And then magic happens…

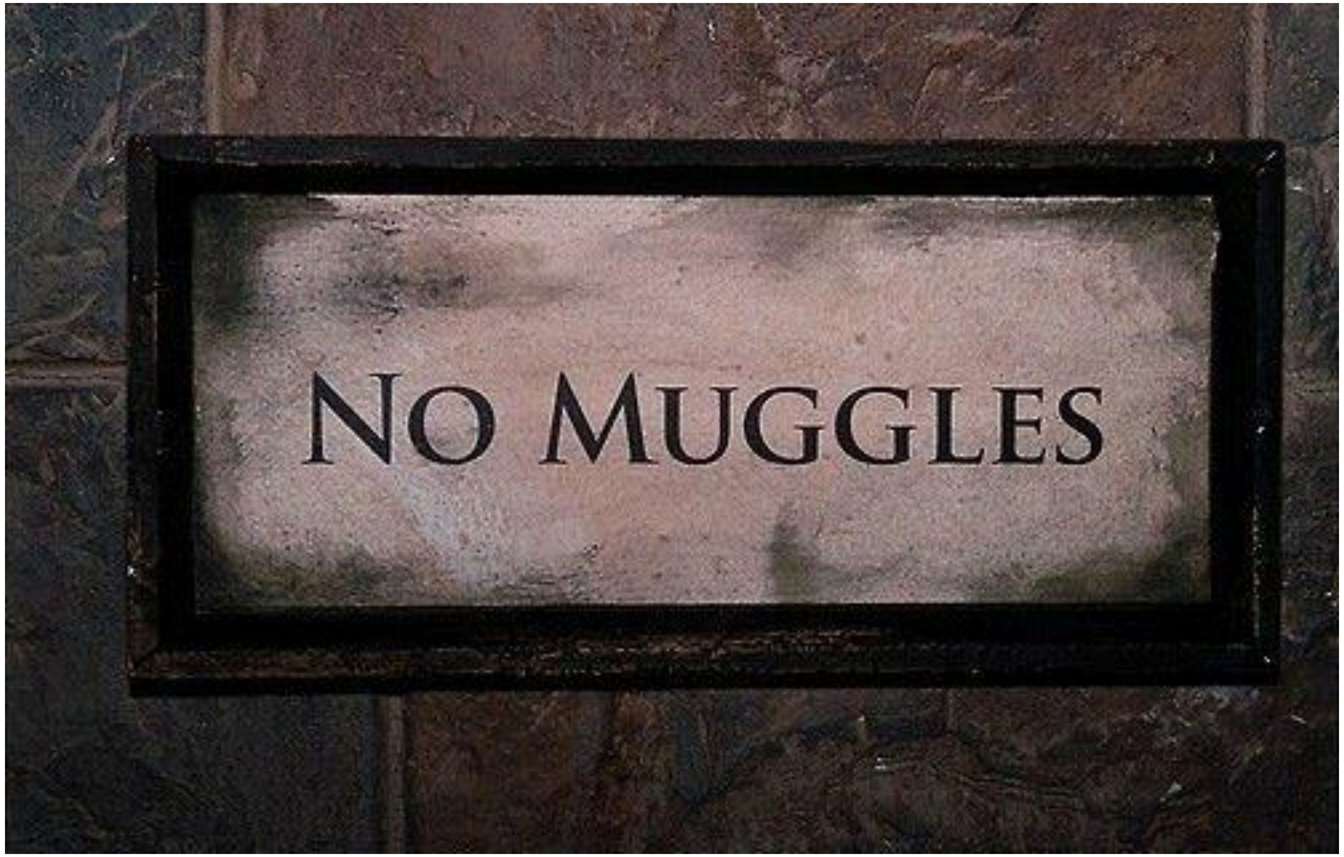
@Chris_Evelo

Maastricht University
WikiPathways team
ELIXIR Interoperability team
Open PHACTS

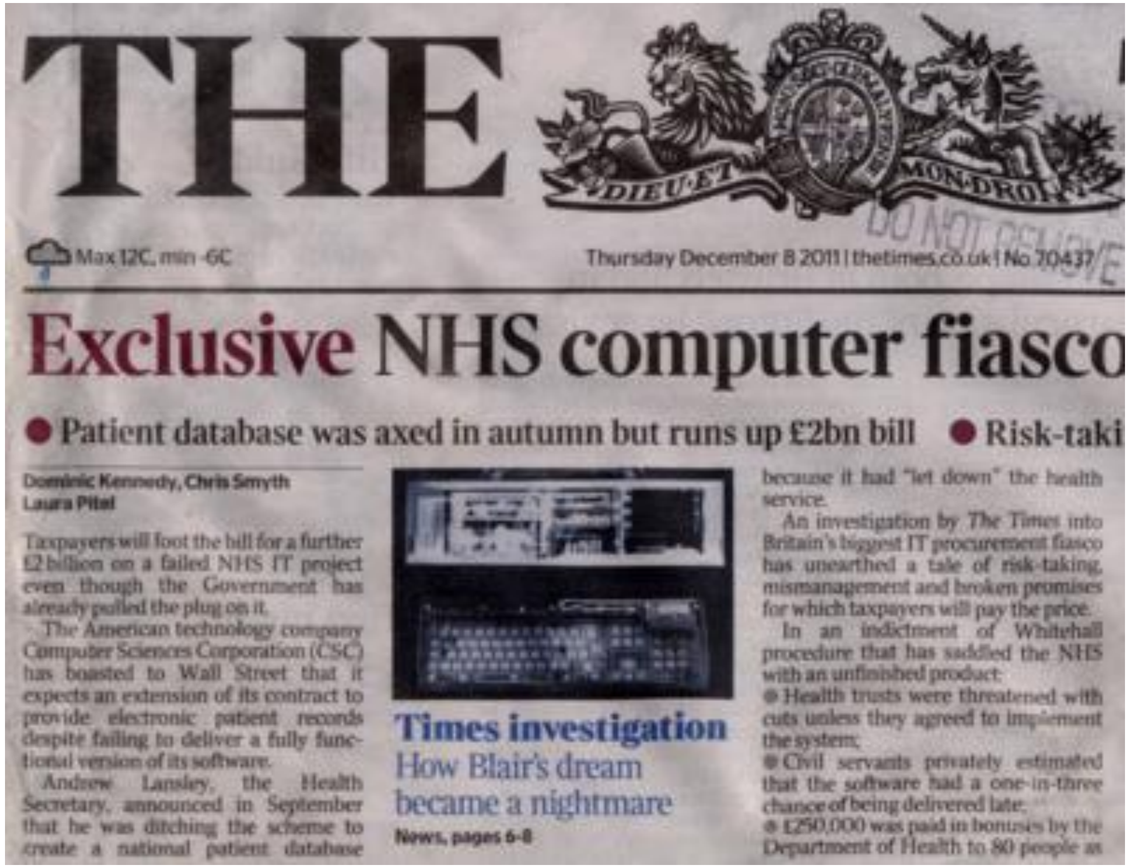
Combined CHARME – EMBnet and NETTAB 2016 Workshop

Reproducibility, standards and SOP in bioinformatics

October 25-26 2016, Rome



If we don’t do the magic


DIEU ET MON DROI

THE

DO NOT REMOVE

Max 12C, min -6C

Thursday December 8 2011 | thetimes.co.uk | No 70437

Exclusive NHS computer fiasco

● Patient database was axed in autumn but runs up £2bn bill ● Risk-taki

Dominic Kennedy, Chris Smyth
Laura Pitel

Taxpayers will foot the bill for a further £2 billion on a failed NHS IT project even though the Government has already pulled the plug on it.

The American technology company Computer Sciences Corporation (CSC) has boasted to Wall Street that it expects an extension of its contract to provide electronic patient records despite failing to deliver a fully functional version of its software.

Andrew Lansley, the Health Secretary, announced in September that he was ditching the scheme to create a national patient database because it had "let down" the health service.

An investigation by *The Times* into Britain's biggest IT procurement fiasco has unearthed a tale of risk-taking, mismanagement and broken promises for which taxpayers will pay the price.

In an indictment of Whitehall procedure that has saddled the NHS with an unfinished product:

- Health trusts were threatened with cuts unless they agreed to implement the system;
- Civil servants privately estimated that the software had a one-in-three chance of being delivered late;
- £250,000 was paid in bonuses by the Department of Health to 80 people as

Times investigation
How Blair's dream became a nightmare
News, pages 6-8

Lessons Learned from caBIG™

NCTR

January 21, 2010

Warren Kibbe
Northwestern University
wakibbe@northwestern.edu

caBIG, the hard parts

- Progress is hard to measure
- Delivering exactly what people want is *not* the mission of caBIG
- Overspecifying, model-first, and semantic inflexibility are all enemies of caBIG's need to be agile and flexible
- The products are harder to use than anyone would like, but that is the nature of the space, so far…

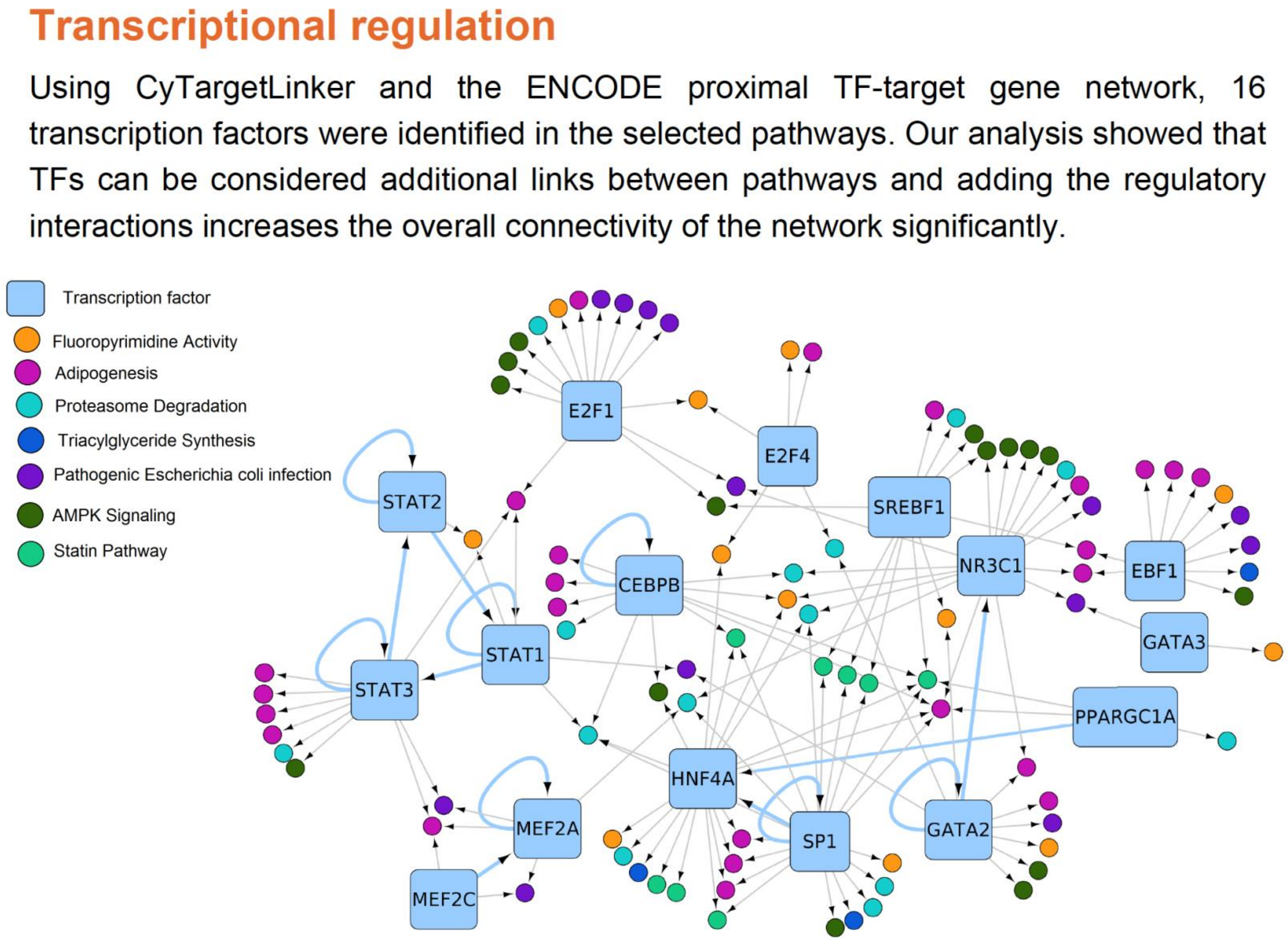
caBIG, revisited

- Fuzzy semantics and semantic web technologies will be embraced
- Complete specification is an unrealistic goal
- Architect for change

