How data sharing leads to knowledge

M. Scott Marshall, Ph.D.
W3C HCLS IG co-chair
Leiden University Medical Center
University of Amsterdam
http://staff.science.uva.nl/~marshall
http://www.w3.org/blog/hcls

Motivation

→Science is based on **knowledge***:*

knowledge capture, knowledge sharing, i.e. communication of findings.

→Semantic Web provides a basis for knowledge sharing through machine-readable and *reason-able* annotation of resources.

What is *knowledge* ?

"data", "information", "facts", "knowledge"

Knowledge is a statement
that can be tested for truth.

(by a machine)

Otherwise, computing can't add much

RDF : a web format for knowledge

RDF is a W3C language to express statements.

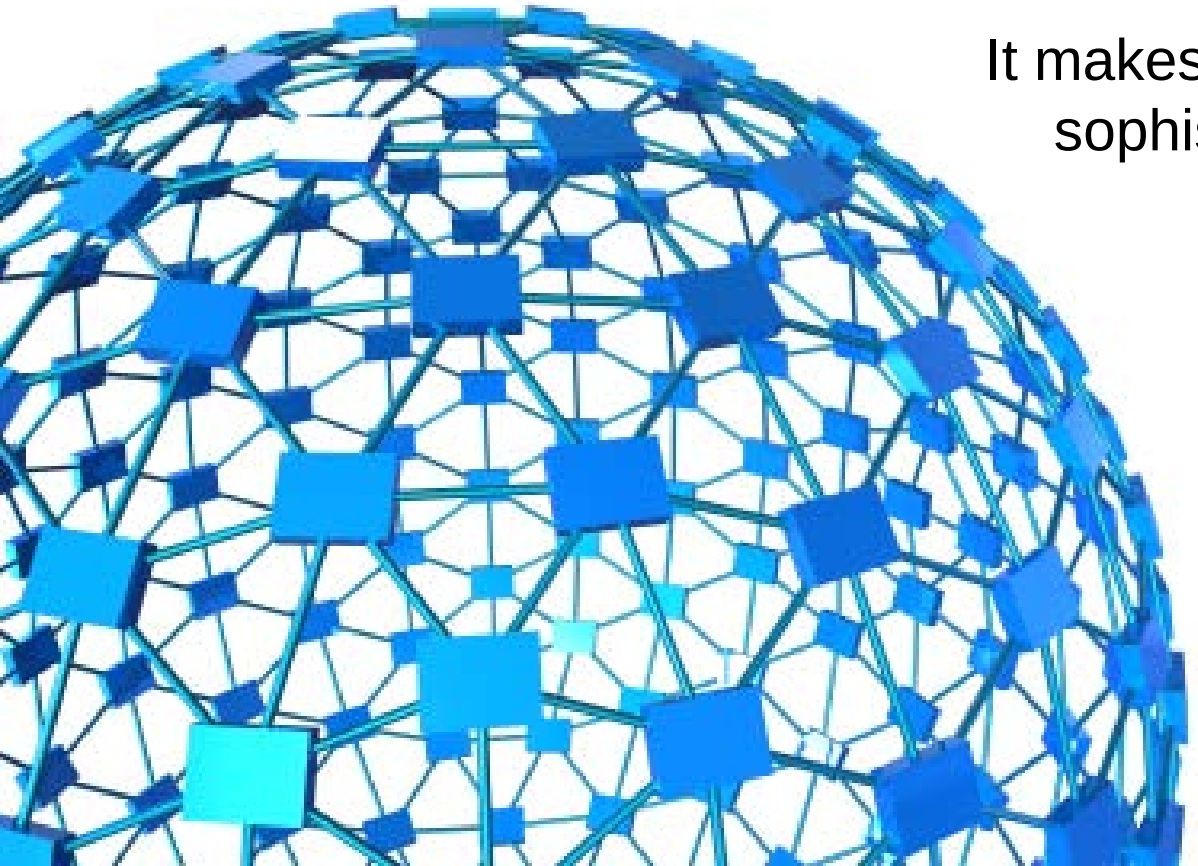
RDF Triple:

Graph of Knowledge:

The Semantic Web is the New Global Web of Knowledge

It is about standards for publishing, sharing and querying knowledge drawn from diverse sources

It makes possible the answering sophisticated questions using background knowledge



W3C Semantic Web

Source: Michel Dumontier

Where is biomedical knowledge?

Can be extracted from:

- People
- Literature
- Diagrams
- Clinical reports
- Databases
- Excel sheets
- …

Most of these sources of biomedical knowledge are ***not*** machine-readable

Many tasks are still a challenge!

With existing Web and Health IT:

- Find and integrate information
  - "Although a plethora of resources (tools, databases, materials) for neuroscientists is now available on the web, finding these resources among the **billions of possible** web pages continues to be a challenge." [M. Martone, NCBO Seminar Series, 4 Nov 2009]
- Make multiple inferences based on background knowledge
  - to obtain more complete answers
  - to discover knowledge

Source: Christine Golbreich

Examples

- in a medical record system

  “find all patients whose radiology exhibits a fracture of femur”

- in genomic data

  “find all genes annotated with a molecular function or any of its descendants and which is associated with any form of a given disease” (see genes associated with *muscular dystrophy* [Sahoo et al. 2007])

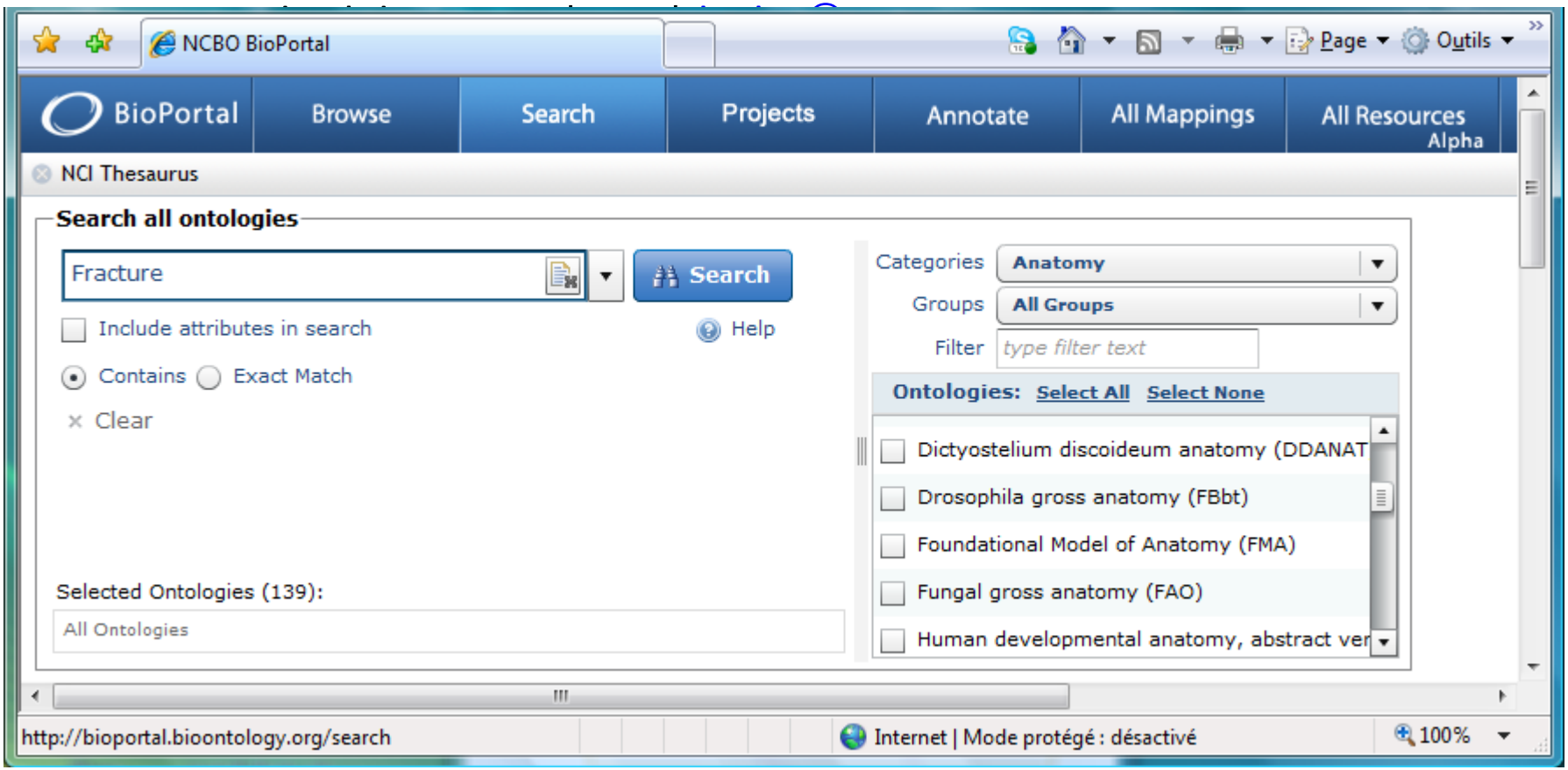
- find, share, annotate images

Source: Christine Golbreich

Pistoia Alliance
Vocabulary Services Initiative

"The life sciences industry currently operates in an environment where few of the basic components of its study (e.g. genes, proteins, cells, diseases, biomarkers, assays, drugs and technologies) are described using consistent, universally agreed-upon vocabularies."

Biological and medical ontologies



Source: Christine Golbreich

Some of the forces at work

- Pharmaceutical industry changing strategy
  - David Cox (Pfizer) Strategy: Academic / Industry partnership, wellness: rare variants that *protect* against disease
  - Pistoia Alliance, Vocabulary Services Initiative
- Personalized Medicine and EHRs
- US NIH NCBCs: NCBO and I2B2
- NCI Semantic Infrastructure
- European Innovative Medicine Initiatives (IMI)

