Integration of the scientific literature into the Semantic Web: Facts from biomedical data resources

Dietrich Rebholz-Schuhmann - rebholz@ebi.ac.uk

Samuel Croset – croset@ebi.ac.uk

Christoph Grabmüller - grabmuel@ebi.ac.uk

December 8th, 2011

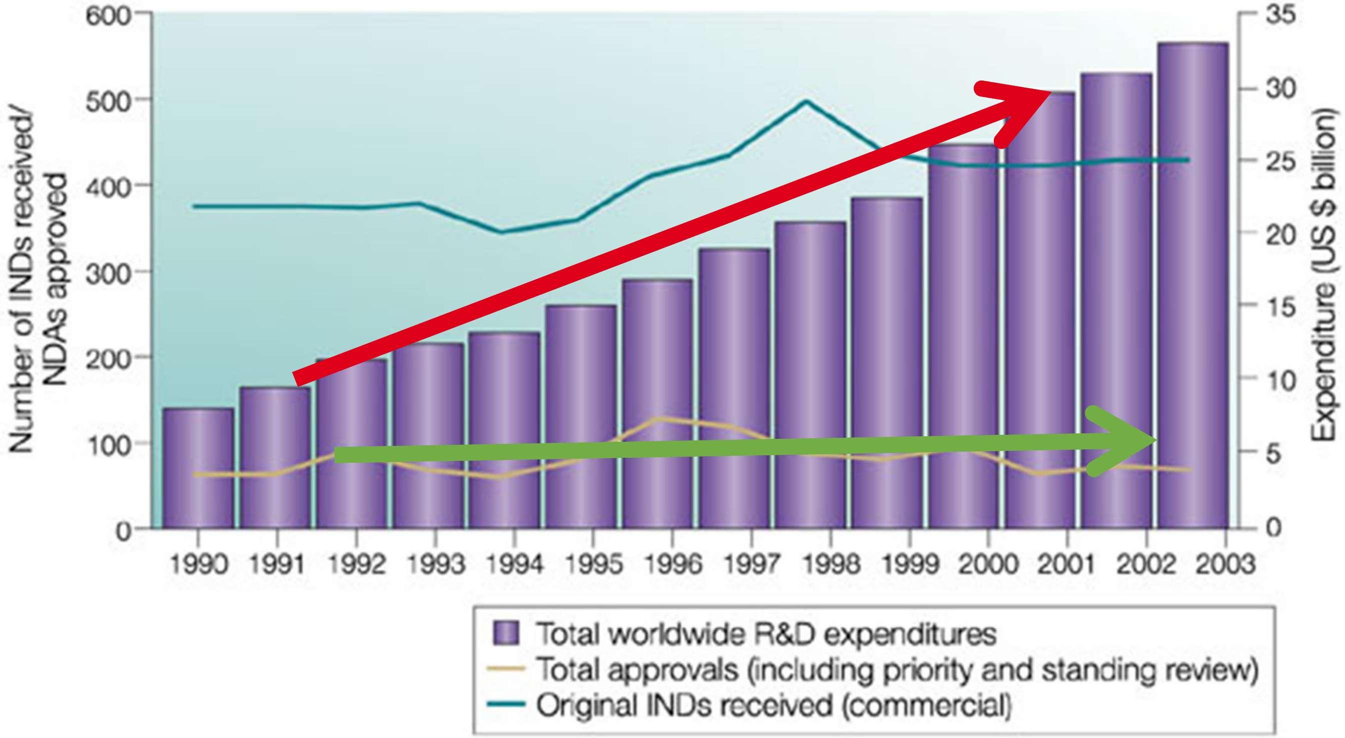
Objectives of the tutorial:

- How is Semantic Web applied to Biology?
- What is required to implement the Semantic Web?
- How does scientific literature fit into the Semantic Web?
- How to put raw text into RDF?
- How to query the linked data resources?

Outline

- Why the Semantic Web in Biology?
- What have we done? SESL Prototype
    - Data in RDF
    - Integration
- What are the outcomes?
- What next?