Comfort Zone
Where the
MAGIC
Happens

2015 Jeannel King

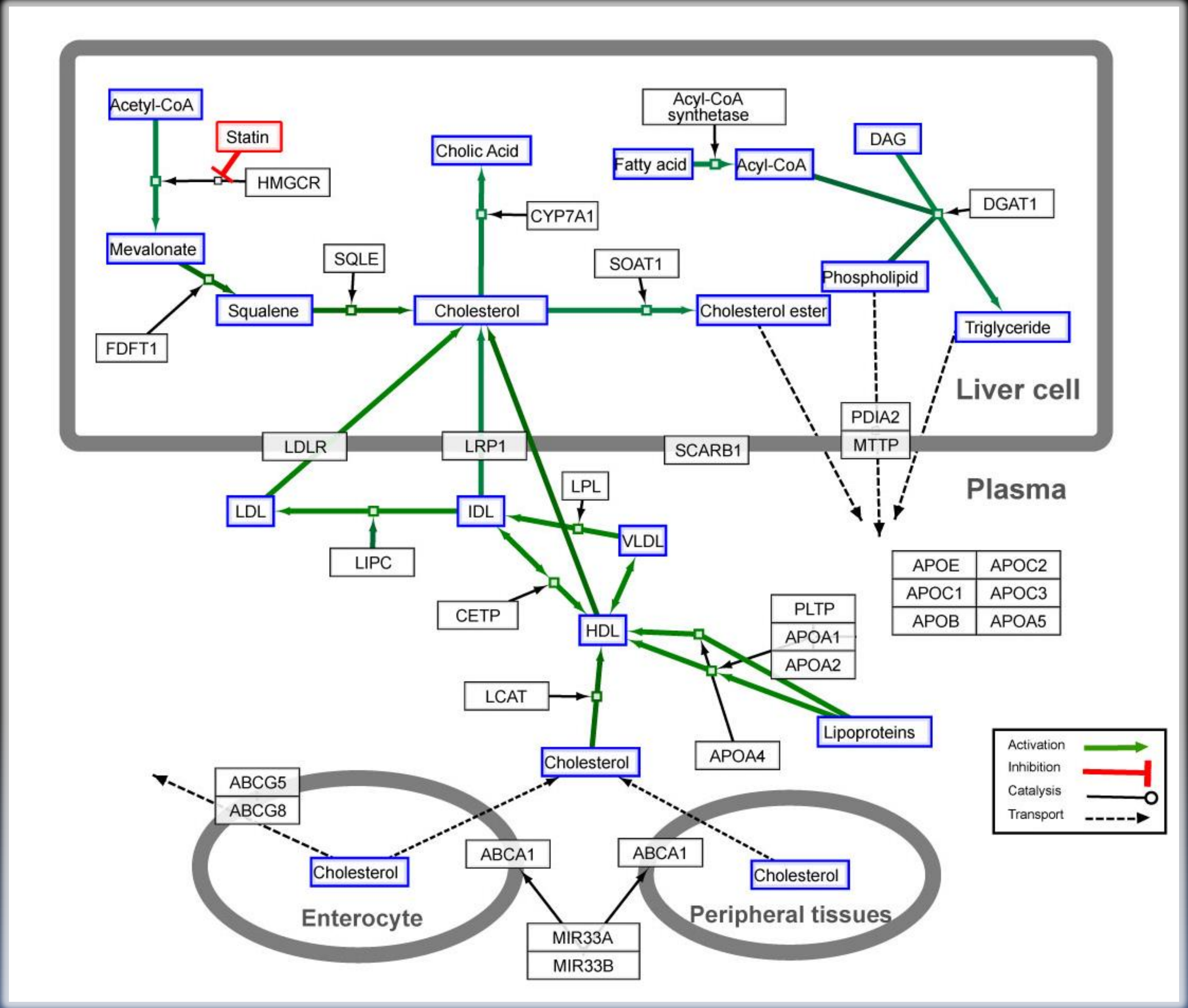
We can do things like this…



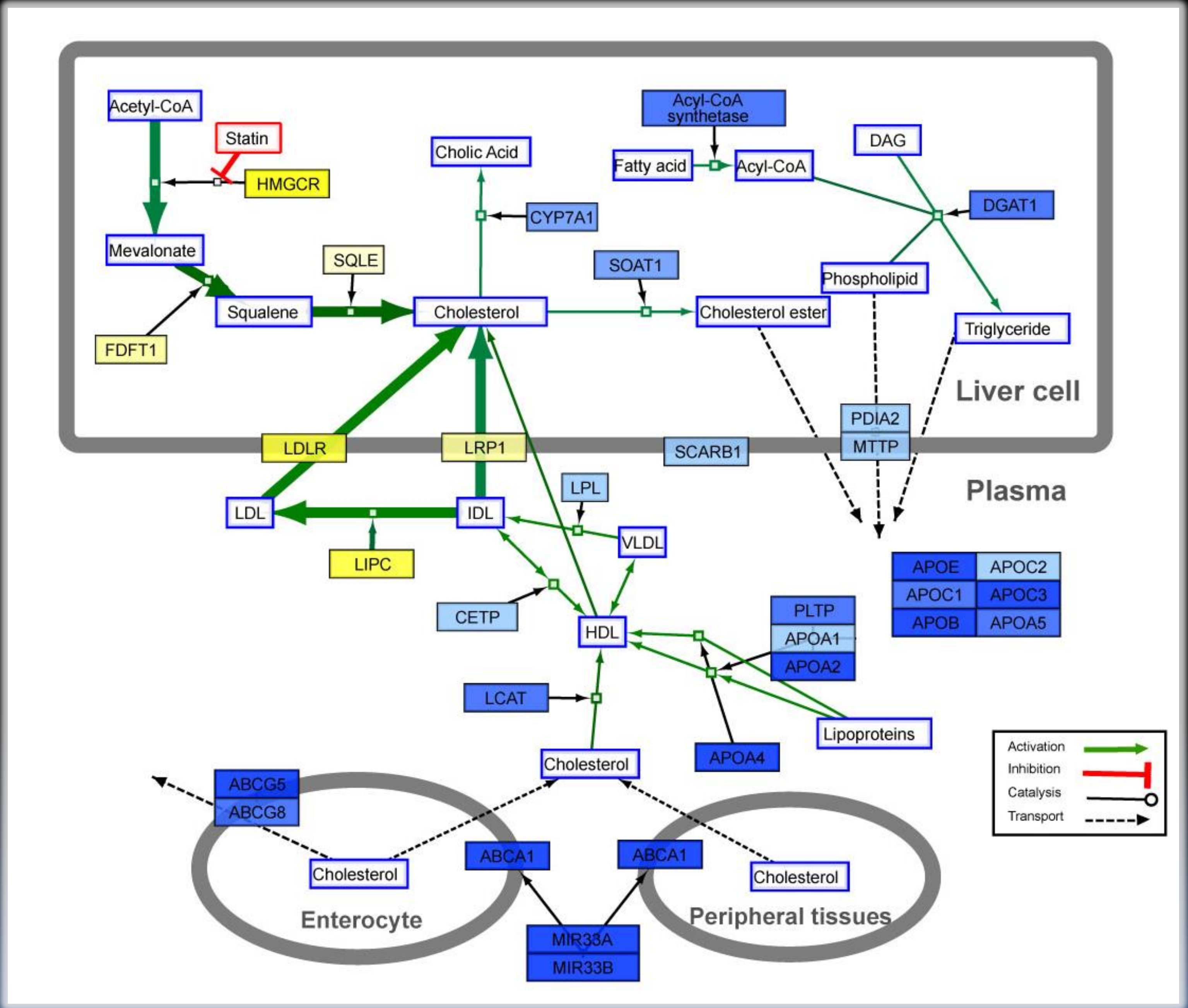
Transcriptional regulation

Using CyTargetLinker and the ENCODE proximal TF-target gene network, 16 transcription factors were identified in the selected pathways. Our analysis showed that TFs can be considered additional links between pathways and adding the regulatory interactions increases the overall connectivity of the network significantly.

■ Transcription factor
● Fluoropyrimidine Activity
● Adipogenesis
● Proteasome Degradation
● Triacylglyceride Synthesis
● Pathogenic Escherichia coli infection
● AMPK Signaling
● Statin Pathway

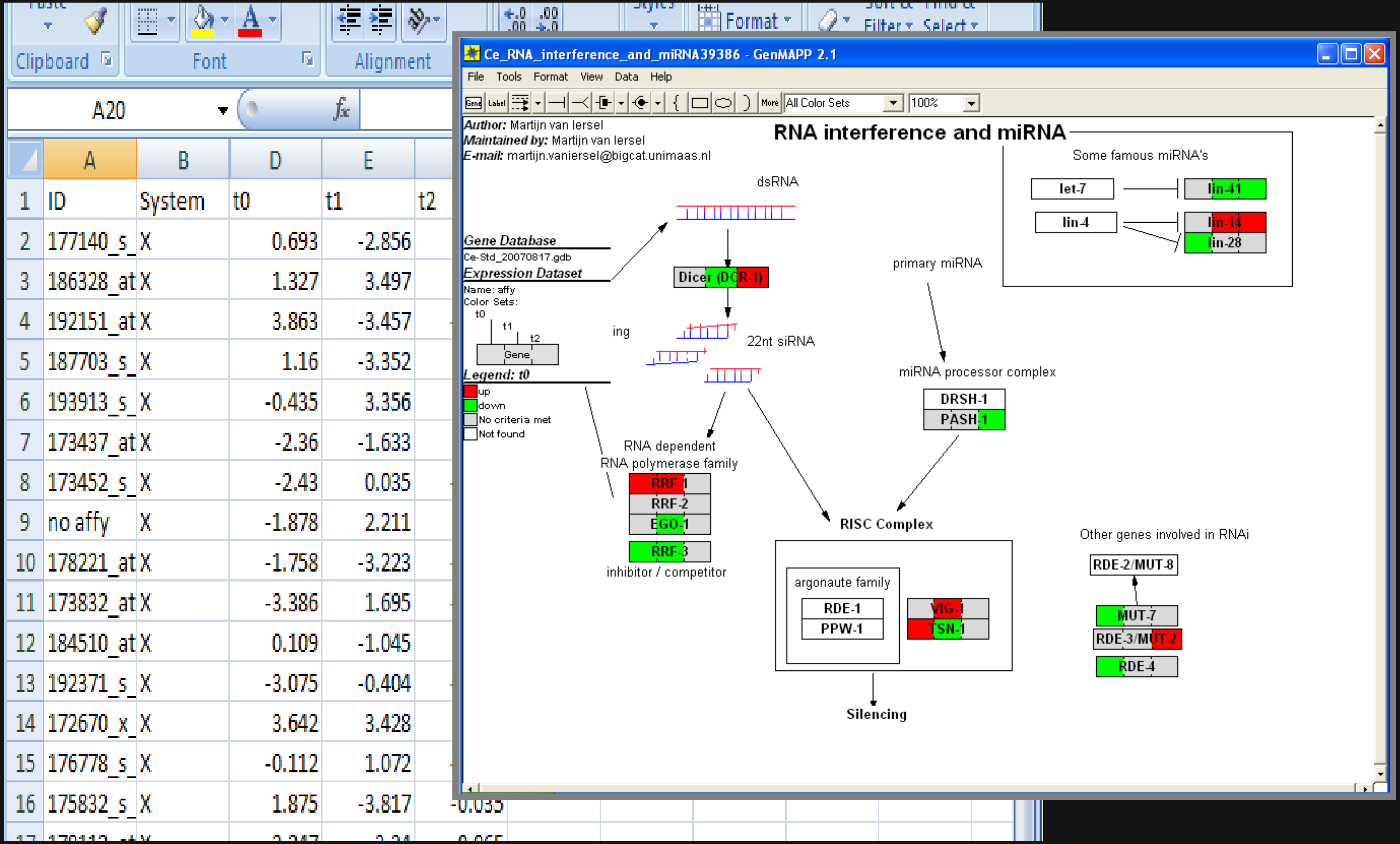


http://www.wikipathways.org/index.php/Pathway:WP430



http://www.wikipathways.org/index.php/Pathway:WP430

How to do data visualization?