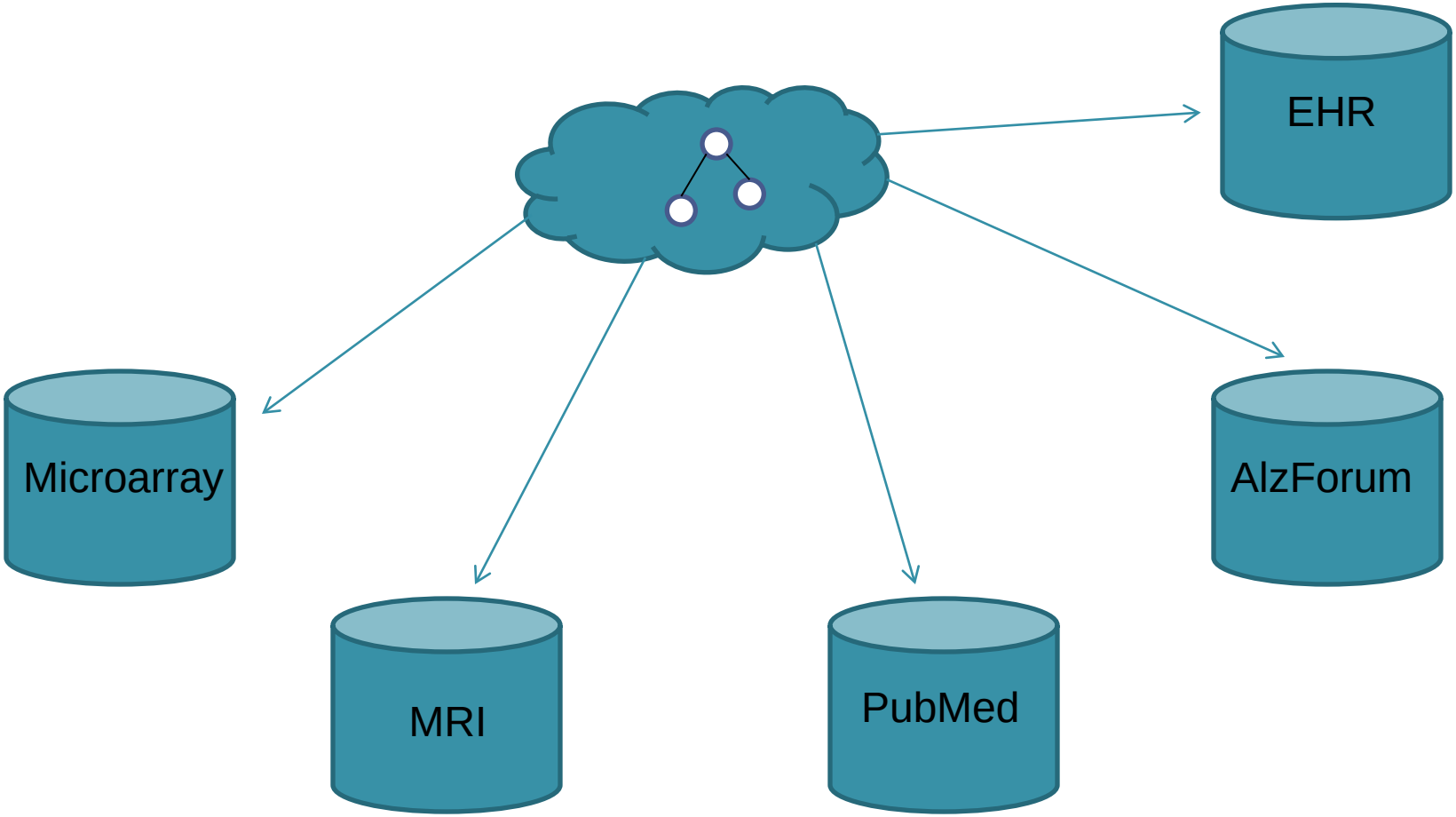
Background of the HCLS IG

- Originally chartered in 2005
  - Chairs: Eric Neumann and Tonya Hongsermeier
- Re-chartered in 2008
  - Chairs: Scott Marshall and Susie Stephens
  - Team contact: Eric Prud'hommeaux
- Broad industry participation
  - Over 100 members
  - Mailing list of over 600
- Background Information
  - http://www.w3.org/blog/hcls
  - http://esw.w3.org/topic/HCLSIG

Mission of HCLS IG

- The mission of HCLS is to develop, advocate for, and support the use of Semantic Web technologies for
  - Biological science
  - **Translational medicine**
  - Health care
- These domains stand to gain tremendous benefit by adoption of Semantic Web technologies, as they depend on the **interoperability of information** from many domains and processes for efficient decision support

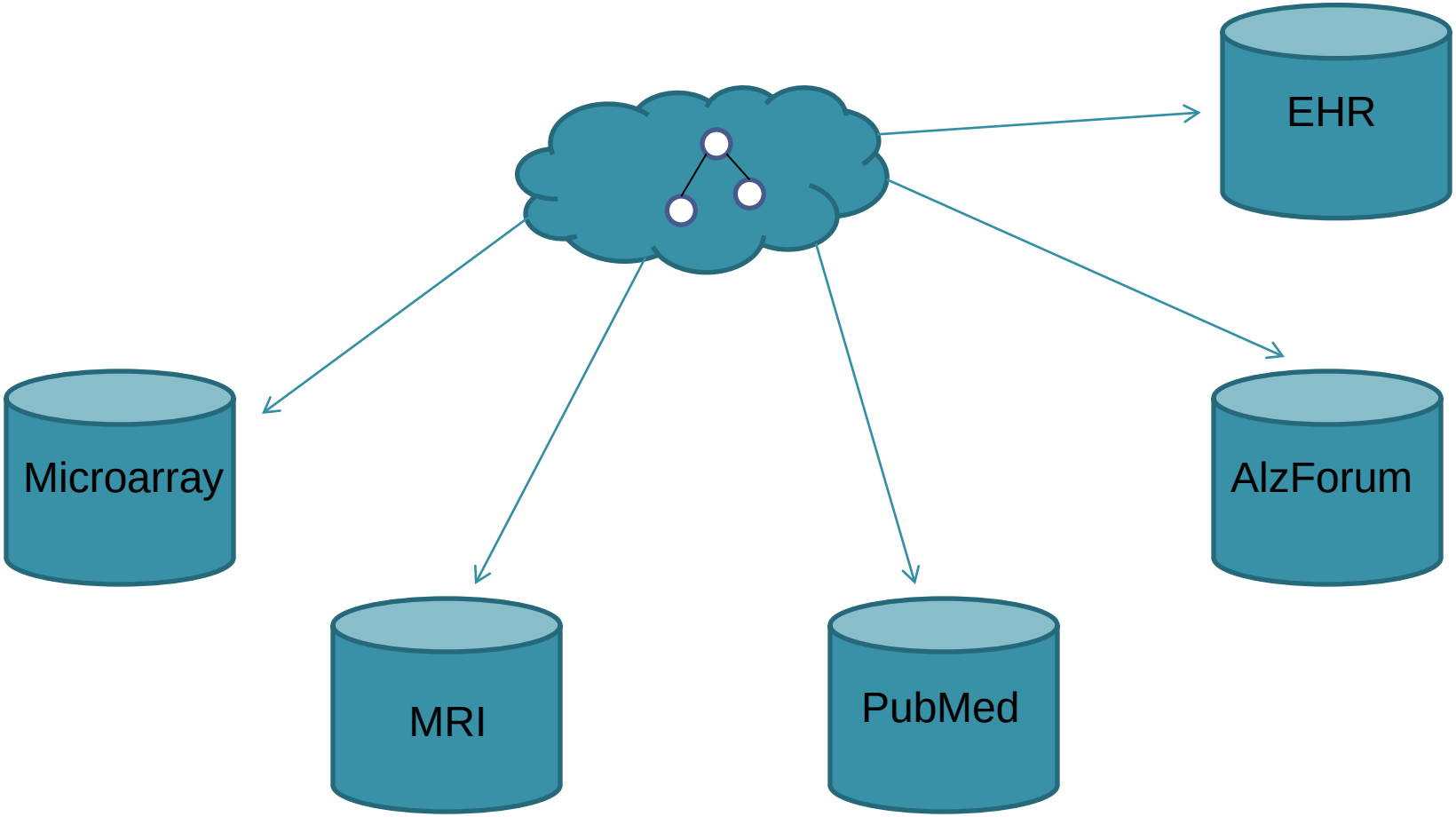
Translating across domains



Current Task Forces

- **BioRDF** – federating (neuroscience) knowledge bases
  - M. Scott Marshall (Leiden University Medical Center / University of Amsterdam)
- **Clinical Observations Interoperability** – patient recruitment in trials
  - Vipul Kashyap (Cigna Healthcare)
- **Linking Open Drug Data** – aggregation of Web-based drug data
  - Susie Stephens (Johnson & Johnson)
- **Translational Medicine Ontology** – high level patient-centric ontology
  - Michel Dumontier (Carleton University)
- **Scientific Discourse** – building communities through networking
  - Tim Clark (Harvard University)
- **Terminology** – Semantic Web representation of existing resources
  - John Madden (Duke University)

BioRDF: Translating across domains



Provenance

- Data context (can be experimental context)
- Represent knowledge so that
  - others can discover where a fact (or triple) came from
  - and evaluate how to use it

– link facts to data as evidence

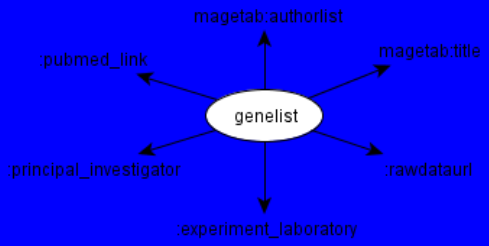
Provenance types are perspectives on the data

institutional level

Is the laboratory where the experiment was performed trusthworthy?

What other experiments were performed by the same group/lab?

Where was the experiment published and where can the raw data be accessed?

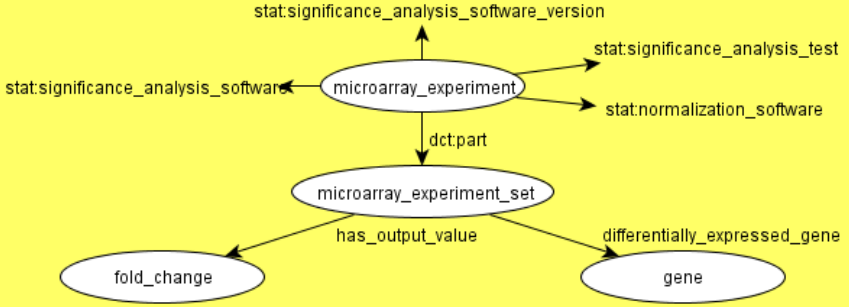


data analysis and significance level

Which algorithm was used to determine genes that are differentially expressed in Alzheimer's Disease patients?

What is the p-value associated with gene expression fold increase or decrease in the test case vs the control?

What was the statistical cutoff to determine that a gene is differentially expressed?

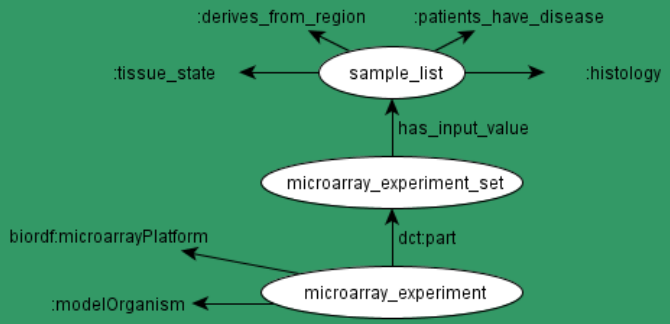


experimental context level

Find all experiments involving the Entorhinal cortex of Alzheimer's Disease patients

Are the same genes differentially expressed in the Hippocampus of Homo sapiens and Mus musculus?

What experiments were performed using neurobriillary tangles of the Entorhinal cortex layer II?