A productivity problem



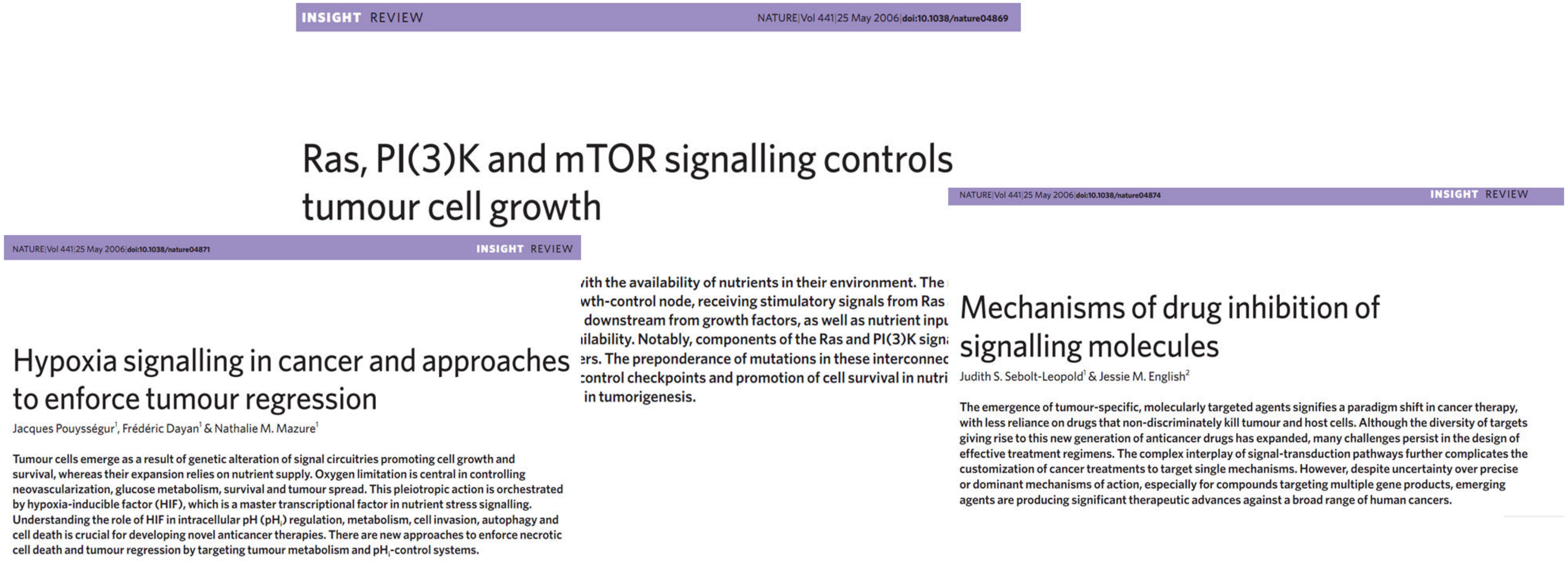
Nature Reviews | Drug Discovery

DRUG REPOSITIONING:
IDENTIFYING AND DEVELOPING
NEW USES FOR EXISTING DRUGS

Ted T. Ashburn and Karl B. Thor

Diseases mechanisms are complex

- Cancer, Alzheimer, Diabetes, Ageing, etc…
- Different types of entities: Molecules, proteins, genes, cell types, phenotype, environment, etc…



INSIGHT REVIEW NATURE|Vol 441|25 May 2006|doi:10.1038/nature04869

Ras, PI(3)K and mTOR signalling controls tumour cell growth

vith the availability of nutrients in their environment. The
wth-control node, receiving stimulatory signals from Ras
downstream from growth factors, as well as nutrient inpu
ilability. Notably, components of the Ras and PI(3)K signa
ers. The preponderance of mutations in these interconnec
control checkpoints and promotion of cell survival in nutri
in tumorigenesis.