	A	B	D	E	F
1	ID	System	t0	t1	t2
2	177140_s_	X	0.693	-2.856	
3	186328_at	X	1.327	3.497	
4	192151_at	X	3.863	-3.457	
5	187703_s_	X	1.16	-3.352	
6	193913_s_	X	-0.435	3.356	
7	173437_at	X	-2.36	-1.633	
8	173452_s_	X	-2.43	0.035	
9	no affy	X	-1.878	2.211	
10	178221_at	X	-1.758	-3.223	
11	173832_at	X	-3.386	1.695	
12	184510_at	X	0.109	-1.045	
13	192371_s_	X	-3.075	-0.404	
14	172670_x_	X	3.642	3.428	
15	176778_s_	X	-0.112	1.072	
16	175832_s_	X	1.875	-3.817	-0.035

Ce_RNA_interference_and_miRNA39386 - GenMAPP 2.1

Author: Martijn van Iersel
Maintained by: Martijn van Iersel
E-mail: martijn.vaniersel@bigcat.unimaas.nl

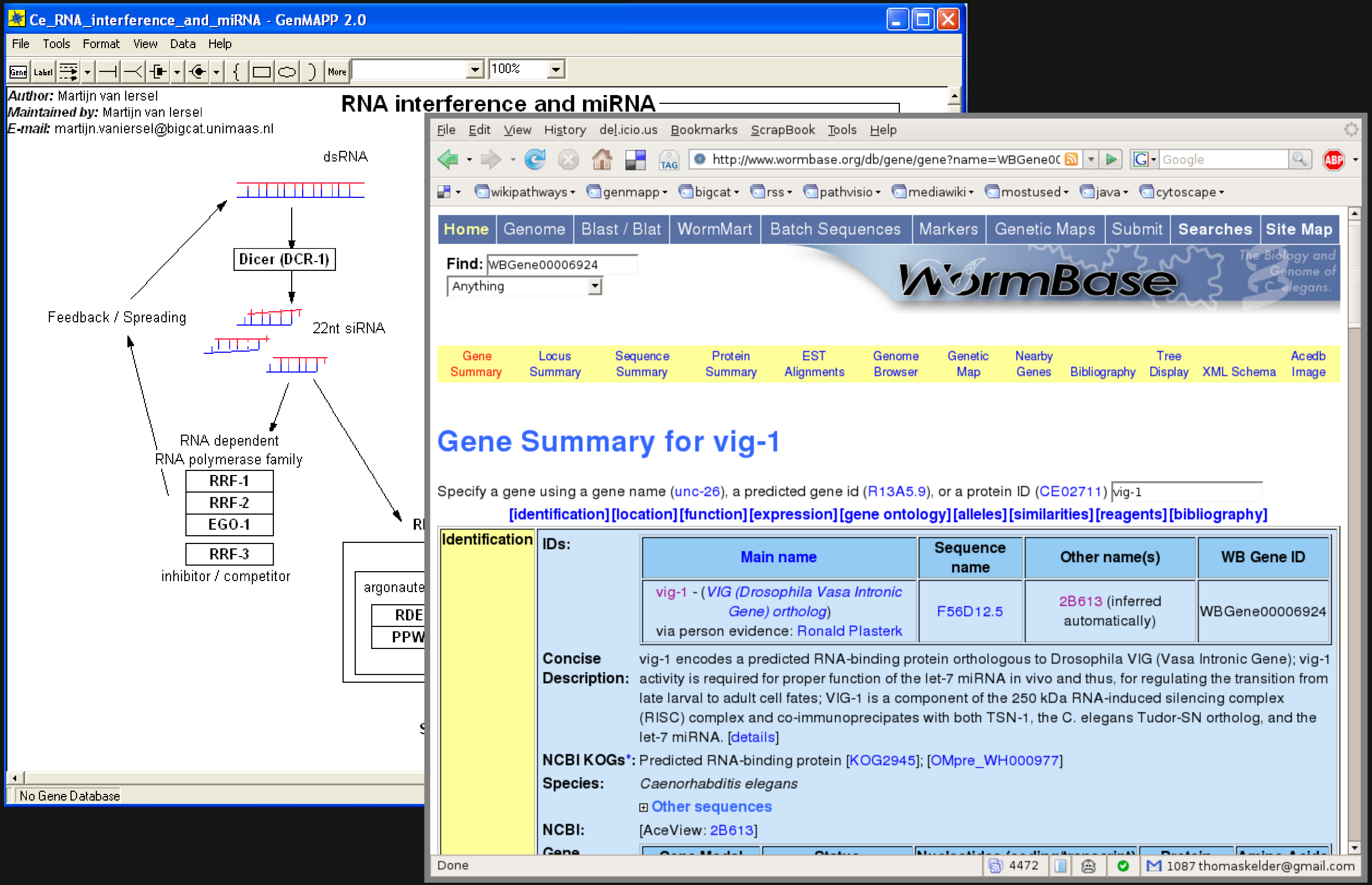
Gene Database
Ce-Std_20070817.gdb
Expression Dataset
Name: affy
Color Sets:

Legend: t0
🟥 up
🟩 down
□ No criteria met
□ Not found

]

Solve
My
Problem

Connect to Genome Databases



Author: Martijn van Iersel
Maintained by: Martijn van Iersel
E-mail: martijn.vaniersel@bigcat.unimaas.nl

Gene Summary for vig-1

Specify a gene using a gene name (unc-26), a predicted gene id (R13A5.9), or a protein ID (CE02711) vig-1

[identification] [location] [function] [expression] [gene ontology] [alleles] [similarities] [reagents] [bibliography]

Identification

IDs:

Main name	Sequence name	Other name(s)	WB Gene ID
vig-1 - (*VIG (Drosophila Vasa Intronic Gene) ortholog*) via person evidence: Ronald Plasterk	F56D12.5	2B613 (inferred automatically)	WBGene00006924

Concise Description: vig-1 encodes a predicted RNA-binding protein orthologous to Drosophila VIG (Vasa Intronic Gene); vig-1 activity is required for proper function of the let-7 miRNA in vivo and thus, for regulating the transition from late larval to adult cell fates; VIG-1 is a component of the 250 kDa RNA-induced silencing complex (RISC) complex and co-immunoprecipates with both TSN-1, the C. elegans Tudor-SN ortholog, and the let-7 miRNA. [details]

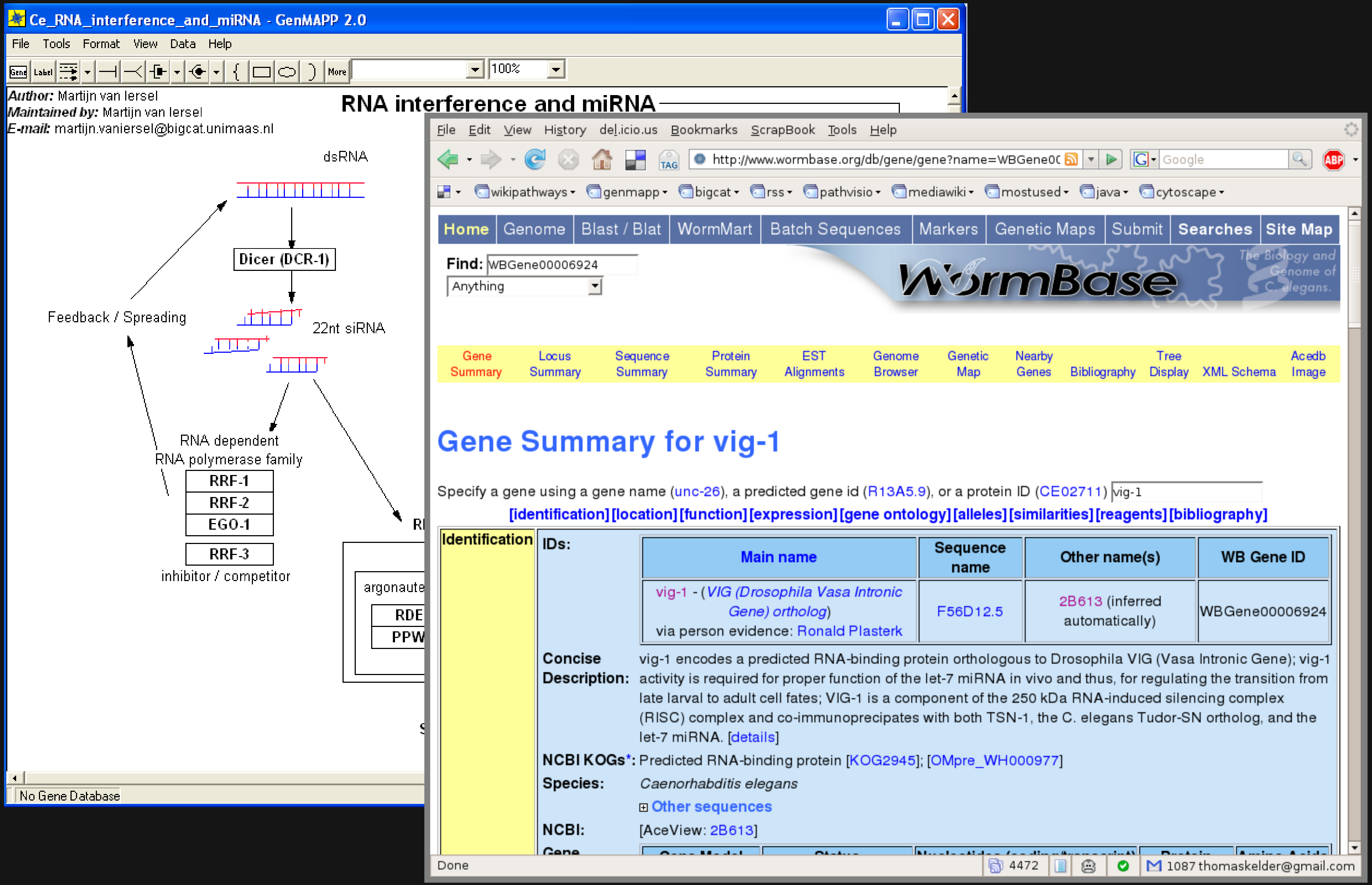
NCBI KOGs*: Predicted RNA-binding protein [KOG2945]; [OMpre_WH000977]

Species: *Caenorhabditis elegans*

⊞ Other sequences

NCBI: [AceView: 2B613]

Connect to Genome Databases



Ce_RNA_interference_and_miRNA - GenMAPP 2.0
File Tools Format View Data Help
Gene Label More 100%
dsRNA
Dicer (DCR-1)
22nt siRNA
Feedback / Spreading
RNA dependent RNA polymerase family
RRF-1
RRF-2
EGO-1
RRF-3
inhibitor / competitor
argonaute
RDE
PPW
No Gene Database
File Edit View History del.icio.us Bookmarks ScrapBook Tools Help
http://www.wormbase.org/db/gene/gene?name=WBGene0(
Google
wikipathways genmapp bigcat rss pathvisio mediawiki mostused java cytoscape
Home Genome Blast / Blat WormMart Batch Sequences Markers Genetic Maps Submit Searches Site Map
Find: WBGene00006924
Anything
WormBase The Biology and Genome of C. elegans.
Gene Summary Locus Summary Sequence Summary Protein Summary EST Alignments Genome Browser Genetic Map Nearby Genes Bibliography Tree Display XML Schema Acedb Image
Done
4472
1087 thomaskelder@gmail.com

Author: Martijn van Iersel
Maintained by: Martijn van Iersel
E-mail: martijn.vaniersel@bigcat.unimaas.nl

RNA interference and miRNA

Gene Summary for vig-1

Specify a gene using a gene name (unc-26), a predicted gene id (R13A5.9), or a protein ID (CE02711) vig-1

[identification] [location] [function] [expression] [gene ontology] [alleles] [similarities] [reagents] [bibliography]

Identification

IDs:

Main name	Sequence name	Other name(s)	WB Gene ID
vig-1 - (*VIG (Drosophila Vasa Intronic Gene) ortholog*) via person evidence: Ronald Plasterk	F56D12.5	2B613 (inferred automatically)	WBGene00006924

Concise Description: vig-1 encodes a predicted RNA-binding protein orthologous to Drosophila VIG (Vasa Intronic Gene); vig-1 activity is required for proper function of the let-7 miRNA in vivo and thus, for regulating the transition from late larval to adult cell fates; VIG-1 is a component of the 250 kDa RNA-induced silencing complex (RISC) complex and co-immunoprecipitates with both TSN-1, the C. elegans Tudor-SN ortholog, and the let-7 miRNA. [details]

NCBI KOGs*: Predicted RNA-binding protein [KOG2945]; [OMpre_WH000977]

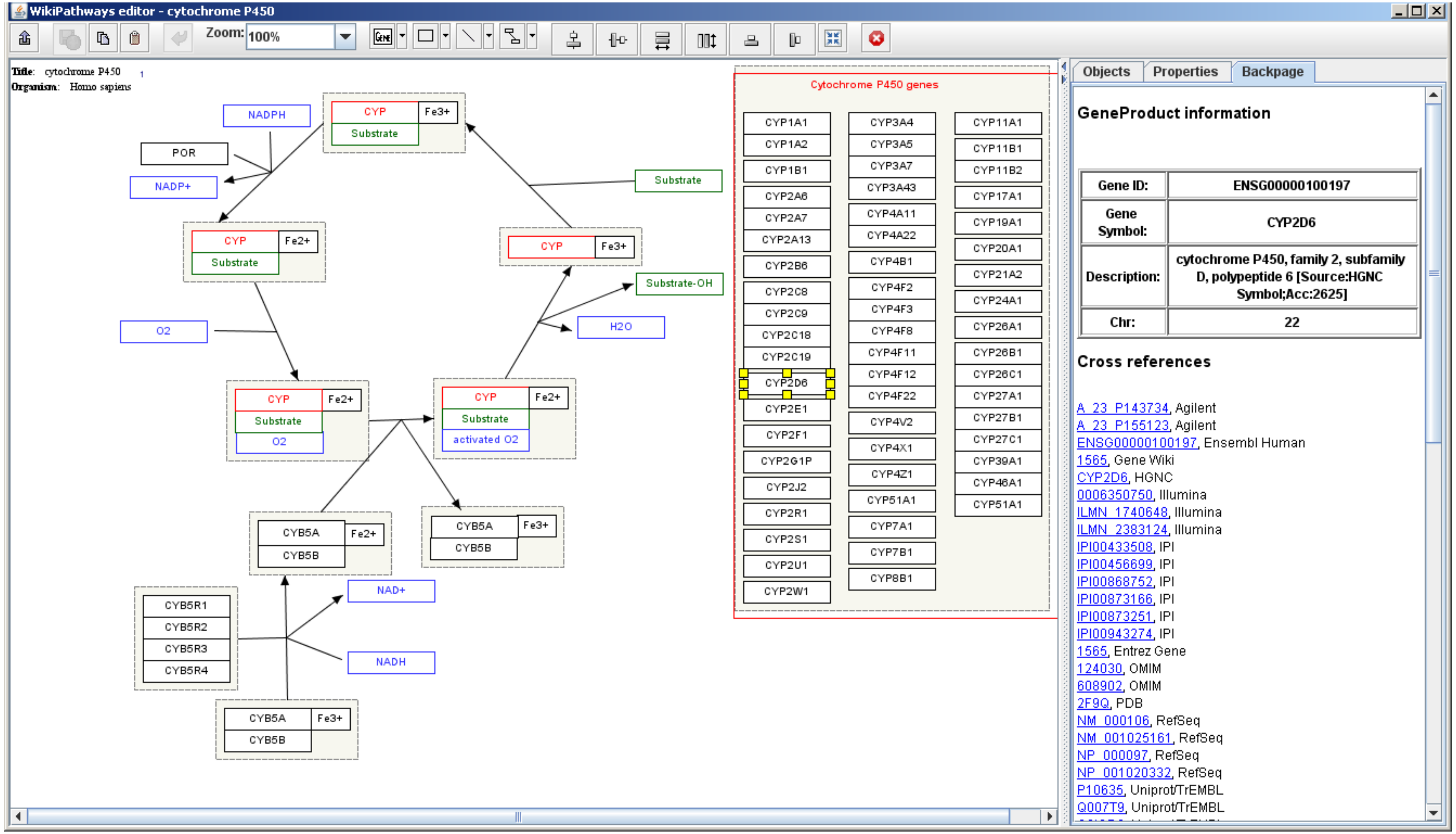
Species: *Caenorhabditis elegans*

⊞ Other sequences

NCBI: [AceView: 2B613]