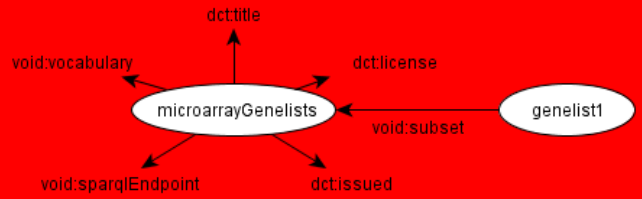


dataset description level

Perform the query using a specific version of the RDF genelist

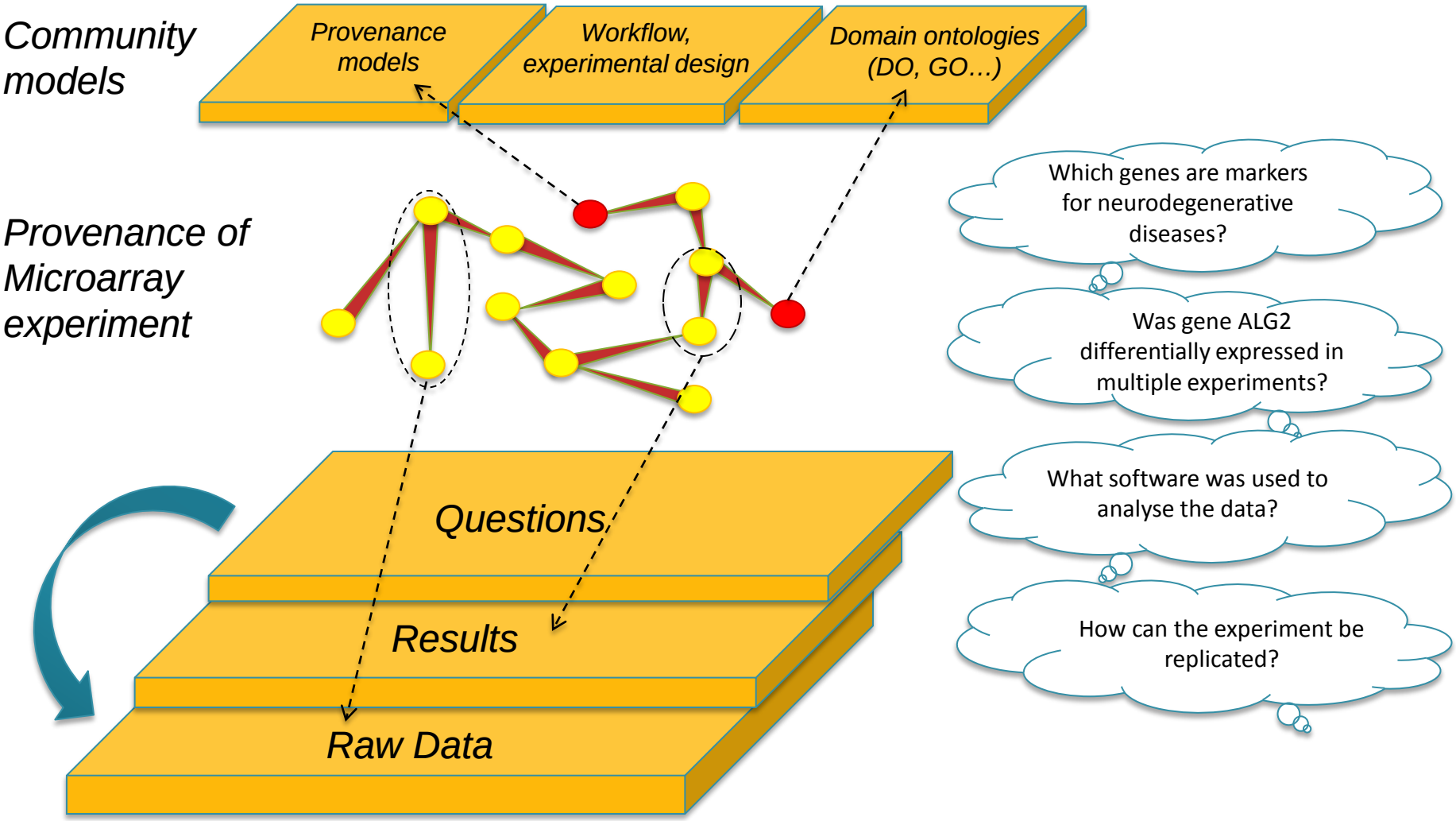
Get of summary of the datasets available for query (e.g. gene count, study type)

Find more data about differentially expressed genes in other datasets by federating the query



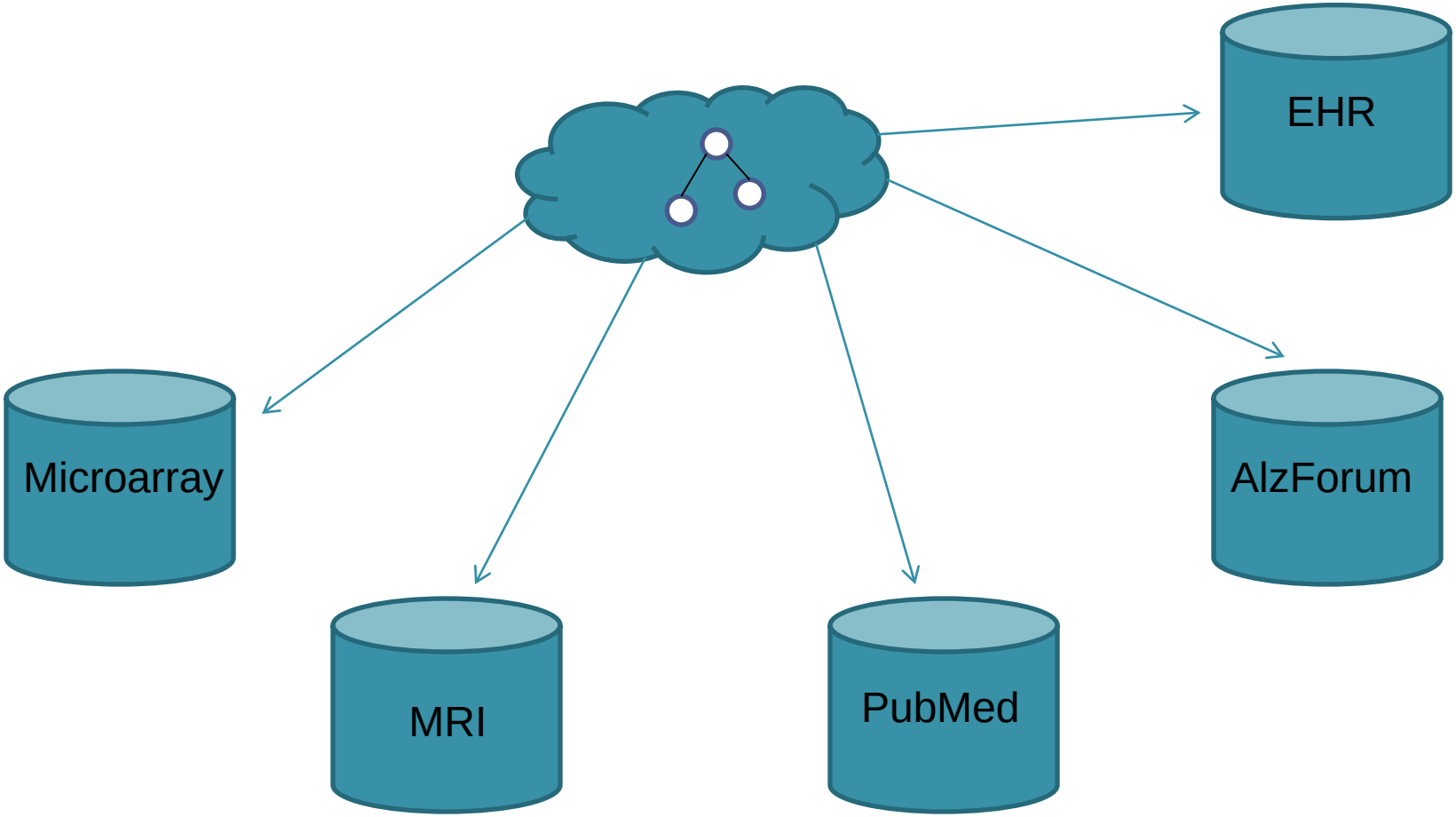
Source: Helena Deus

A Bottom-up Approach

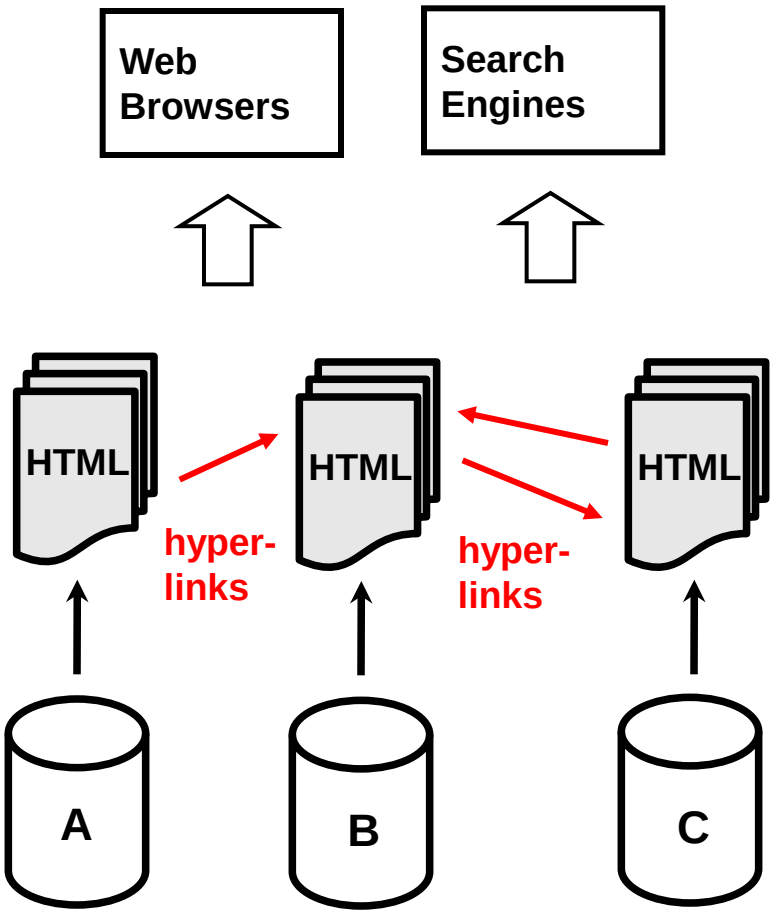


Source: Helena Deus

LODD: Translating across domains



The Classic Web

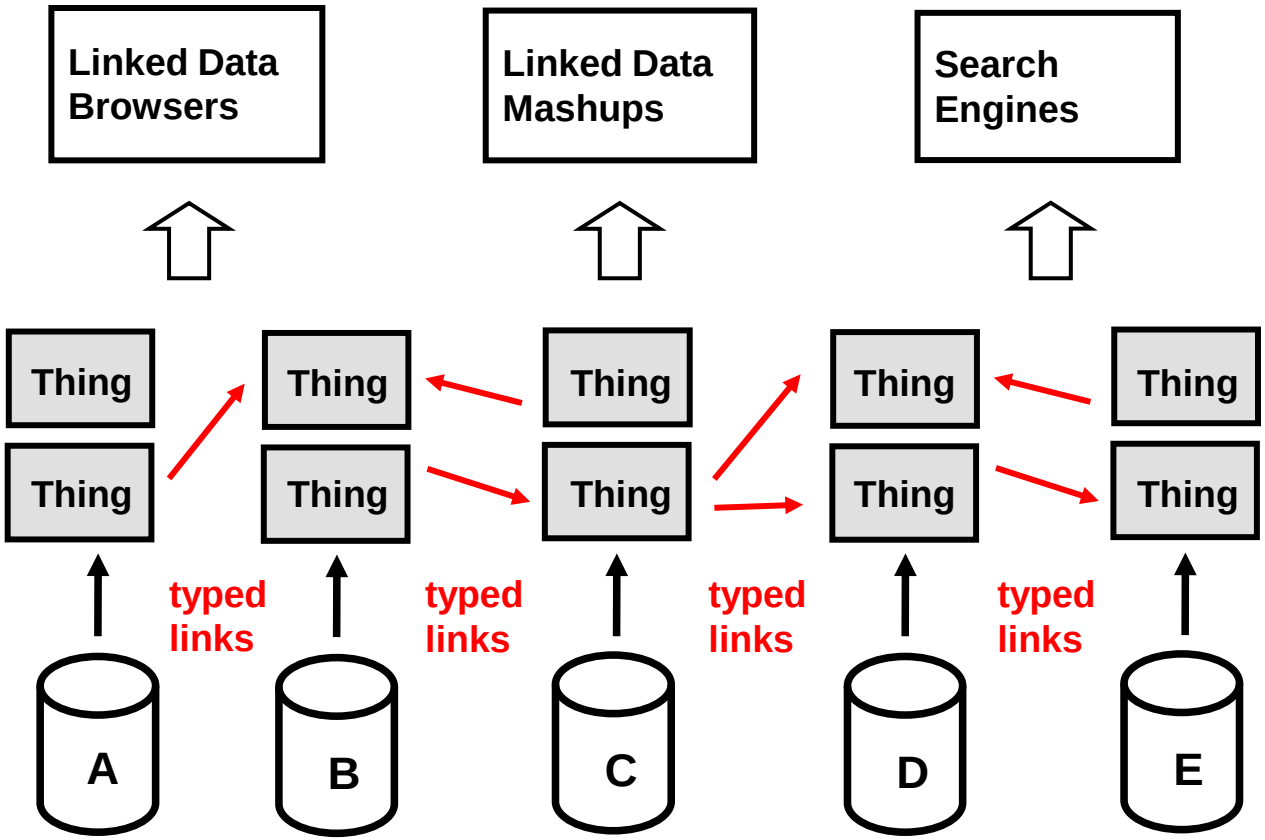


- Single information space
- HTML describes presentation
- Built on URIs
  - globally unique IDs
  - retrieval mechanism
- Built on Hyperlinks
  - are the glue that holds everything together