NATURE|Vol 441|25 May 2006|doi:10.1038/nature04871 INSIGHT REVIEW

Hypoxia signalling in cancer and approaches to enforce tumour regression

Jacques Pouysségur[1], Frédéric Dayan[1] & Nathalie M. Mazure[1]

Tumour cells emerge as a result of genetic alteration of signal circuitries promoting cell growth and survival, whereas their expansion relies on nutrient supply. Oxygen limitation is central in controlling neovascularization, glucose metabolism, survival and tumour spread. This pleiotropic action is orchestrated by hypoxia-inducible factor (HIF), which is a master transcriptional factor in nutrient stress signalling. Understanding the role of HIF in intracellular pH (pH_i) regulation, metabolism, cell invasion, autophagy and cell death is crucial for developing novel anticancer therapies. There are new approaches to enforce necrotic cell death and tumour regression by targeting tumour metabolism and pH_i-control systems.

NATURE|Vol 441|25 May 2006|doi:10.1038/nature04874 INSIGHT REVIEW

Mechanisms of drug inhibition of signalling molecules

Judith S. Sebolt-Leopold[1] & Jessie M. English[2]

The emergence of tumour-specific, molecularly targeted agents signifies a paradigm shift in cancer therapy, with less reliance on drugs that non-discriminately kill tumour and host cells. Although the diversity of targets giving rise to this new generation of anticancer drugs has expanded, many challenges persist in the design of effective treatment regimens. The complex interplay of signal-transduction pathways further complicates the customization of cancer treatments to target single mechanisms. However, despite uncertainty over precise or dominant mechanisms of action, especially for compounds targeting multiple gene products, emerging agents are producing significant therapeutic advances against a broad range of human cancers.

Understanding and treating diseases

Understanding and treating diseases

Pertea et al. BMC Bioinformatics 2011, 12:274
http://www.biomedcentral.com/1471-2105/12/274

RESEARCH ARTICLE

BMC Bioinformatics

Open Access

Detection of lineage-specific evolutionary changes among primate species

Mihaela Pertea, Geo M Pertea and Steven L Salzberg

Abstract

LETTER

doi:10.1038/nature10592

The unusual gamma-ray burst GRB 101225A explained as a minor body falling onto a neutron star