Gene

Backpages link to multiple databases



Cytochrome P450 genes

CYP1A1
CYP1A2
CYP1B1
CYP2A6
CYP2A7
CYP2A13
CYP2B6
CYP2C8
CYP2C9
CYP2C18
CYP2C19
CYP2D6
CYP2E1
CYP2F1
CYP2G1P
CYP2J2
CYP2R1
CYP2S1
CYP2U1
CYP2W1

CYP3A4
CYP3A5
CYP3A7
CYP3A43
CYP4A11
CYP4A22
CYP4B1
CYP4F2
CYP4F3
CYP4F8
CYP4F11
CYP4F12
CYP4F22
CYP4V2
CYP4X1
CYP4Z1
CYP51A1
CYP7A1
CYP7B1
CYP8B1

CYP11A1
CYP11B1
CYP11B2
CYP17A1
CYP19A1
CYP20A1
CYP21A2
CYP24A1
CYP26A1
CYP26B1
CYP26C1
CYP27A1
CYP27B1
CYP27C1
CYP39A1
CYP46A1
CYP51A1

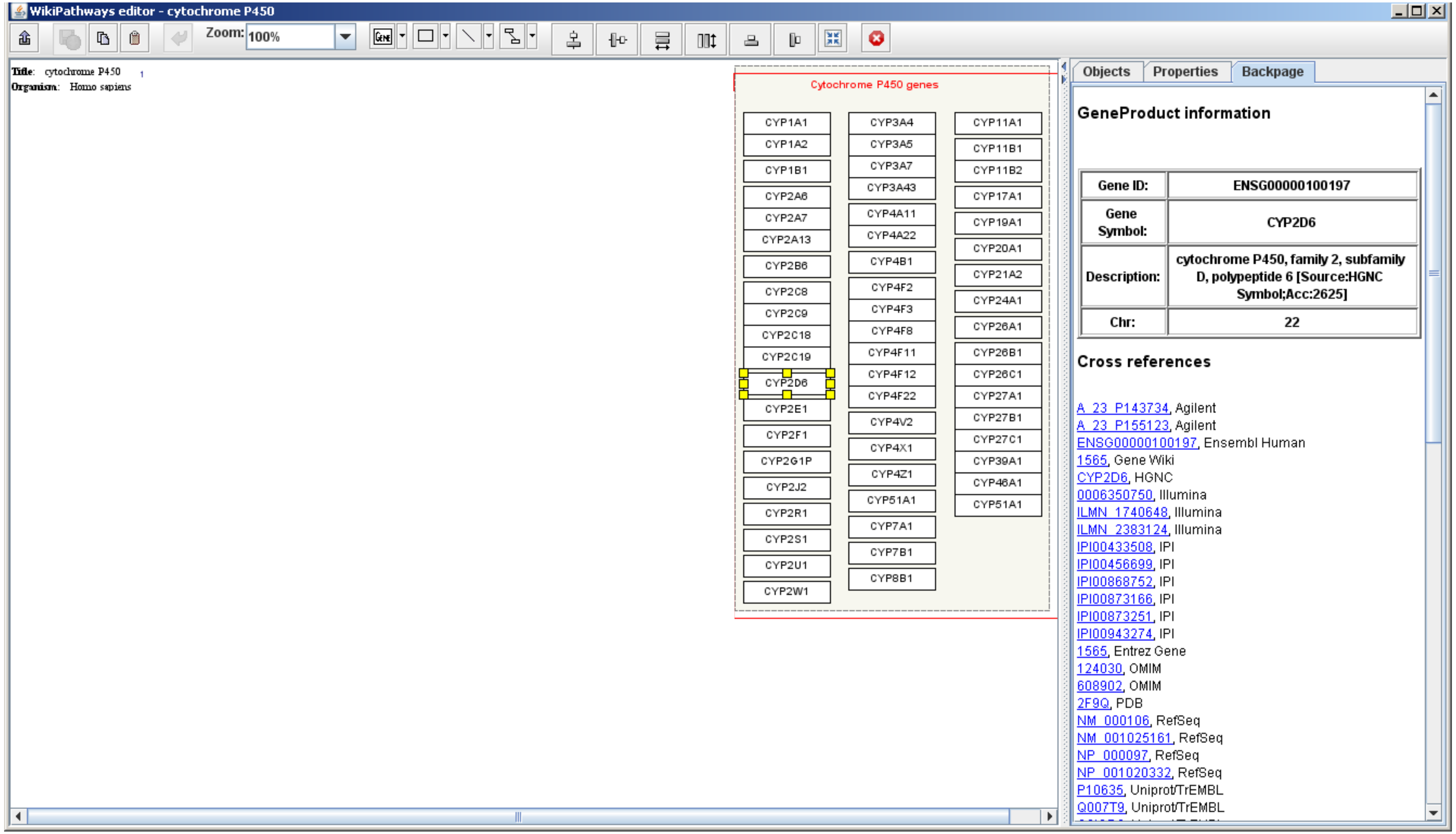
GeneProduct information

Gene ID:	ENSG00000100197
Gene Symbol:	CYP2D6
Description:	cytochrome P450, family 2, subfamily D, polypeptide 6 [Source:HGNC Symbol;Acc:2625]
Chr:	22

Cross references

A_23_P143734, Agilent
A_23_P155123, Agilent
ENSG00000100197, Ensembl Human
1565, Gene Wiki
CYP2D6, HGNC
0006350750, Illumina
ILMN_1740648, Illumina
ILMN_2383124, Illumina
IPI00433508, IPI
IPI00456699, IPI
IPI00868752, IPI
IPI00873166, IPI
IPI00873251, IPI
IPI00943274, IPI
1565, Entrez Gene
124030, OMIM
608902, OMIM
2F9Q, PDB
NM_000106, RefSeq
NM_001025161, RefSeq
NP_000097, RefSeq
NP_001020332, RefSeq
P10635, Uniprot/TrEMBL
Q007T9, Uniprot/TrEMBL

You could do this for gene lists



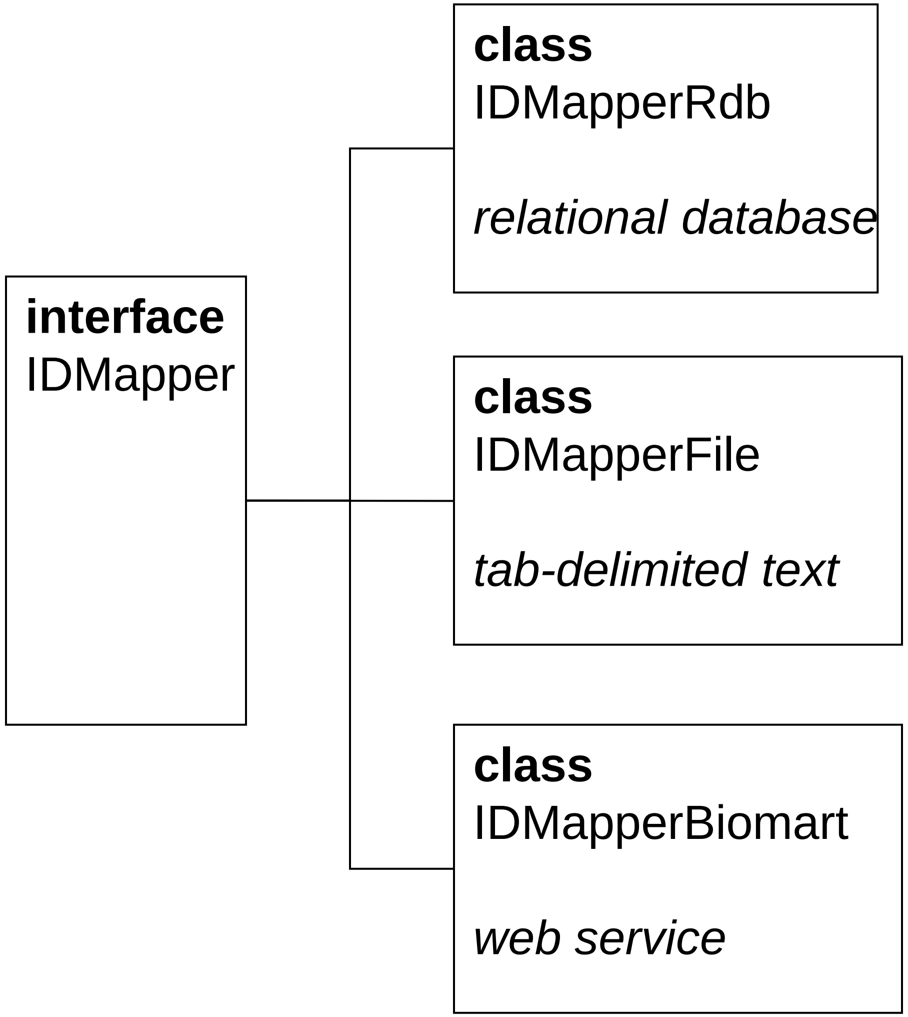
GeneProduct information

Gene ID:	ENSG00000100197
Gene Symbol:	CYP2D6
Description:	cytochrome P450, family 2, subfamily D, polypeptide 6 [Source:HGNC Symbol;Acc:2625]
Chr:	22

Cross references

A_23_P143734, Agilent
A_23_P155123, Agilent
ENSG00000100197, Ensembl Human
1565, Gene Wiki
CYP2D6, HGNC
0006350750, Illumina
ILMN_1740648, Illumina
ILMN_2383124, Illumina
IPI00433508, IPI
IPI00456699, IPI
IPI00868752, IPI
IPI00873166, IPI
IPI00873251, IPI
IPI00943274, IPI
1565, Entrez Gene
124030, OMIM
608902, OMIM
2F9Q, PDB
NM_000106, RefSeq
NM_001025161, RefSeq
NP_000097, RefSeq
NP_001020332, RefSeq
P10635, Uniprot/TrEMBL
Q007T9, Uniprot/TrEMBL