Source: Chris Bizer

Linked Data

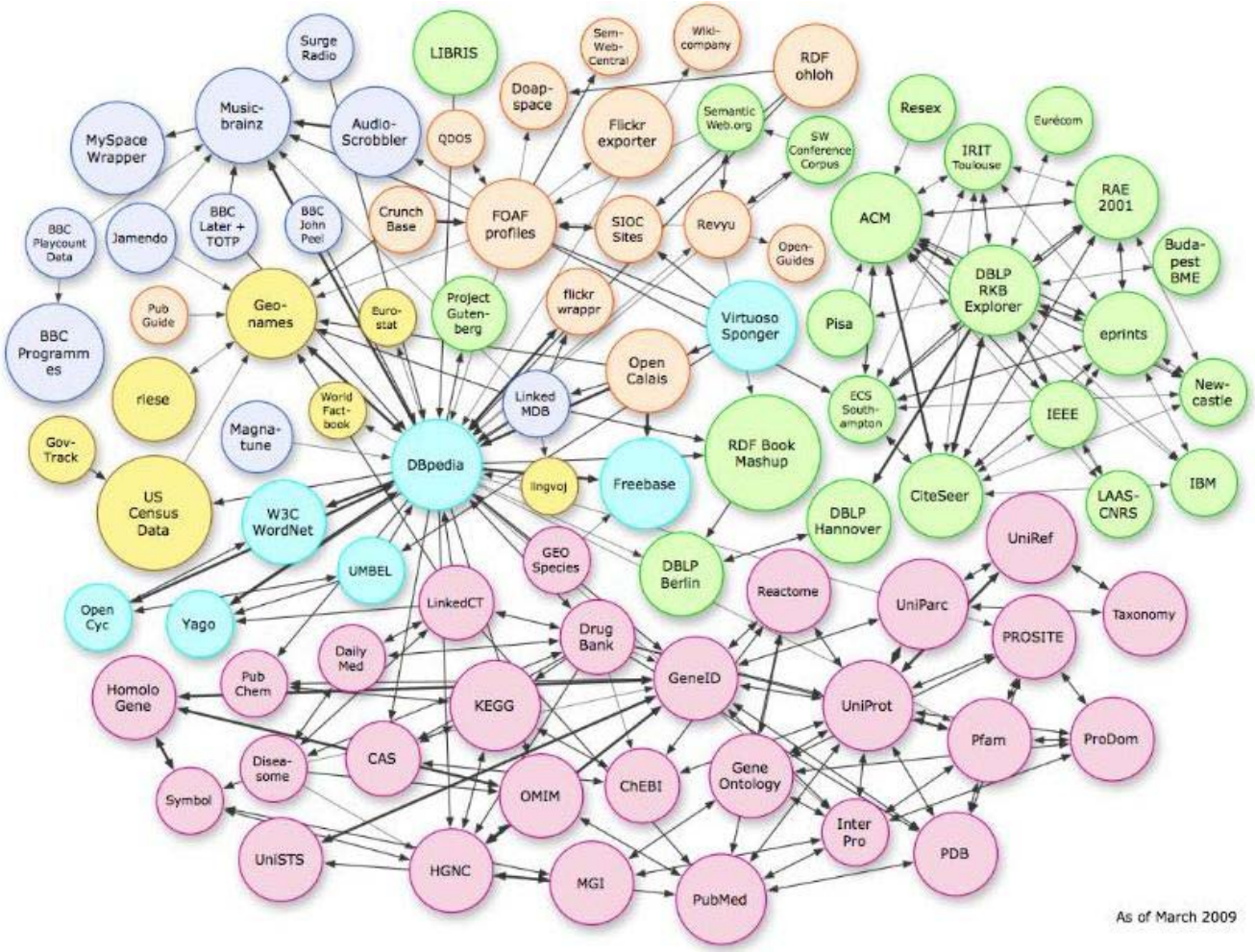
Use Semantic Web technologies to publish structured data on the Web and set links between data from one data source and data from another data sources