1 DECEMBER 2011 | VOL 480 | NATURE | 69

©2011 Macmillan Publishers Limited. All rights reserved

OPEN ACCESS Freely available online

Synopsis

People Gather Only the Information They Need to Make Decisions

Janelle Weaver

November 2011 | Volume 9 | Issue 11 | e1001204

www.plosbiology.org

RESEARCH ARTICLES

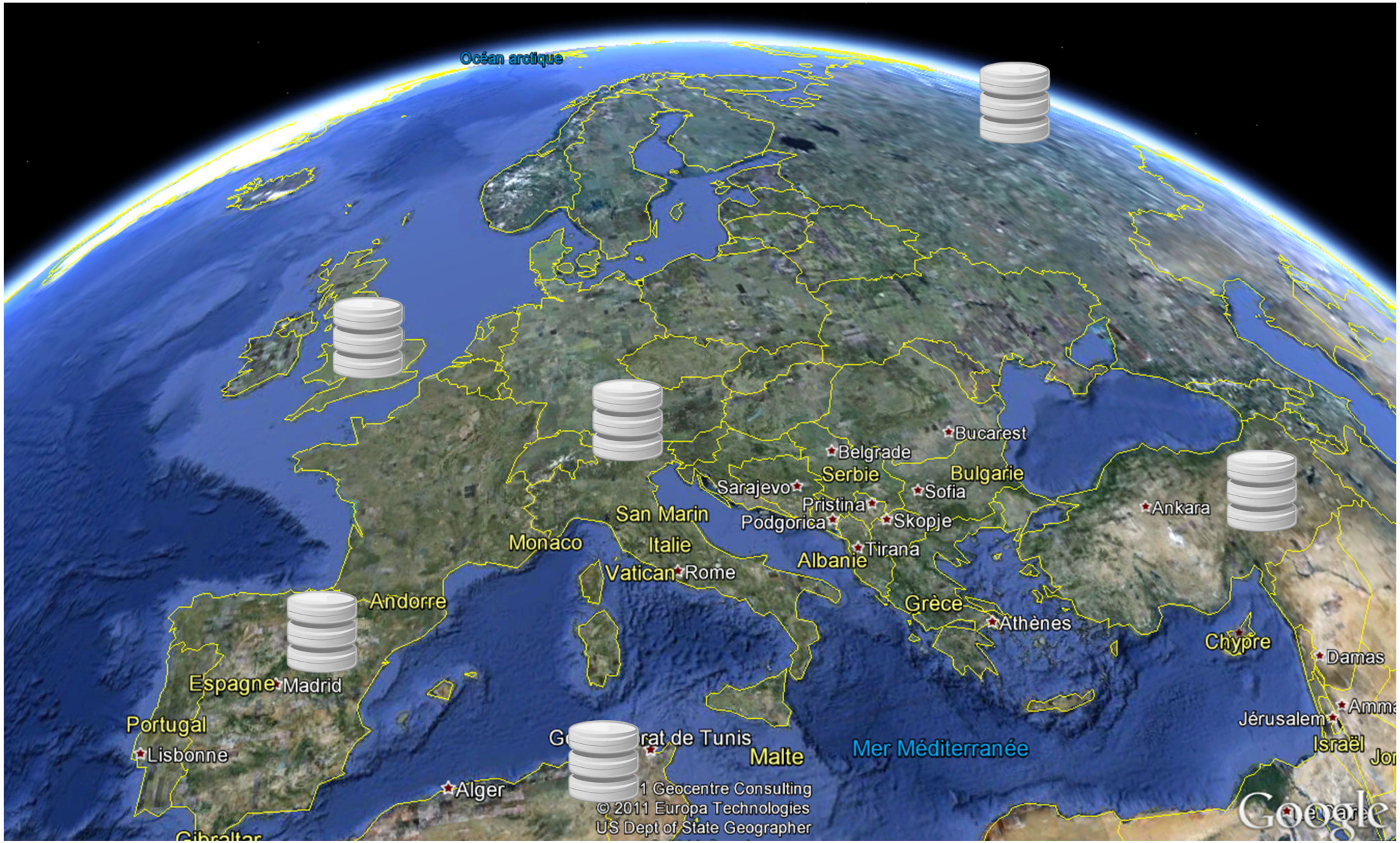
The Cambrian Conundrum: Early Divergence and Later Ecological Success in the Early History of Animals

Douglas H. Erwin, Marc Laflamme, Sarah M. Tweedt, Erik A. Sperling, Davide Pisani, Kevin J. Peterson