BridgeDb: Abstraction Layer



interface IDMapper

class IDMapperRdb

relational database

class IDMapperFile

tab-delimited text

class IDMapperBiomart

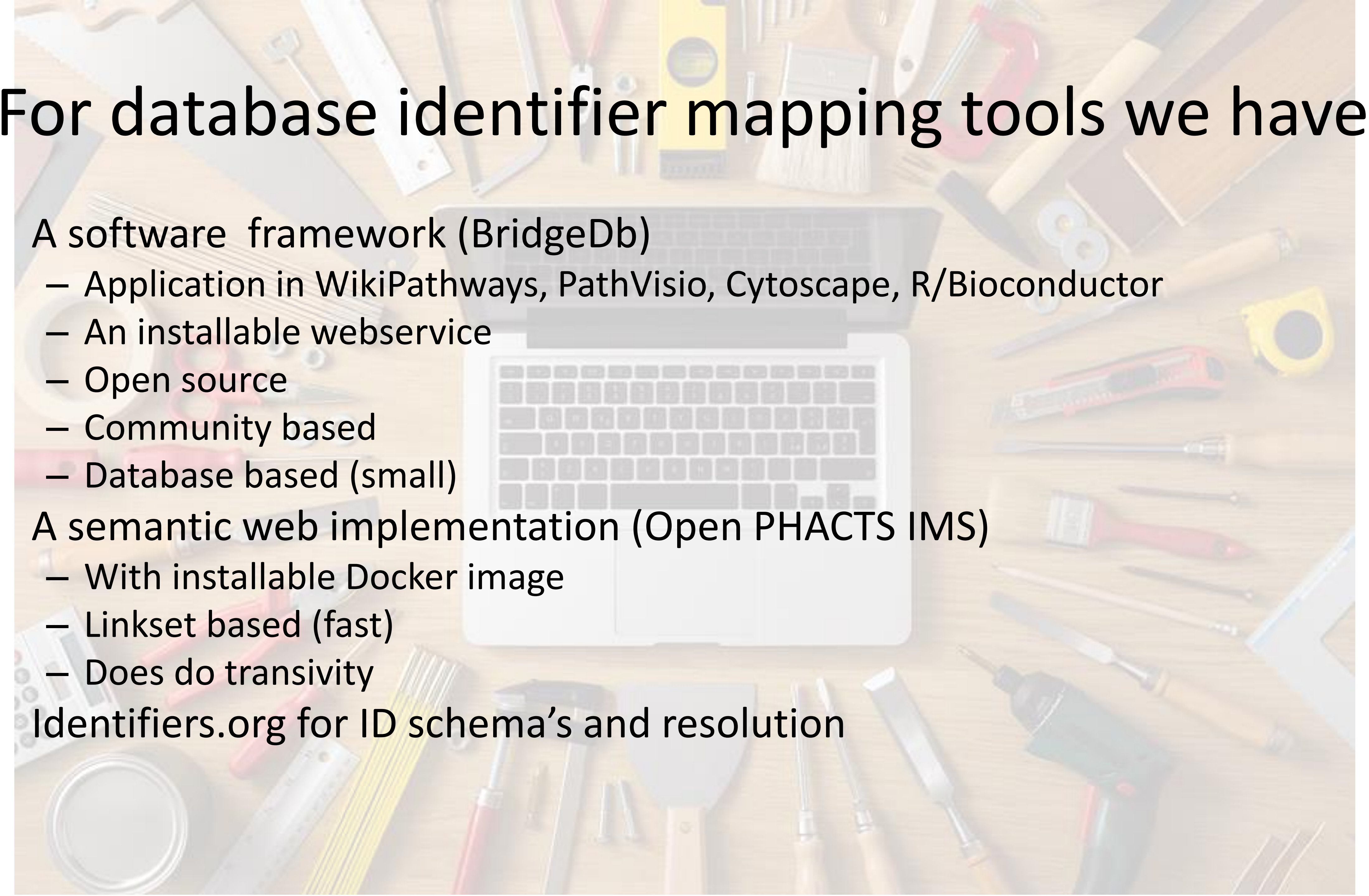
web service

The BridgeDb Framework: Standardized Access to Gene, Protein and Metabolite Identifier Mapping Services. Martijn P van Iersel, Alexander R Pico, Thomas Kelder, Jianjiong Gao, Isaac Ho, Kristina Hanspers, Bruce R Conklin, Chris T Evelo. BMC Bioinformatics 2010, 11: 5.

The magic is only in part in the tool

And mostly in the content

For database identifier mapping tools we have:



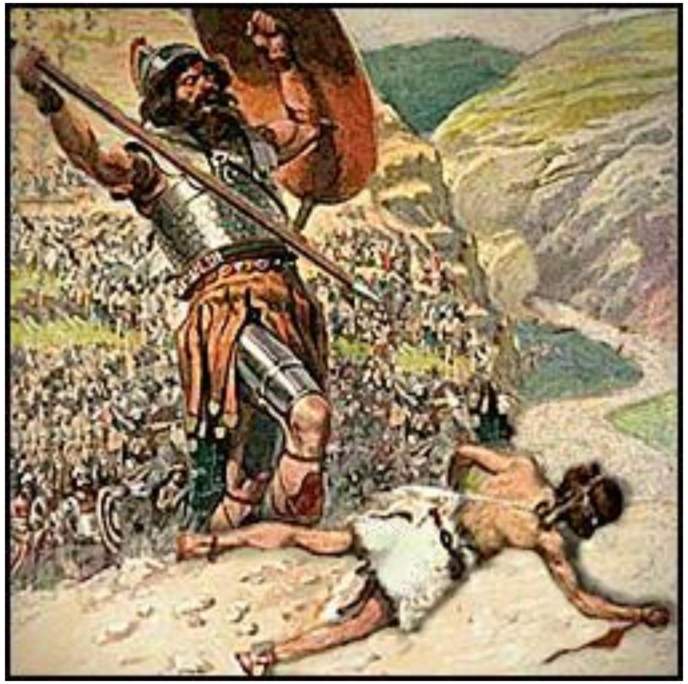
- A software framework (BridgeDb)
  - Application in WikiPathways, PathVisio, Cytoscape, R/Bioconductor
  - An installable webservice
  - Open source
  - Community based
  - Database based (small)
- A semantic web implementation (Open PHACTS IMS)
  - With installable Docker image
  - Linkset based (fast)
  - Does do transivity
- Identifiers.org for ID schema’s and resolution

For ID mappings we have:

- Gene product, ENSEMBL derived, databases for:
  - most health related species
  - Some bacteria
  - Some plants
- Metabolite database, HMDB & CheBi derived, for most "human" metabolites
- Reaction database from RHEA
- Linksets for all of these

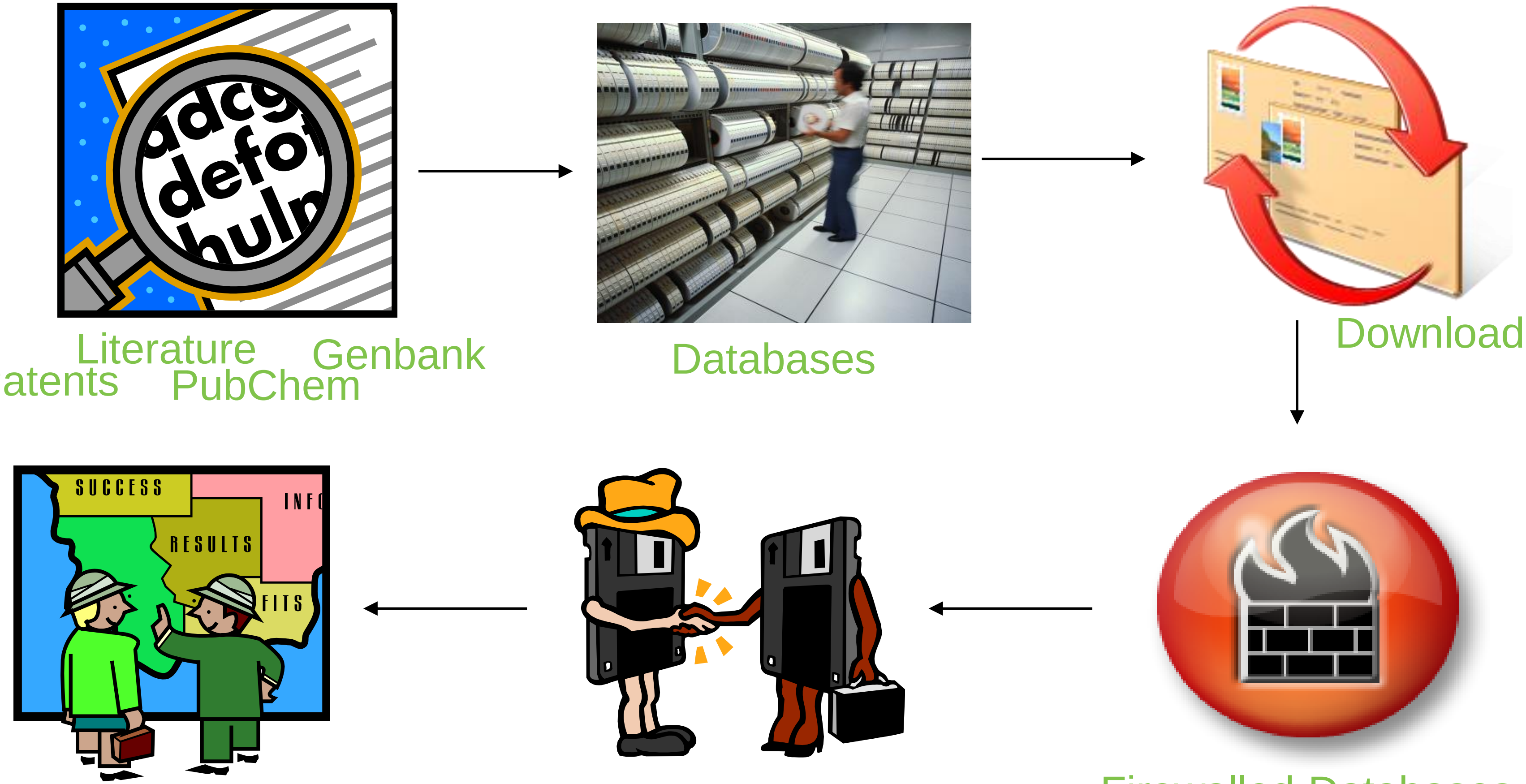
Other relevant:

- MyGene.info (Su lab)
  - also connected to WikiData
- DAVID (from Tigr)
  - recently updated (after years of inactivity)
  - often used (even in that period)
  - Should watch out for Goliath



SPACEX

How do R&D companies use public data?