Source: Chris Bizer

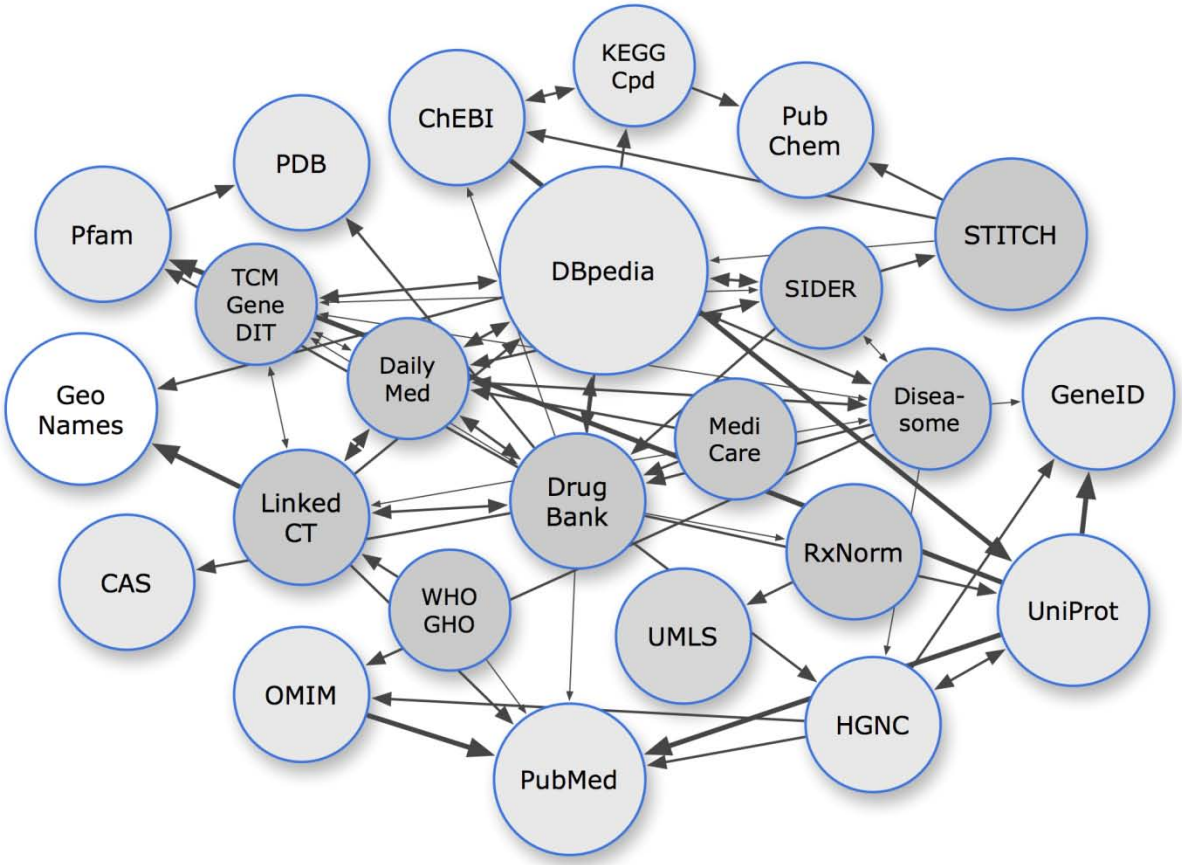
The Linked Data Cloud

LINKINGOPENDATA
W3C SWEO Community Project

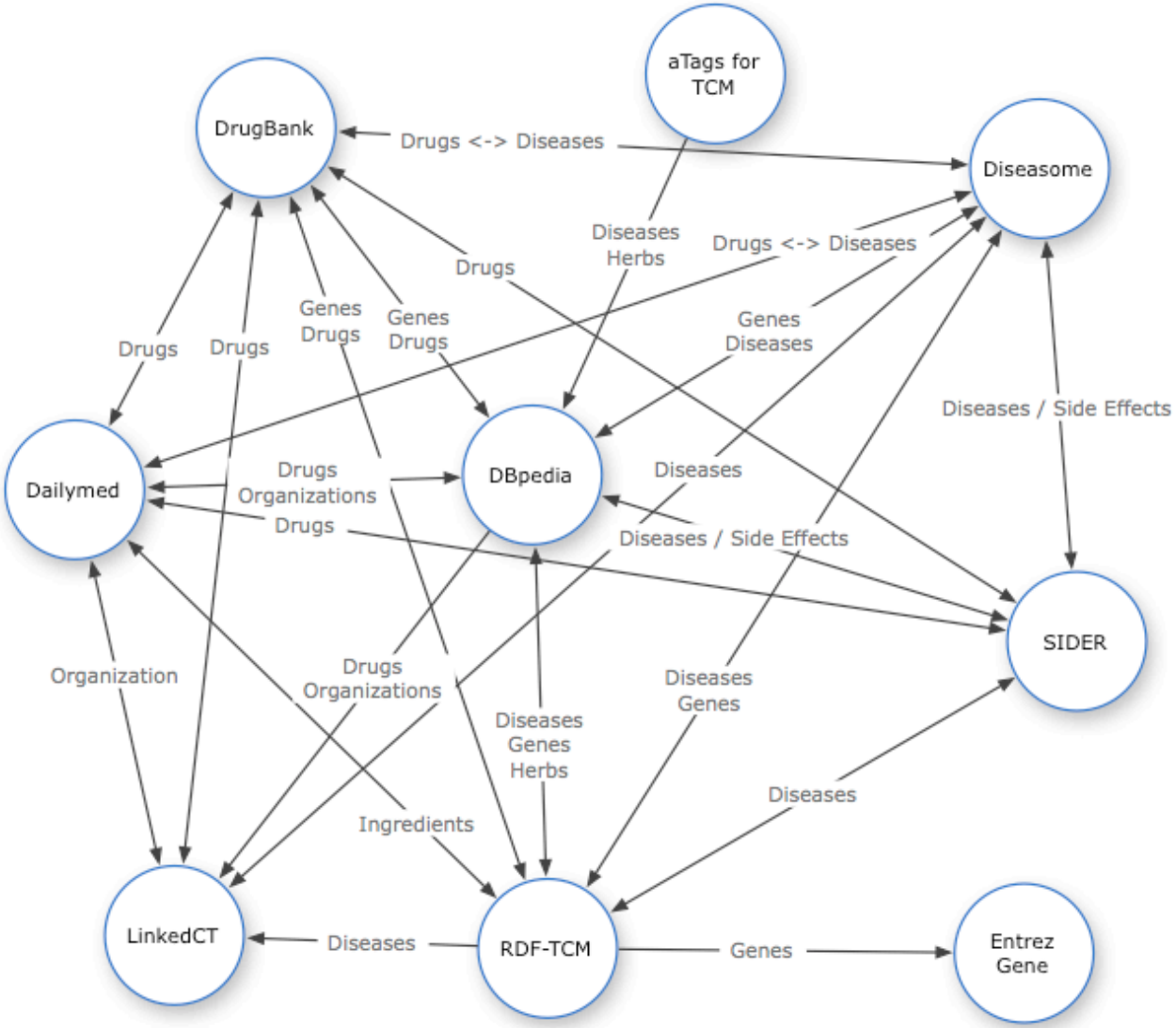


Source: Chris Bizer

LODD

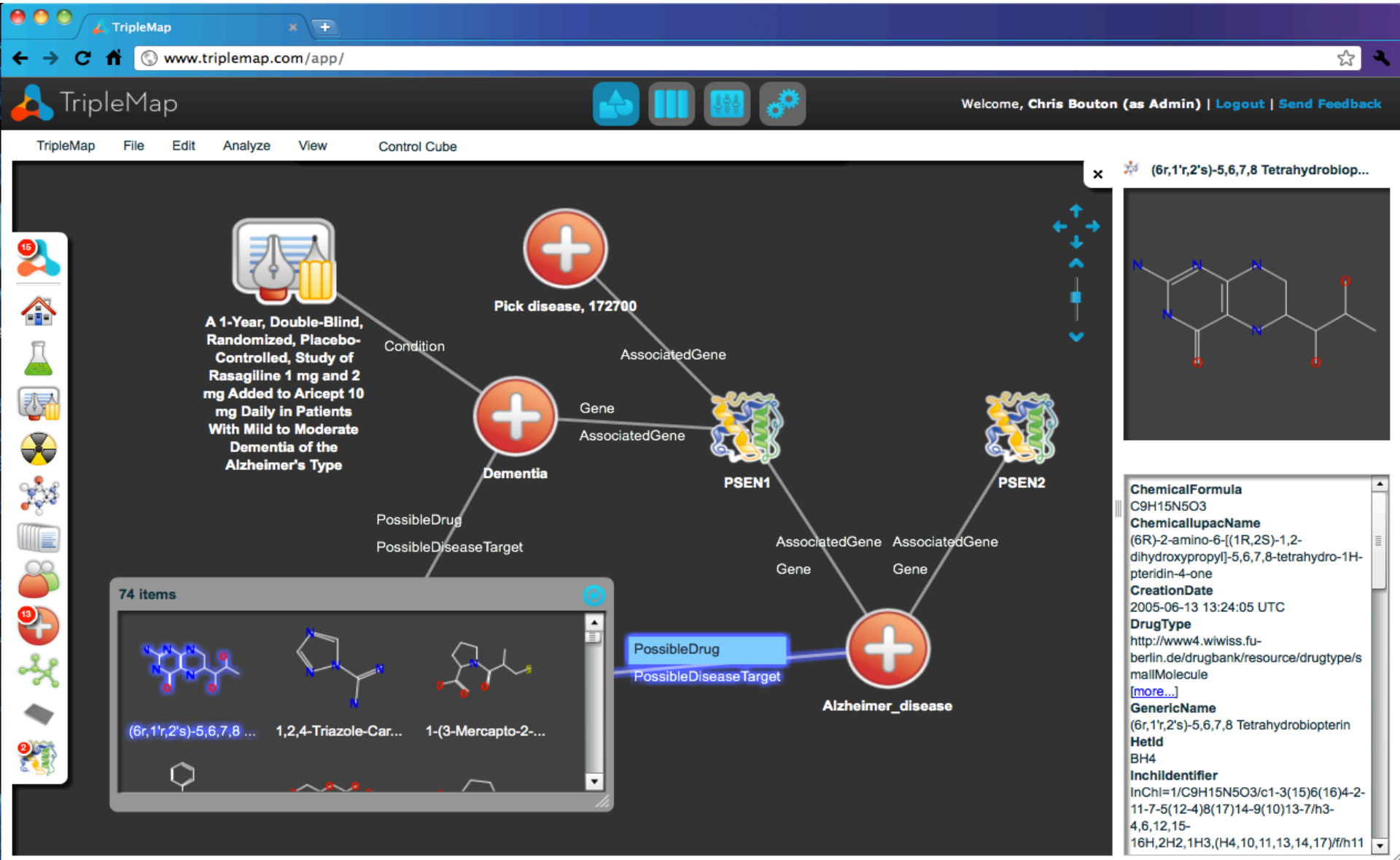

KEGG Cpd
ChEBI
Pub Chem
PDB
Pfam
STITCH
DBpedia
TCM Gene DIT
SIDER
Daily Med
Geo Names
Disea-some
GeneID
Medi Care
Drug Bank
Linked CT
RxNorm
CAS
WHO GHO
UMLS
UniProt
OMIM
HGNC
PubMed

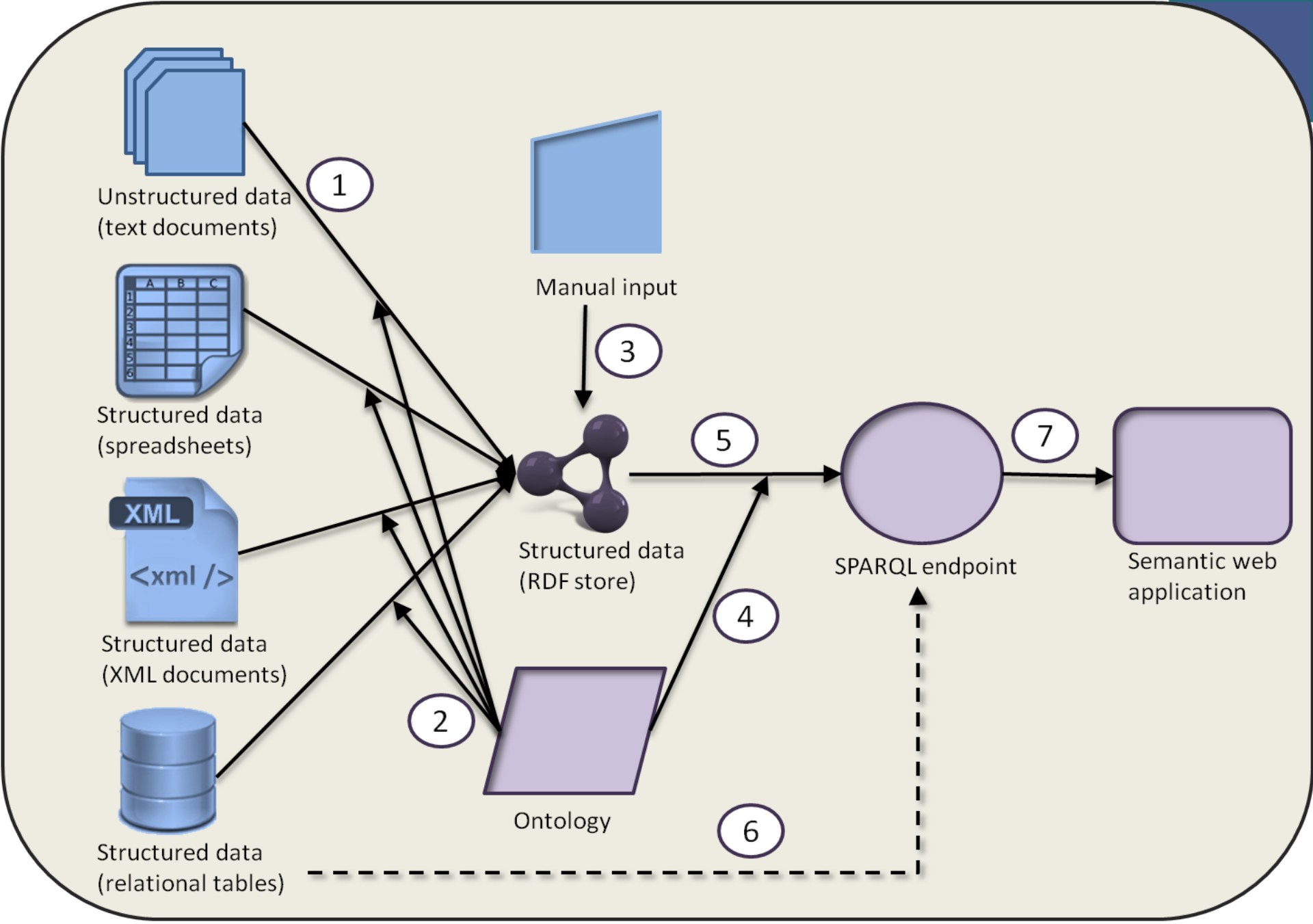
Interlinking in LODD



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

TripleMap

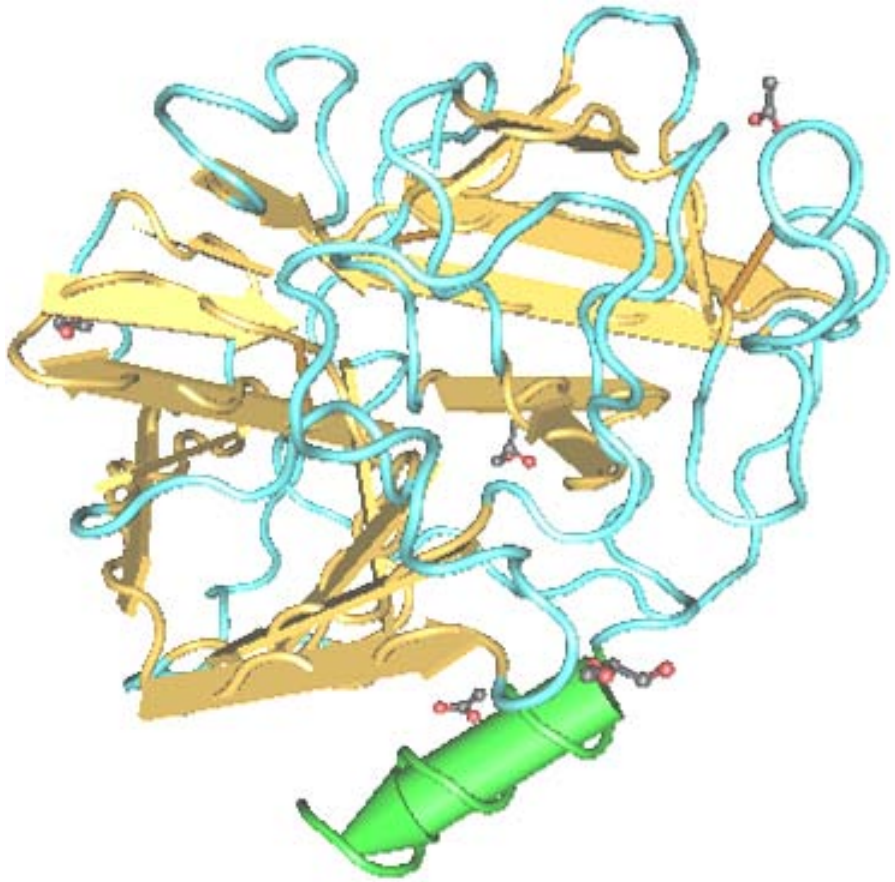

TripleMap
www.triplemap.com/app/
TripleMap
Welcome, Chris Bouton (as Admin) | Logout | Send Feedback
TripleMap File Edit Analyze View Control Cube
A 1-Year, Double-Blind, Randomized, Placebo-Controlled, Study of Rasagiline 1 mg and 2 mg Added to Aricept 10 mg Daily in Patients With Mild to Moderate Dementia of the Alzheimer's Type
Condition
Pick disease, 172700
AssociatedGene
Dementia
Gene
AssociatedGene
PSEN1
PSEN2
PossibleDrug
PossibleDiseaseTarget
AssociatedGene
Gene
AssociatedGene
Gene
74 items
(6r,1'r,2's)-5,6,7,8 ...
1,2,4-Triazole-Car...
1-(3-Mercapto-2-...
PossibleDrug
PossibleDiseaseTarget
Alzheimer_disease
(6r,1'r,2's)-5,6,7,8 Tetrahydrobiop...
ChemicalFormula
C9H15N5O3
ChemicalIupacName
(6R)-2-amino-6-[(1R,2S)-1,2-dihydroxypropyl]-5,6,7,8-tetrahydro-1H-pteridin-4-one
CreationDate
2005-06-13 13:24:05 UTC
DrugType
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugtype/smallMolecule
[more...]
GenericName
(6r,1'r,2's)-5,6,7,8 Tetrahydrobiopterin
HetId
BH4
InchIdentifier
InChI=1/C9H15N5O3/c1-3(15)6(16)4-2-11-7-5(12-4)8(17)14-9(10)13-7/h3-4,6,12,15-16H,2H2,1H3,(H4,10,11,13,14,17)/f/h11