SCIENCE

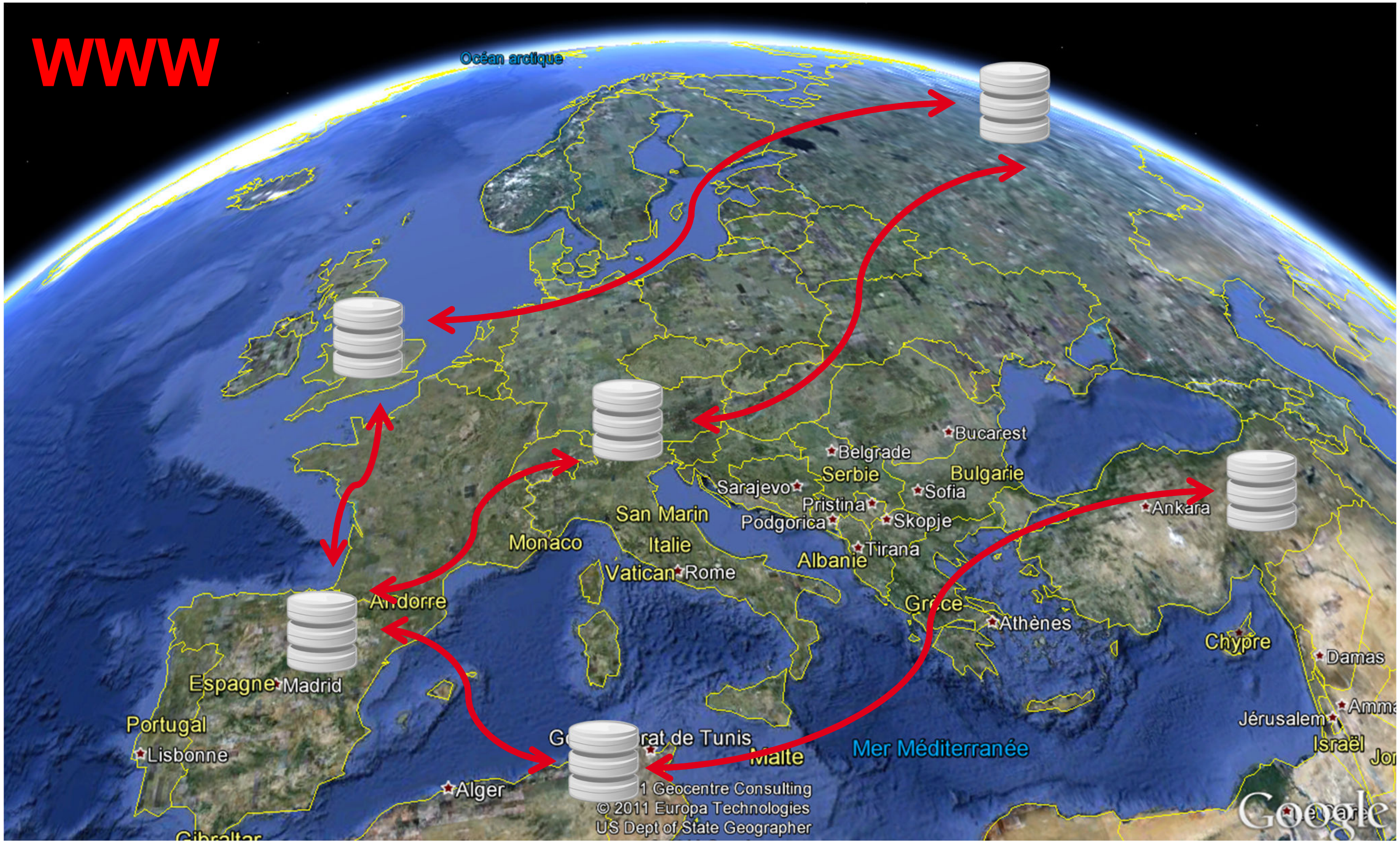
www.sciencemag.org

The data is far away

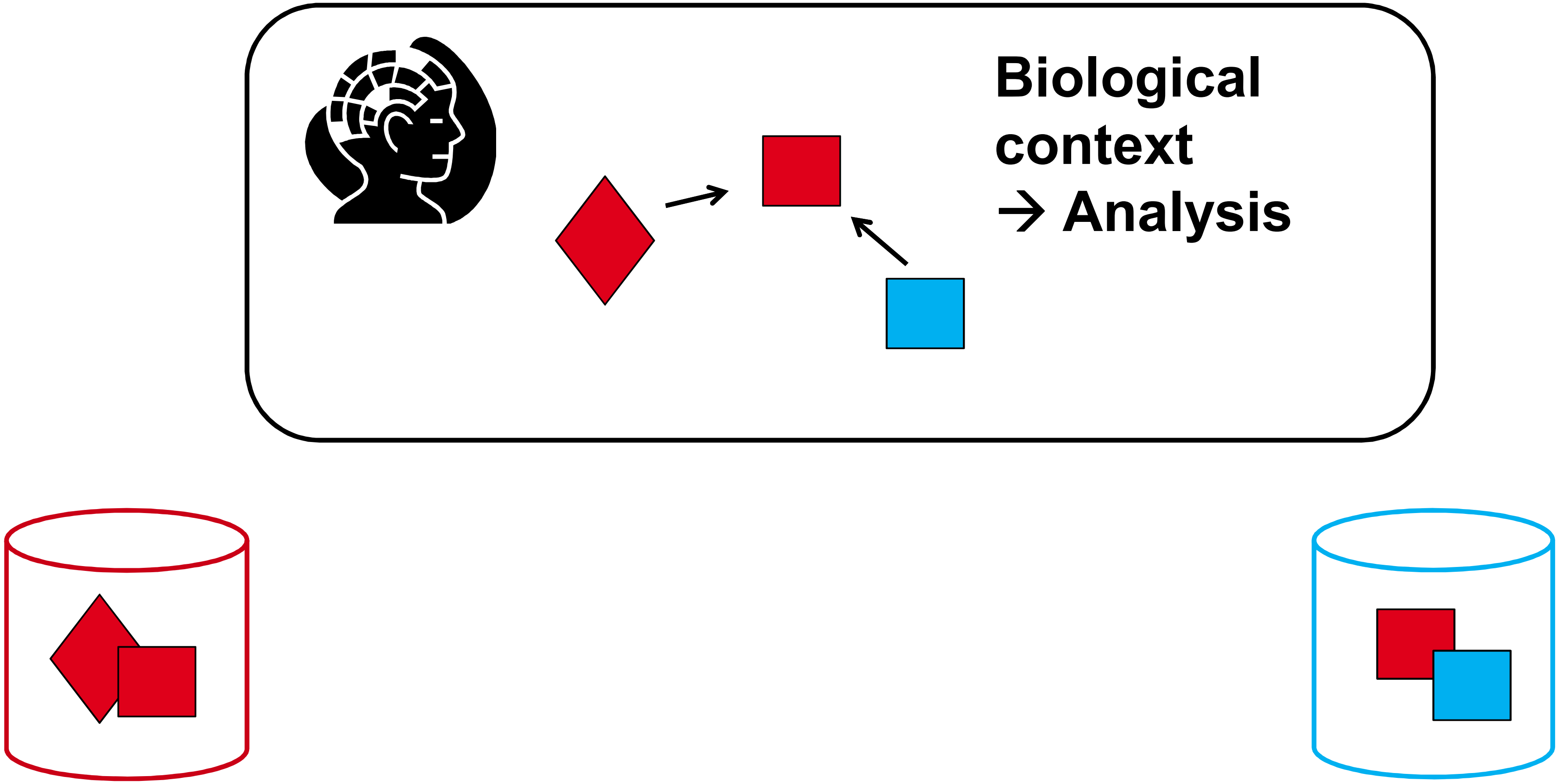