Literature
Patents
Genbank
PubChem

Databases

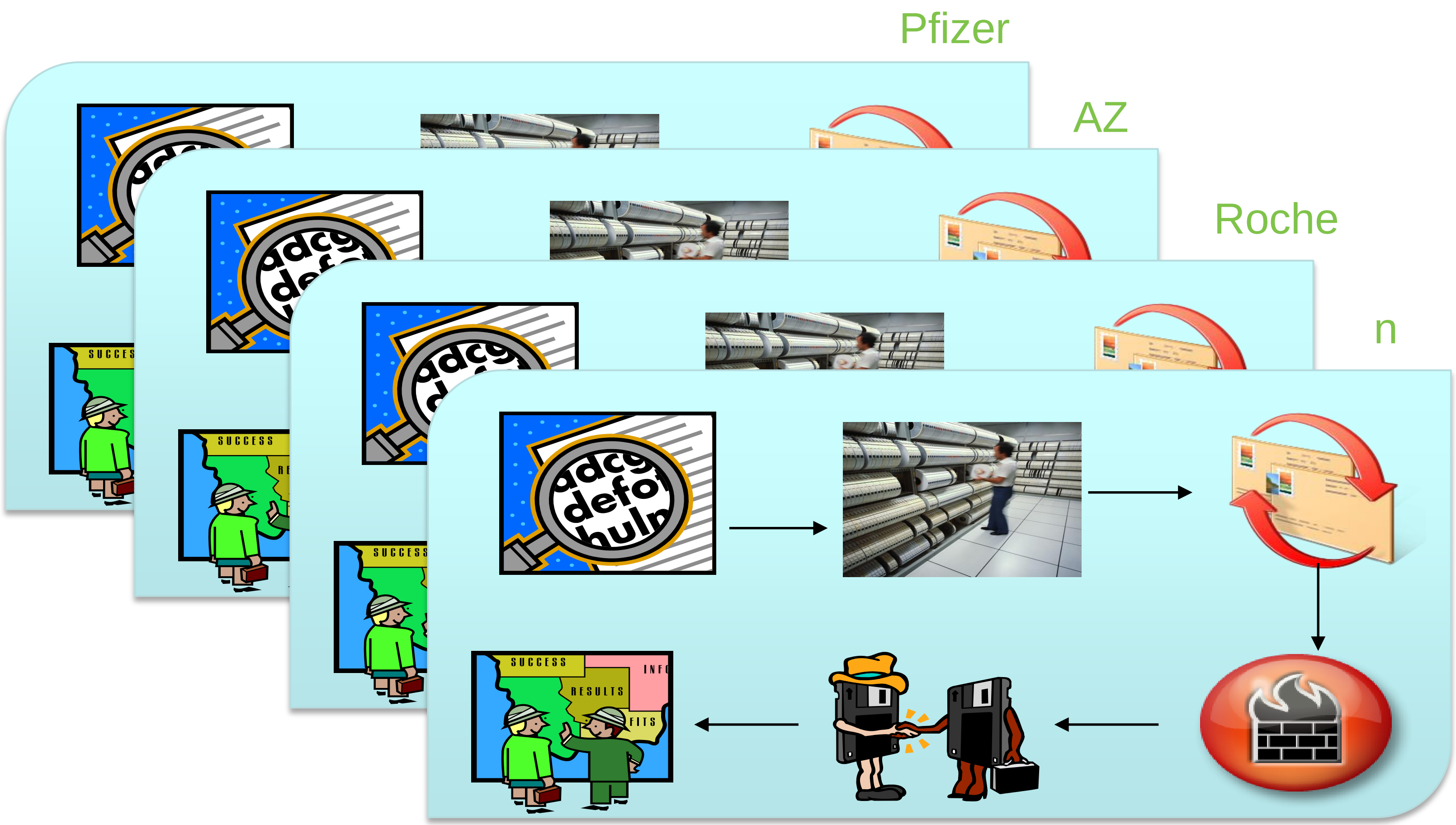
Downloads

Firewalled Databases

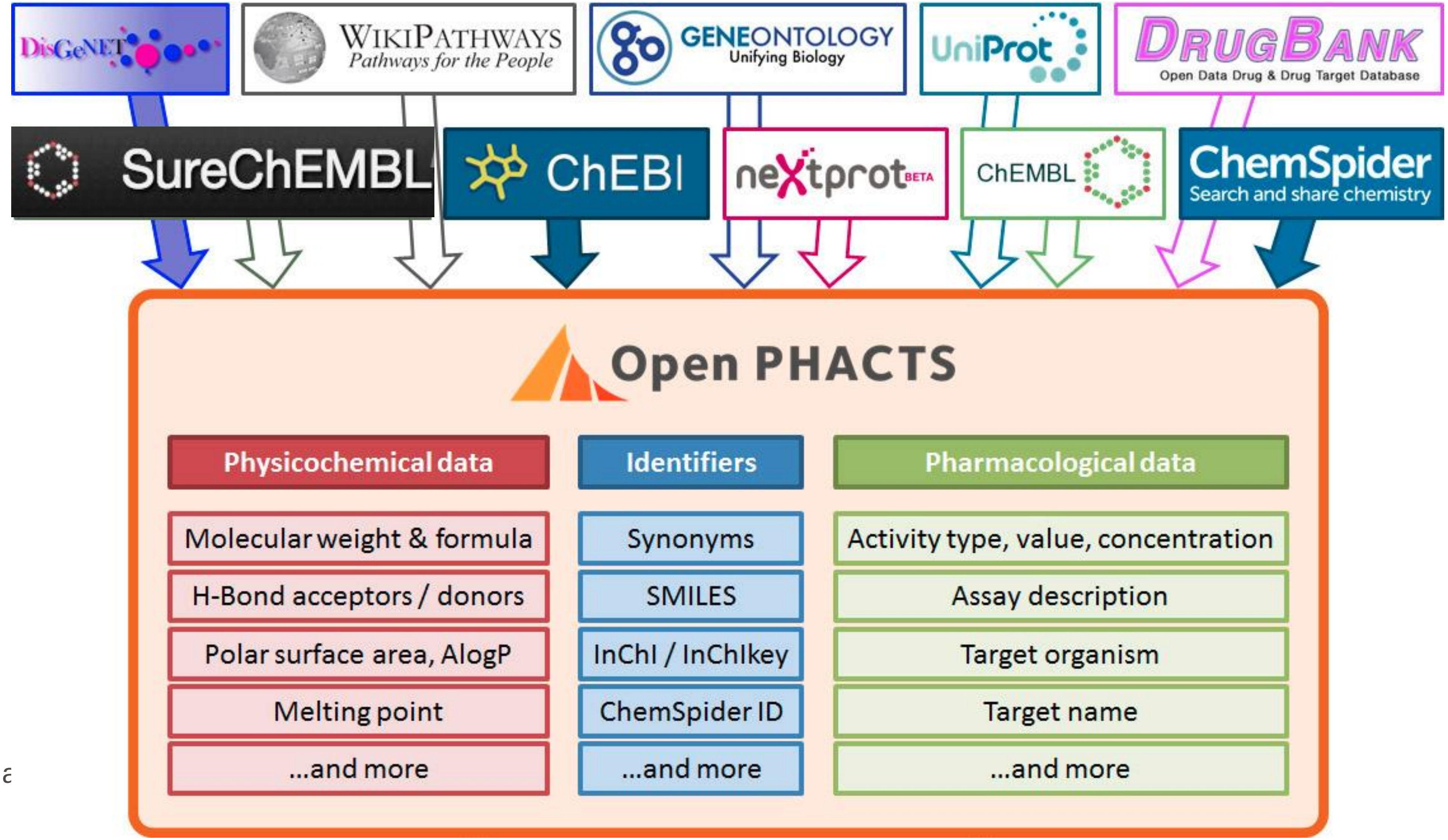
Data Integration

Data Analysis

How do pharma companies use public data?

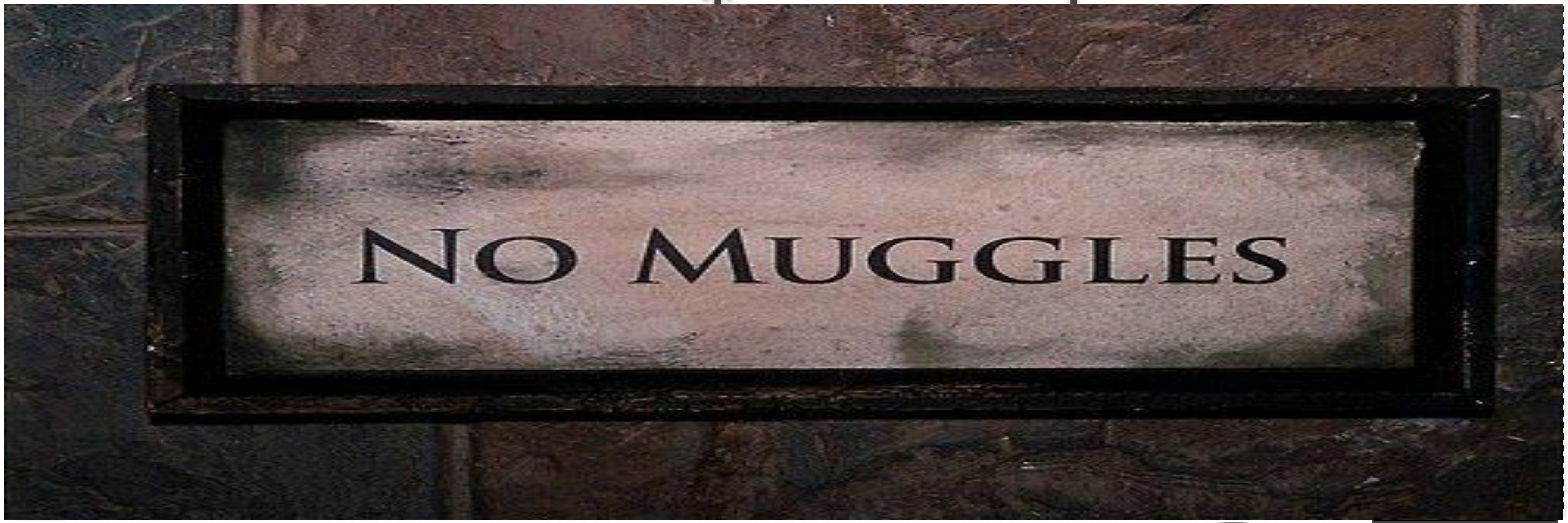


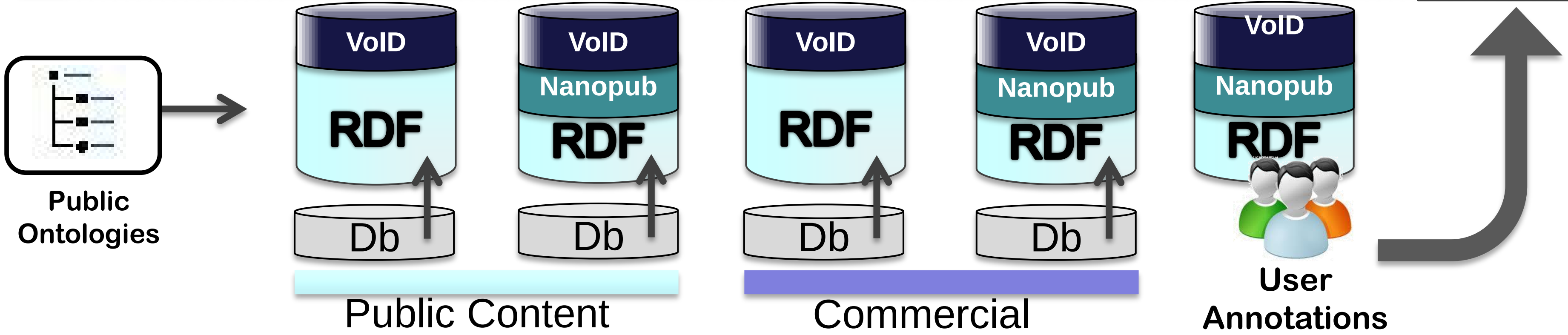
Open PHACTS
Open Pharmacological Space



Physicochemical data	Identifiers	Pharmacological data
Molecular weight & formula	Synonyms	Activity type, value, concentration
H-Bond acceptors / donors	SMILES	Assay description
Polar surface area, AlogP	InChI / InChIkey	Target organism
Melting point	ChemSpider ID	Target name
...and more	...and more	...and more