Unstructured data
(text documents)
1
Manual input
3
Structured data
(spreadsheets)
5
7
XML
<xml />
Structured data
(XML documents)
Structured data
(RDF store)
SPARQL endpoint
Semantic web
application
4
2
Ontology
Structured data
(relational tables)
6

Homonyms

PSA

- **P**rostate **S**pecific **A**ntigen
- **PS**oriatic **A**rthritis
- alpha-2,8-**P**oly**S**ialic **A**cid
- **P**oly**S**ubstance **A**buse
- **P**icryl **S**ulfonic **A**cid
- **P**olymeric **S**ilicic **A**cid
- **P**artial **S**ensory **A**gnosia
- **P**oultry **S**cience **A**ssociation

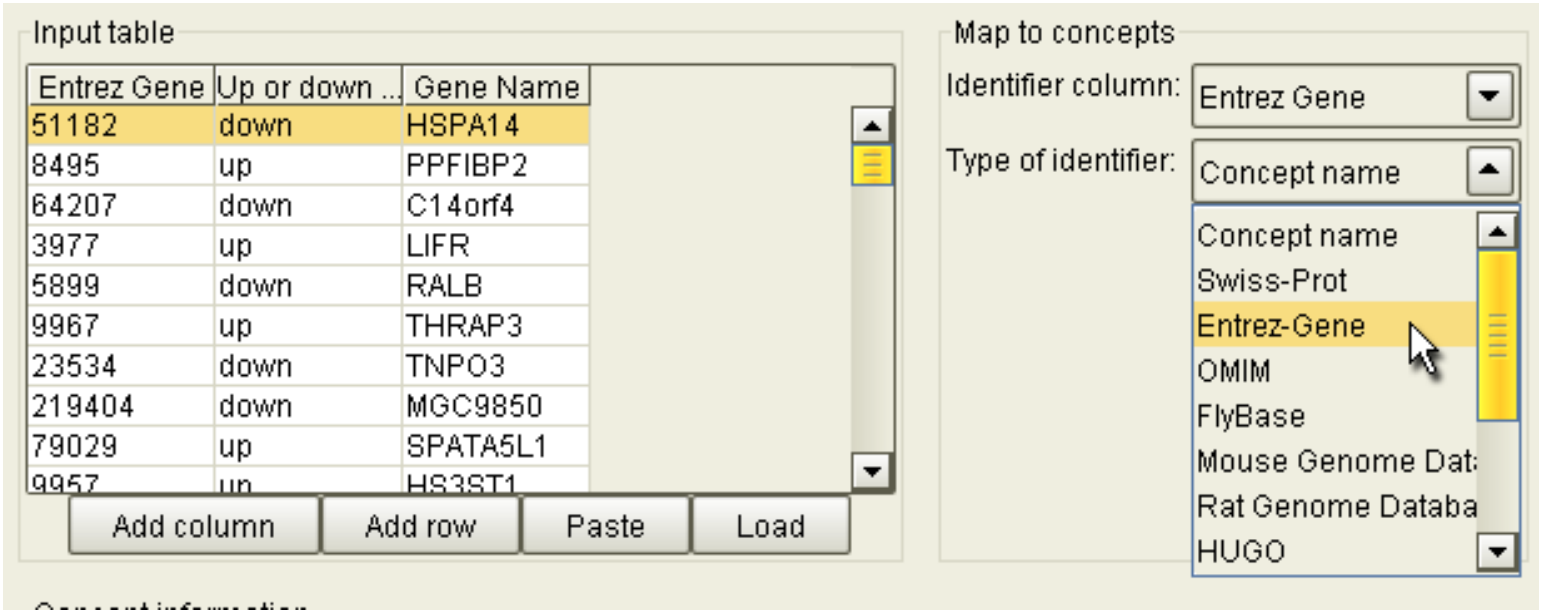


Source: Martijn Schuemie

Shared Identifiers

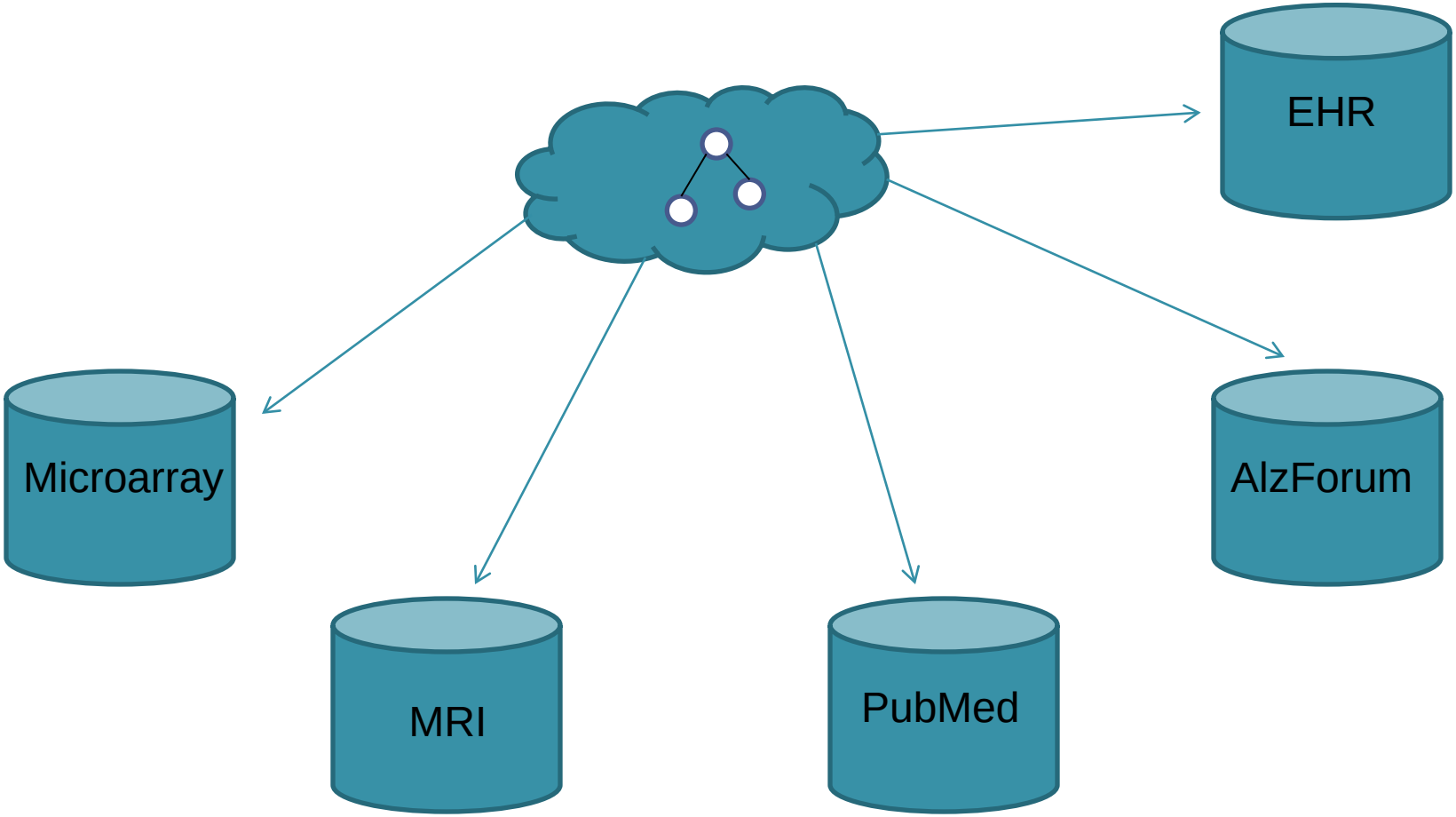
- Must use common URI's in order to link data
- Provenance related identifiers still needed:
  - Identifiers for people (researchers)
  - Identifiers for diseases
  - Identifiers for terms (Terminology servers)
  - Identifiers for programs, processes, workflows
  - Identifiers for chemical compounds
- Shared Names http://sharednames.org
- Bio2RDF

