
(s)-rolipram	Depression
⋮	⋮
irbesartan	Hypertension
lisinopril	Hypertension
lisinopril	Diabetes Mellitus, Insulin-Dependent
nifedipine	Hypertension
perindopril	Proteinuria
perindopril	Diabetes Mellitus, Non-Insulin-Dependent
perindopril	Cerebrovascular Accident
perindopril	Cardiovascular Diseases
perindopril	Dementia
perindopril	Hypertension
perindopril	Memory Disorders
pravastatin	Coronary Arteriosclerosis

TMO

- home:
  - http://esw.w3.org/topic/HCLSIG/PharmaOntology
- source code / TMO:
  - http://www.w3.org/2001/sw/hcls/ns/transmed
  - http://code.google.com/p/translationalmedicineontology/
- data sources (text search & SPARQL endpoint):
  - http://tm.semanticscience.org/fct
  - http://tm.semanticscience.org/sparql
- example queries:
  - http://esw.w3.org/topic/HCLSIG/PharmaOntology/Queries

Comments

Personal Wishlist – Manifesto-Worthy?

- Data providers should make their data available in RDF, with SPARQL endpoints.
- SuperMapper should do all of our mapping work, specify which kind of mapping is used, and record relevant provenance data.
- Federated query should be enabled with access policy mediation.

[See also comments on sameAs and provenance in talk by James McCusker & Deborah Mc Guinness.]

[Talk to Eric Prud'hommeaux about concept for access policy mediation for SPARQL endpoints.]

Summary

Plus

- Several pharma/drug/translational medicine relevant data are available as linked data set.
- A first TMO candidate has been developed.
- The TMO project is a great example of a collaboration between industry, academia, and W3C HCLS in the pre-competitive space.

Minus

- More intuitive and tailored interfaces to linked data are needed.
- There is a lack of freely available clinical data.

Future TMO work:

- Tighter ontology/data integration.
- Revisit mapping procedures.
- Flexible integration of candidate domain ontologies/vocabularies.
- Interfaces.

[Data & know-how sharing, compare also talk by Vijay Bulusu]

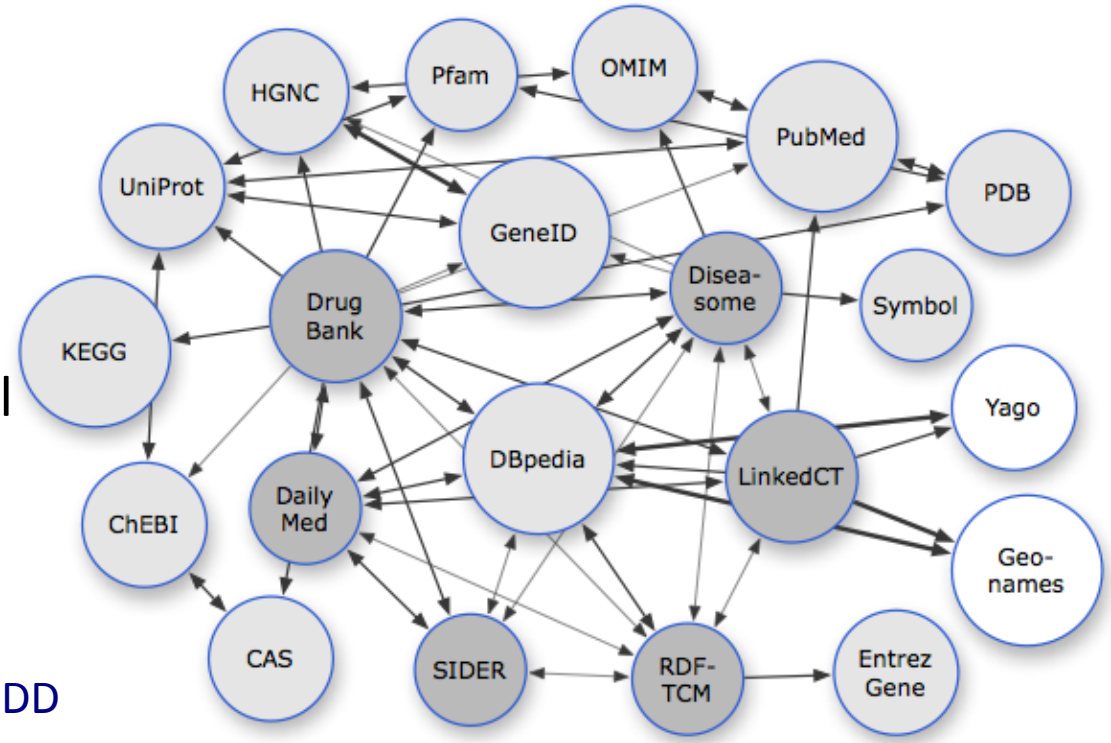
[W3C HCLSIG, see talks from Tutorial session and by Susie Stephens]

LODD
Mission & Linked Data Cloud

LODD …

… focuses on linking various sources of drug data – ranging from data describing the impact of drugs on gene expression, through to clinical trial results – to answer interesting scientific and business questions.

http://esw.w3.org/topic/HCLSIG/LODD



LODD data in the Linked Data cloud …

… are represent in dark gray Collectively, the data sets consist (August 2009) of over 8 million RDF triples, which are interlinked by more than 370,000 RDF links.

LODD

- home:
  - http://esw.w3.org/topic/HCLSIG/LODD
- data sources (with SPARQL endpoints list):
  - http://esw.w3.org/topic/HCLSIG/LODD/Data
  - http://hcls.deri.org/sparql
- examples
  - http://www4.wiwiss.fu-berlin.de/lodd/topquestions/

Acknowledgements

- TMO
  - Colin Batchelor, Christine Denney, Christopher Domarew, Michel Dumontier, Anja Jentzsch, Joanne Luciano, Susie Stephens, Patricia L. Whetzel
  - Bosse Andersson, Olivier Bodenreider, Tim Clark, Lee Harland, Vipul Kashyap, Peter Kos, Julia Kozlovsky, James McGurk, Chimezie Ogbuji, Eric Prud'hommeaux, Matthias Samwald, Lynn Schriml, Jun Zhao
- LODD
  - Bosse Anderssen, TN Bhat, Chris Bizer, Don Doherty, Michel Dumontier, Anja Jentzsch, Oktie Hassanzadeh, Scott Marshall, Glen Newton, Eric Prud'hommeaux, Matthias Samwald, Susie Stephens, Kristin Tolle, Egon Willighagen, Jun Zhao
  - Eli Lilly
- W3C / Semantic Web for Health Care and Life Sciences Interest Group