Apps




VoID
RDF
Db

VoID
Nanopub
RDF
Db

VoID
RDF
Db

VoID
Nanopub
RDF
Db

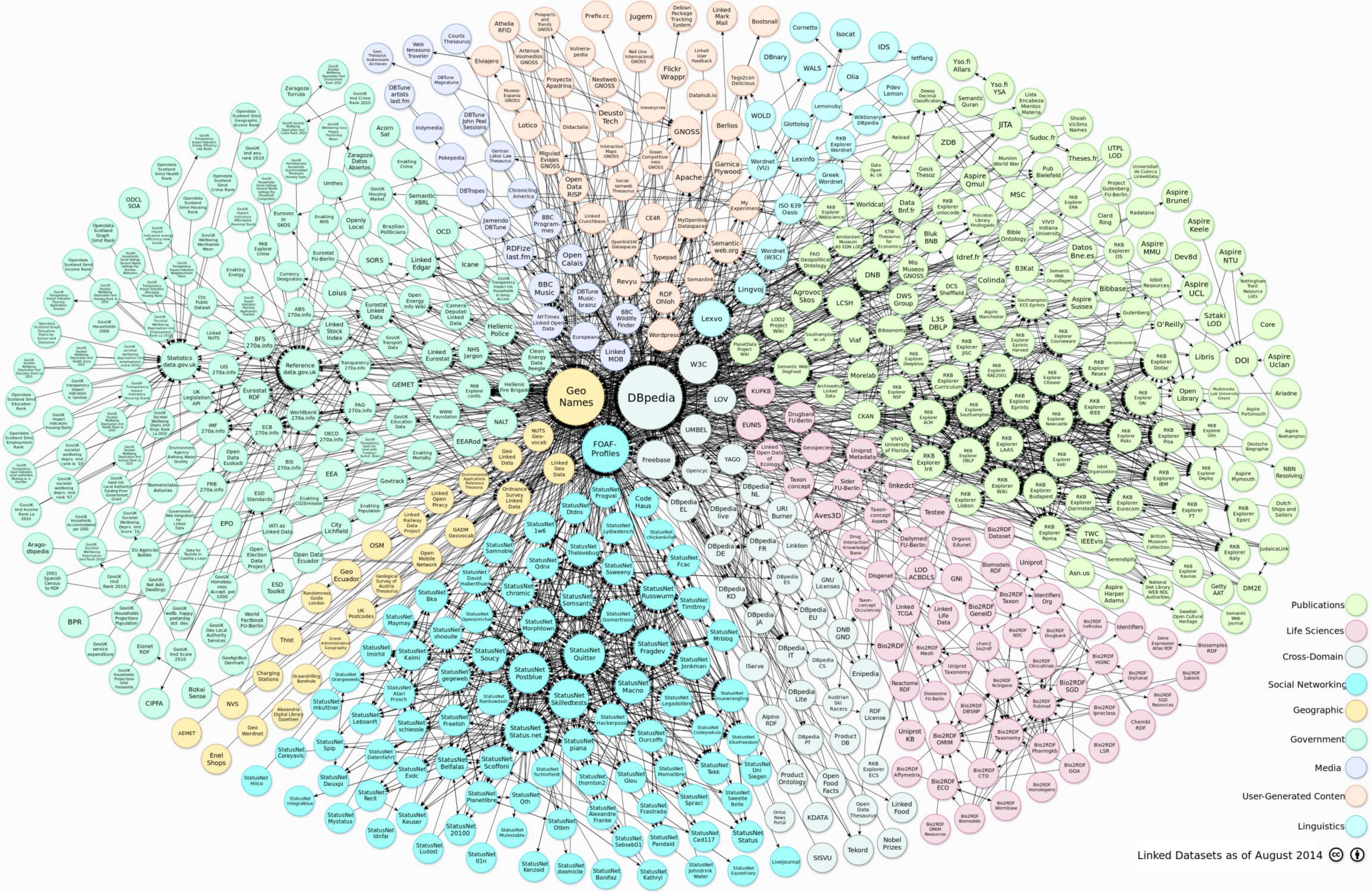
VoID
Nanopub
RDF

Public Ontologies

Public Content

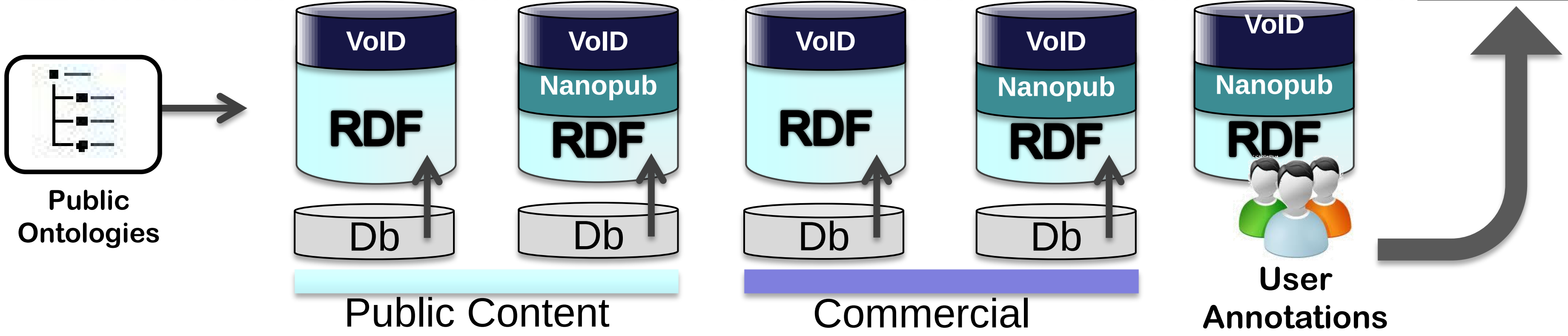
Commercial

User Annotations



Linked Datasets as of August 2014

Apps



Public Ontologies

VoID
RDF
Db

VoID
Nanopub
RDF
Db

VoID
RDF
Db

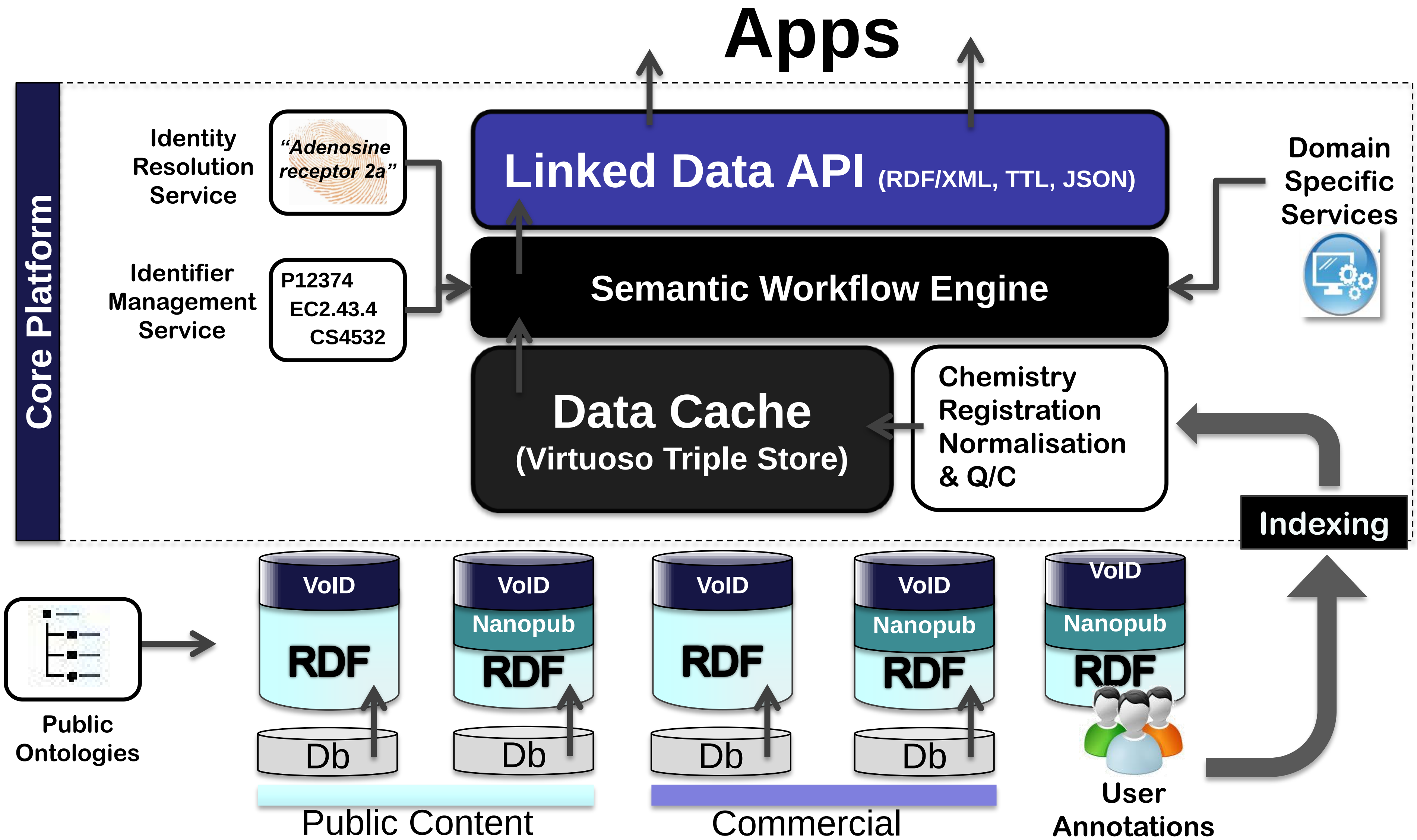
VoID
Nanopub
RDF
Db

VoID
Nanopub
RDF

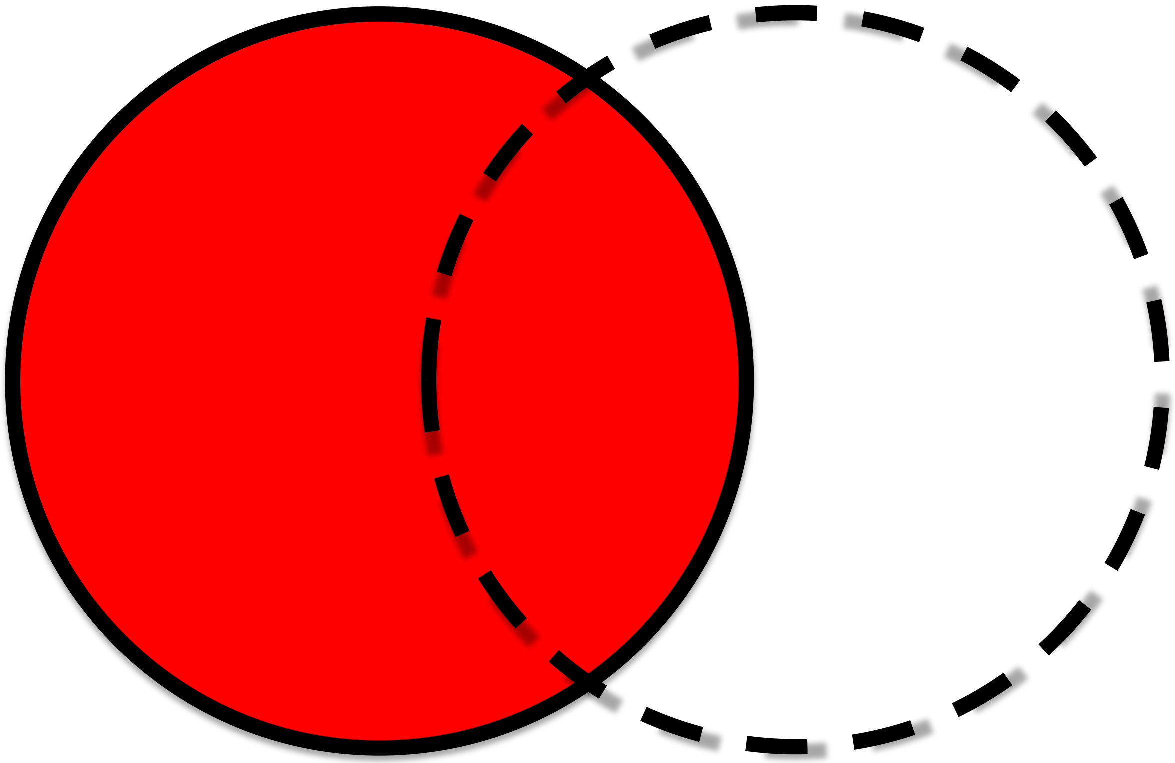
Public Content

Commercial

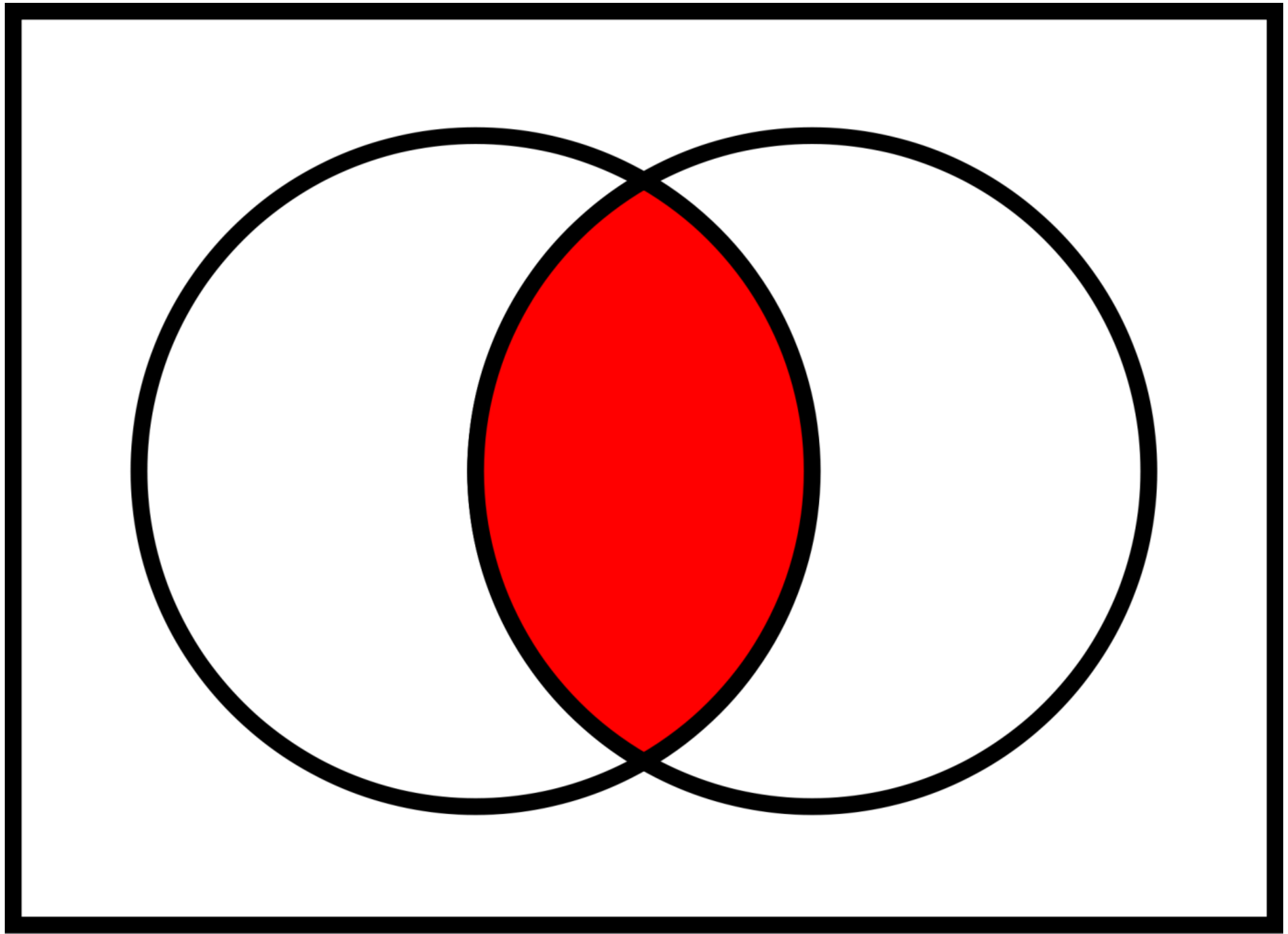
User Annotations


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
Public Content
Commercial
User Annotations