Early semantic commitment: Map input data to concepts

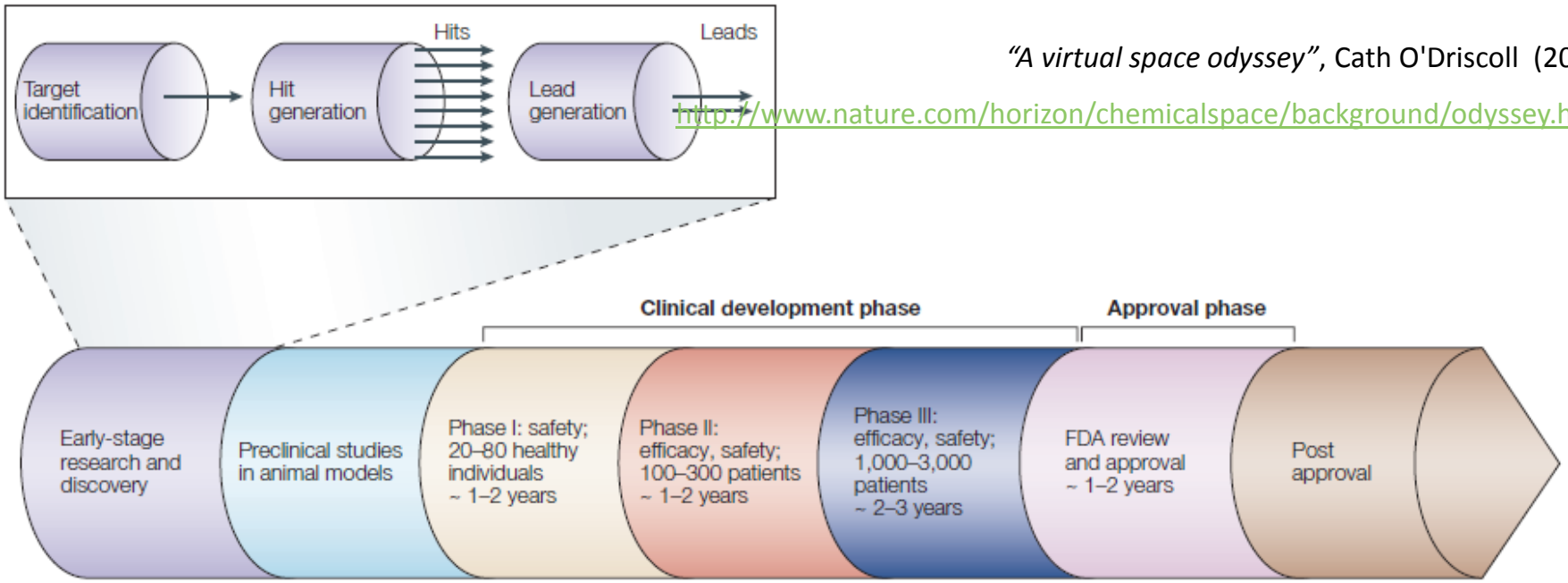


Screenshot Anni: Martijn Schuemie

TMO: Translating across domains



Questions & Problems
The Drug Development Pipeline



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

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

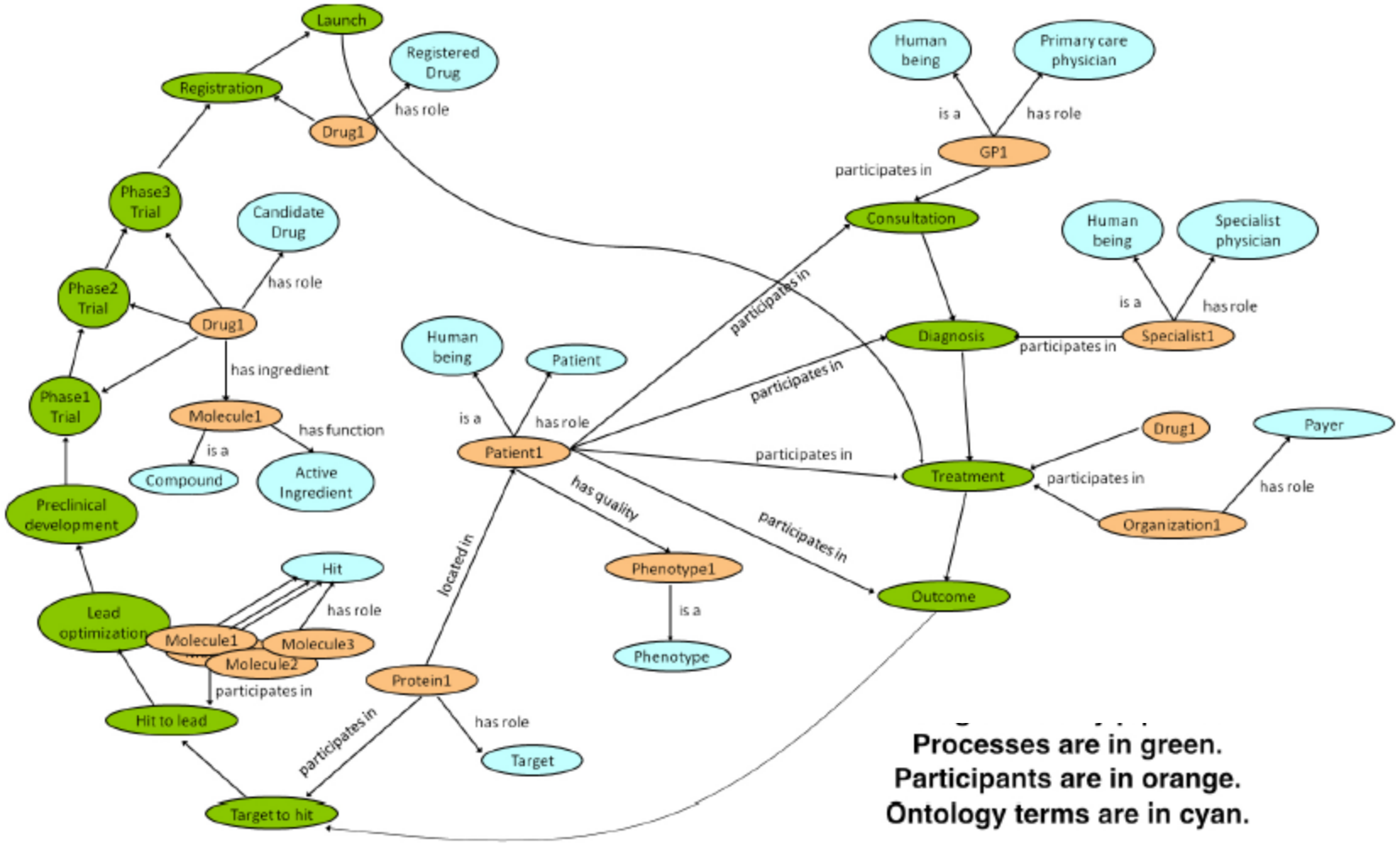
- The road is long, and costly.
- How do we contain costs and develop better drugs?

Source: Elgar Pichler

Translational Medicine Ontology Mission

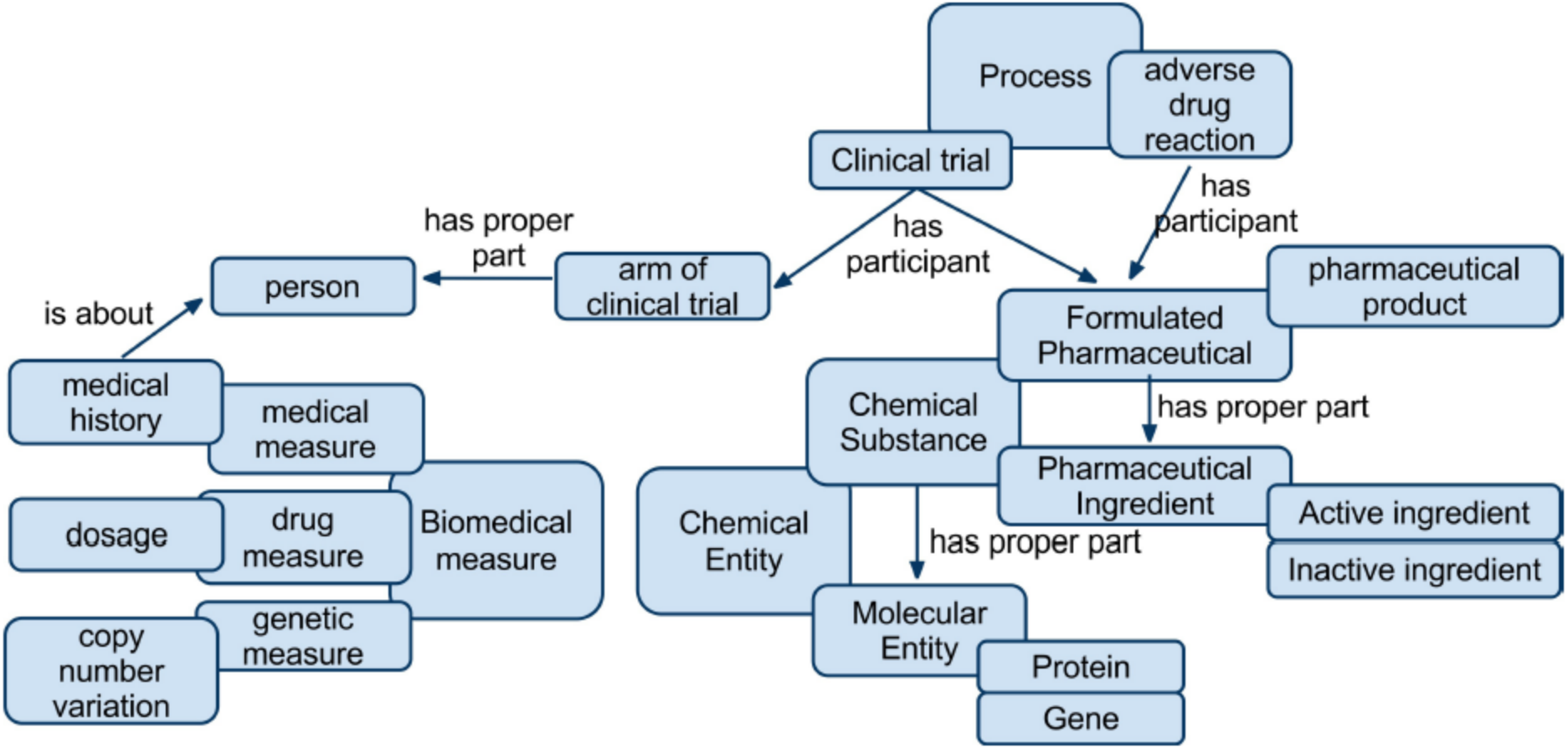
- Focuses on the development of a **high level patient-centric ontology for the pharmaceutical industry**. The ontology should enable data integration across **discovery research**, **hypothesis management**, **experimental studies**, **compounds**, **formulation**, **drug development**, **market size**, **competitive data**, **population data**, etc. 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**.

Scope of the TMO



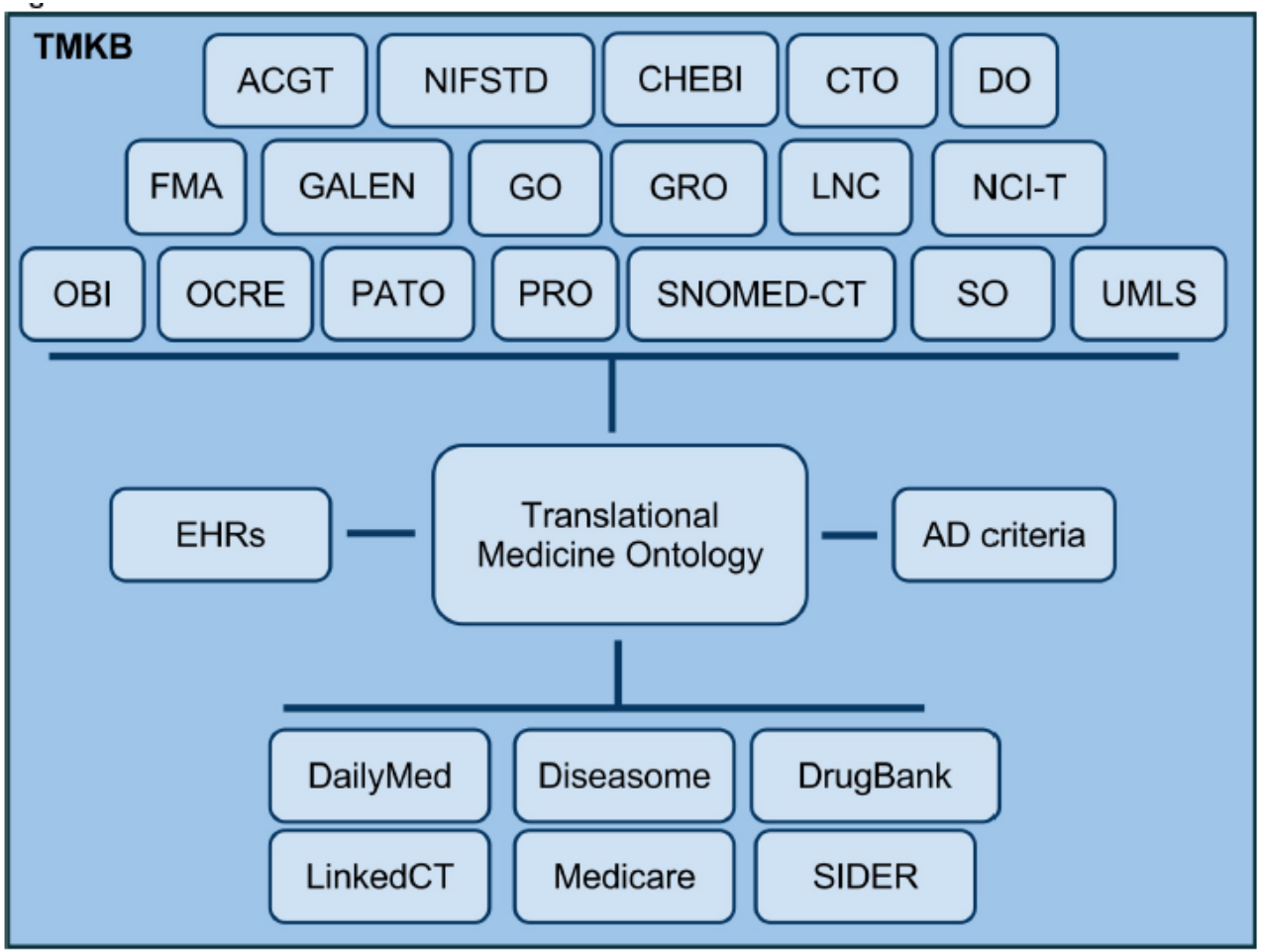
Source: Susie Stephens

TMO Structure



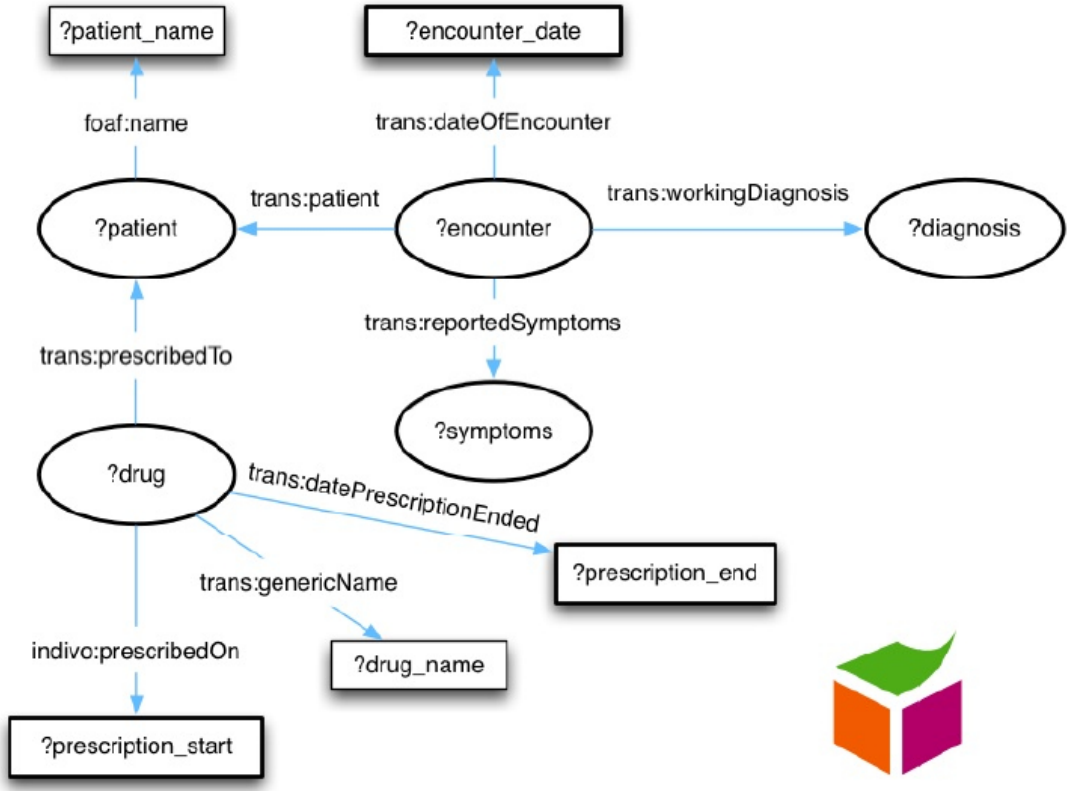
Source: Susie Stephens

Translational Medicine KB



Source: Susie Stephens

TMO Query



How many patients experienced side effects while taking Donepezil?

Source: Susie Stephens

Discovery Questions and Answers

What genes are associated with or implicated in AD?	Diseasome and PharmGKB indicate at least 97 genes have some association with AD.
Which SNPs may be potential AD biomarkers?	PharmGKB reveals 63 SNPs.
Which market drugs might potentially be repurposed for AD because they modulate AD implicated genes?	57 compounds or classes of compounds are used to treat 45 diseases, including AD, diabetes, obesity, and hyper/hypotension

Source: Susie Stephens

Clinical Trials Questions and Answers

Since my patient is suffering from drug-induced side effects for AD treatment, can an AD clinical trial with a different mechanism of action be identified?	Of the 438 drugs linked to AD trials, only 58 are in active trials and only 2 (Doxorubicin and IL-2) have a documented mechanism of action. 78 AD-associated drugs have an established MOA.
Find AD patients without the APOE4 allele as these would be good candidates for the clinical trial involving Bapineuzumab?	Of the 4 patients with AD, only one does not carry the APOE4 allele, and may be a good candidate for the clinical trial.
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.

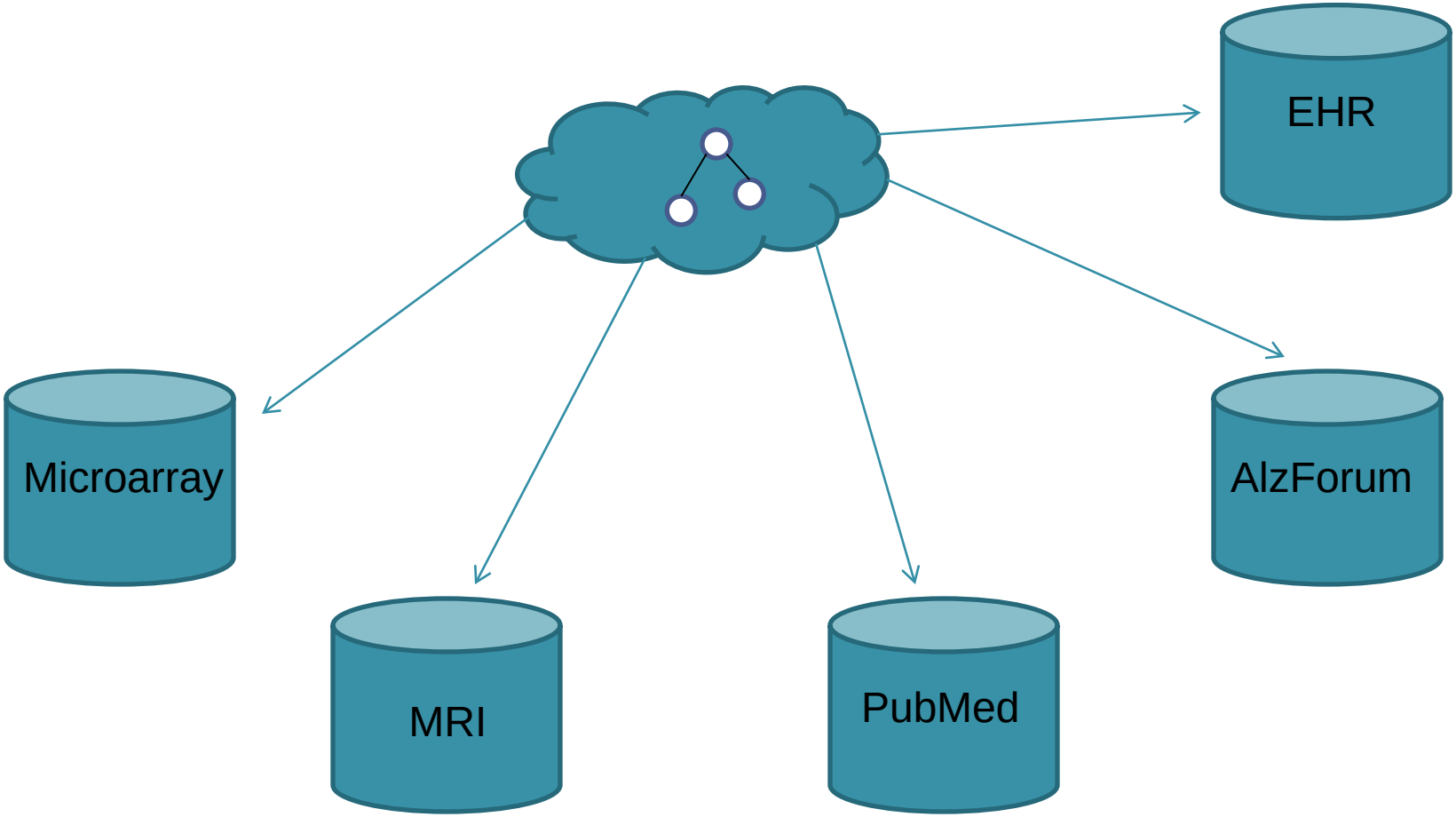
Source: Susie Stephens

Physician Questions and Answers

What are the diagnostic criteria for AD?	There are 12 diagnostic inclusion criteria and 9 exclusion criteria
Does Medicare D cover Dopenezil?	Medicare D covers two brand name formulations of Donepezil: Aricept and Aricept ODT.
Have any AD patients been treated for other neurological conditions?	Patient 2 was found to suffer from AD and depression.

Source: Susie Stephens

Terminology: Translating across domains



Terminology
Ongoing Work

- RDF representation of clinical reports
- Mammogram: Represent both **radiology** and **pathology** report to discover discrepancies
- Use Translational Medicine Ontology, RadLex, SNOMED in the RDF
- Link to data about biomarkers and therapies

- There is a 1.2 cm x 1.3 cm round mass with an indistinct margin in the left breast at 9 o'clock. This round mass is isoechoic.

Source: John Madden

```
# these are pretty commonplace assertion types
# hard to imagine much structural variation here
# can pick a specific vocabulary later
theMassAt9          size                    [dimension

            [val "1.2"^^xsd:float; unit cm],

                                                [val
"1.3"^^xsd:float; unit cm]].


# more sketchy, depends on chosen anatomy vocabulary
theMassAt9          location                [a Location;

            in thePatient,

                    [a Breast; laterality left],

                    [rdfs:label "9:00"]].


# very sketchy
# these are modeled loosey-goosey as value partitions
# need no idea how e.g. Radlex might do this
theMassAt9          shape                       round.
theMassAt9          margin                      indistinct.
```

Source: John Madden

Barriers to data sharing: social, legal, and technical

- *"Biologists would rather share their toothbrush than share a gene name"*
  - Don't want to get "scooped" for a publication and potentially lose years of work and Ph.D. material.
  - Competition for grants
- Need clarity and transparency about threats to patient privacy
- Many data formats, example: CDISC and HL7 RIM
- Most researchers do not feel the need to look at data from neighboring domains (cross-disciplinary studies)

Summary

- The data landscape for personalized medicine is highly fragmented
- Public vocabulary services can be used to connect data sets and make them accessible on the Web
- Data sharing can add value to data through linking
- Best practices for important data sources: microarray data, image data
- Data stewardship – serve data back to community

Acknowledgements

- W3C Health Care and Life Science Interest Group, http://www.w3.org/blog/hcls
- National Center for Biomedical Ontologies
- Concept Web Alliance
- Authors of all contributed slides

The End

"Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house."

– Henri Poincaré,

Science and Hypothesis, 1905