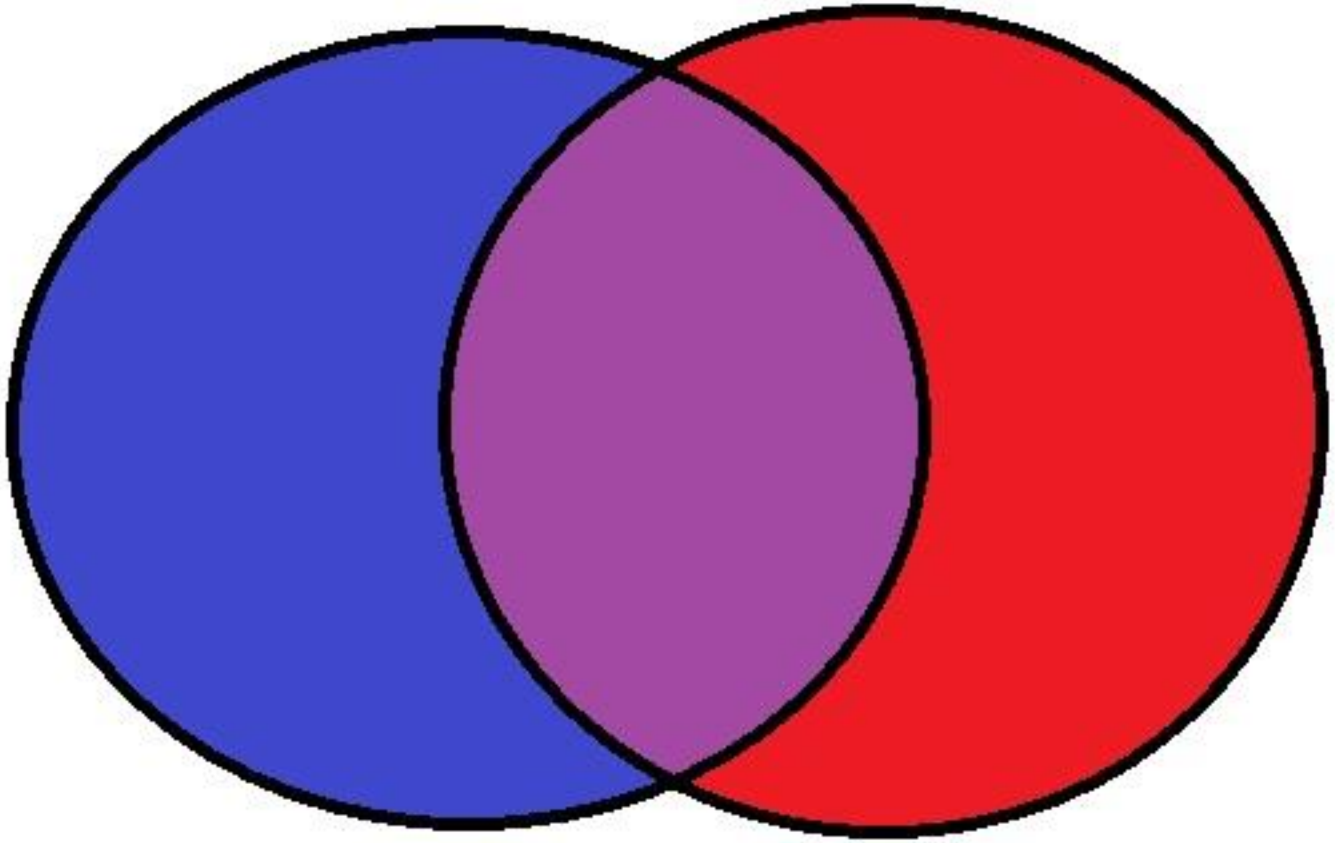
Choose a standard



Link one resource to another



Or use both and map



ID mapping in Open PHACTS:

- Semweb enabled (full URIs)
- Using linksets
- Transitivity (and limits for that)
  - gene -> protein -> has enzyme code
  - Protein -> has enzyme code -> other proteins

This is not just Open PHACTS

Federated SPARQL queries:

e.g. find all genes related to disease, then all pathways with these genes…

Used as hackaton (swat4ls) examples/

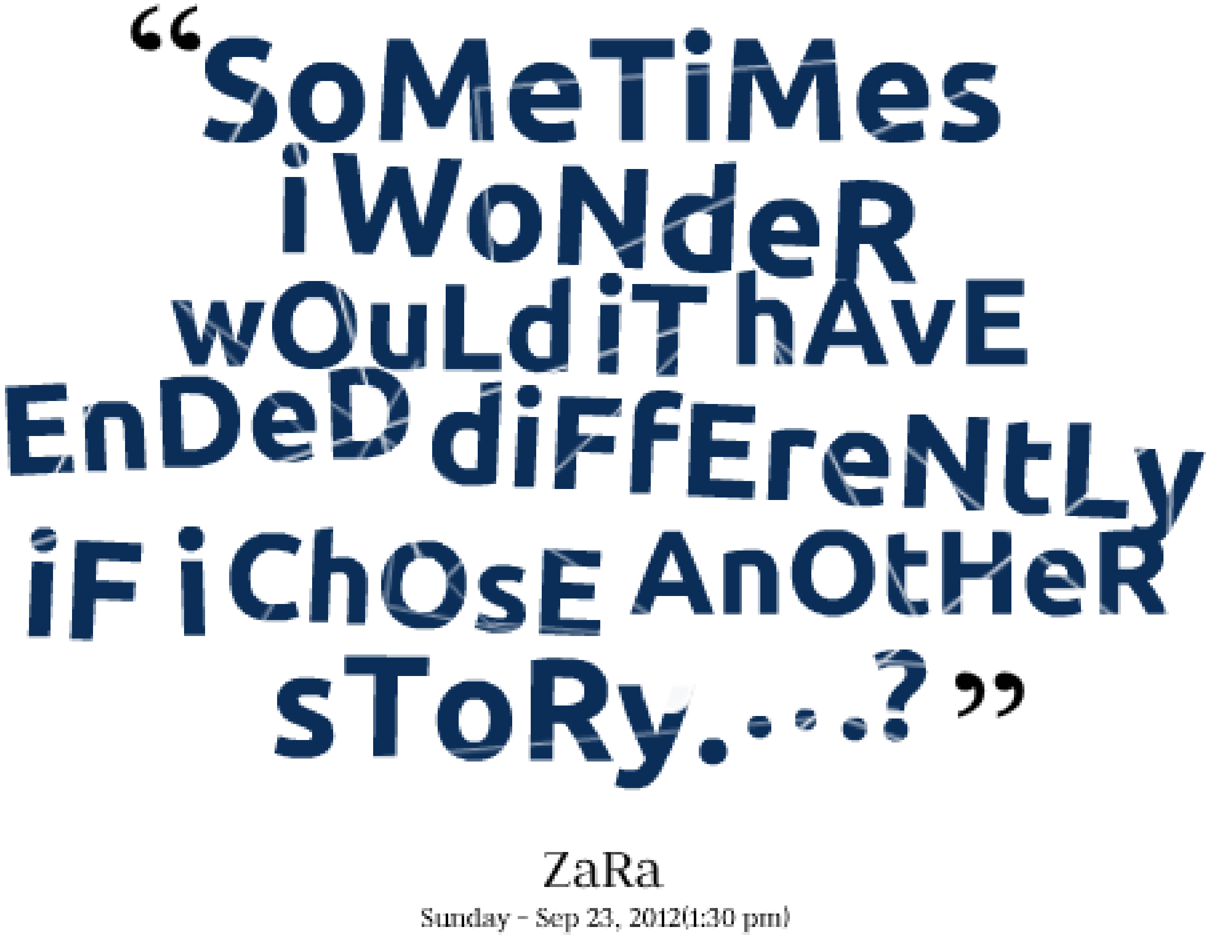
Only works sometimes, by chance

Needs integrated ID mapping!

Metabolite mapping needs

- More mappings! (plant products, drugs, xenobiotics)
- Ontology based mapping (CheBi)
- Because:
  - Palmitic acid is a fatty acid
  - R,R,R-tocopherol is a form of Vitamin E
- And these should (sometimes) map

Also applies to biology: scientific lenses



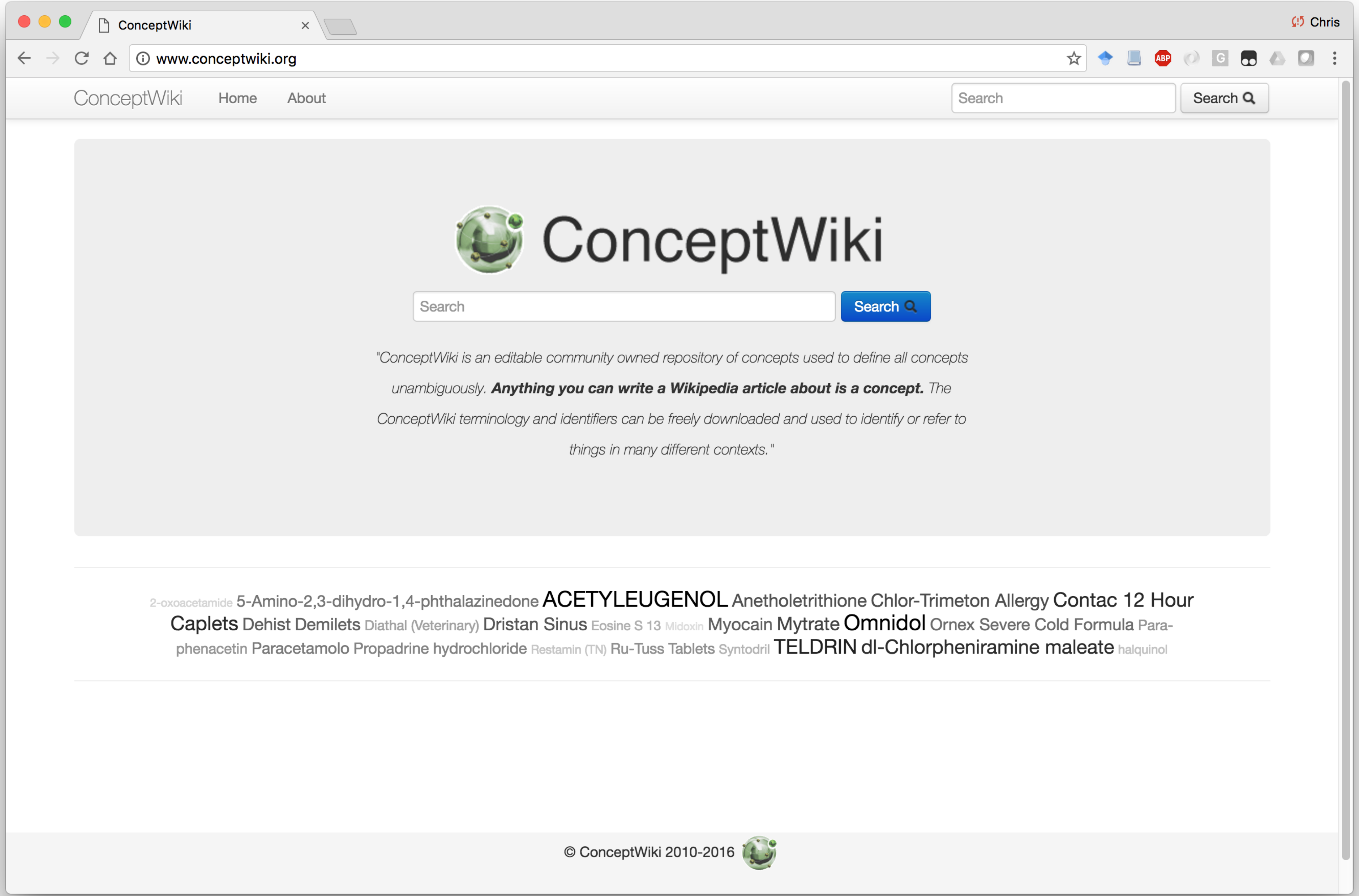
“SoMeTiMes iWoNdeR wOuLd iT hAvE EnDeD diFFEreNtLy iF i ChOsE AnOtHeR sToRy....?”

ZaRa

Sunday - Sep 23, 2012(1:30 pm)

Concepts

- What we need is a big ARTA (also referred to as, synonyms) table
  - That allows typos..
- E.g. synonyms for things found by text mining
- Mapping to ontologies is only a partial solution



"ConceptWiki is an editable community owned repository of concepts used to define all concepts unambiguously. ***Anything you can write a Wikipedia article about is a concept.*** *The ConceptWiki terminology and identifiers can be freely downloaded and used to identify or refer to things in many different contexts."*

2-oxoacetamide 5-Amino-2,3-dihydro-1,4-phthalazinedone ACETYLEUGENOL Anetholetrithione Chlor-Trimeton Allergy Contac 12 Hour Caplets Dehist Demilets Diathal (Veterinary) Dristan Sinus Eosine S 13 Midoxin Myocain Mytrate Omnidol Ornex Severe Cold Formula Para-phenacetin Paracetamolo Propadrine hydrochloride Restamin (TN) Ru-Tuss Tablets Syntodril TELDRIN dl-Chlorpheniramine maleate halquinol

© ConceptWiki 2010-2016

It failed!

Because …

- Try it!
- Calling on a million minds for community annotation …
- But… people work on wikis because it gives them something right away

And we still need it!

- Open PHACTS used an indexer
- These only know what is in there
- Doesn’t for instance allow connecting text mining results

Don’t be afraid to reinvent wheels!

Chemistry mapping

- Structure not ID based
- Allow substructure searches
- Open PHACTS open source ???
- We need it, may have to redo

Ontology mapping

- Many available, even as services
- Often part of met resources
- Meta resources need feedback to original!

Annotation tools use ontology terms

- But do not allow lookup
- Often (typically) do not show what ontology used
- Need frameworks

Special cases

- Proteomics: peptides/fragments to proteins (Dutch proteomics meeting)
- Sequences: typical short sequence resolution (why BLAST the same things over and over?) (Natasa Przulj at NetBio SIG)

Mapping tools are core tools: need funding and sustainability

Delivering

- Not what people want
- But what people need
- For what they want

- Where we will run into the next problem:

If we curate data well, and make all the technical connections real it may still be incomparable. Needs…. brains

People think that mathematics is complicated. Mathematics is the simple bit, it's the stuff we CAN understand. It's cats that are complicated.

— John Horton Conway —

Those who don't believe in magic will never find it.

Roald Dahl in “The Minpins”