EMBL-EBI

The data is NOT far away



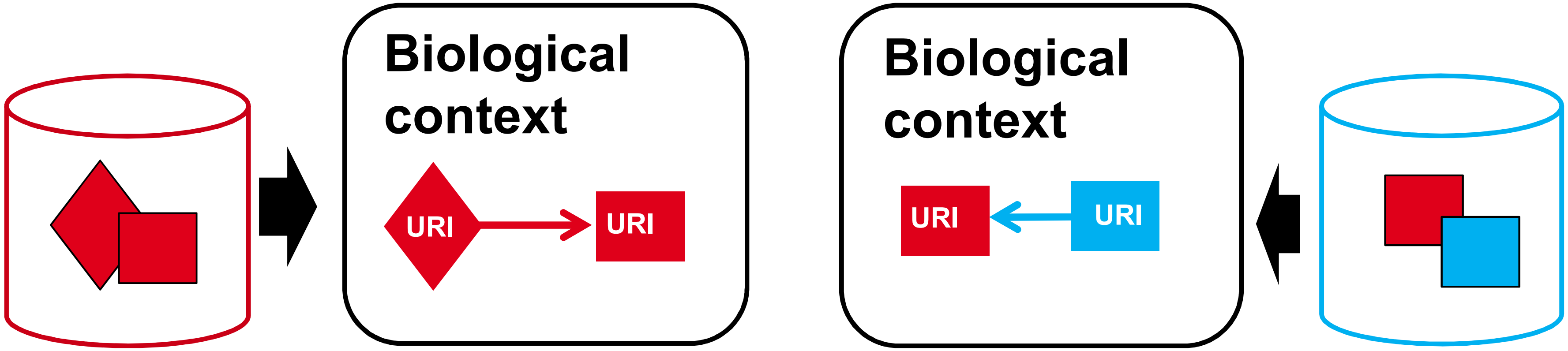
Facts re-use in Biology – Traditional Means



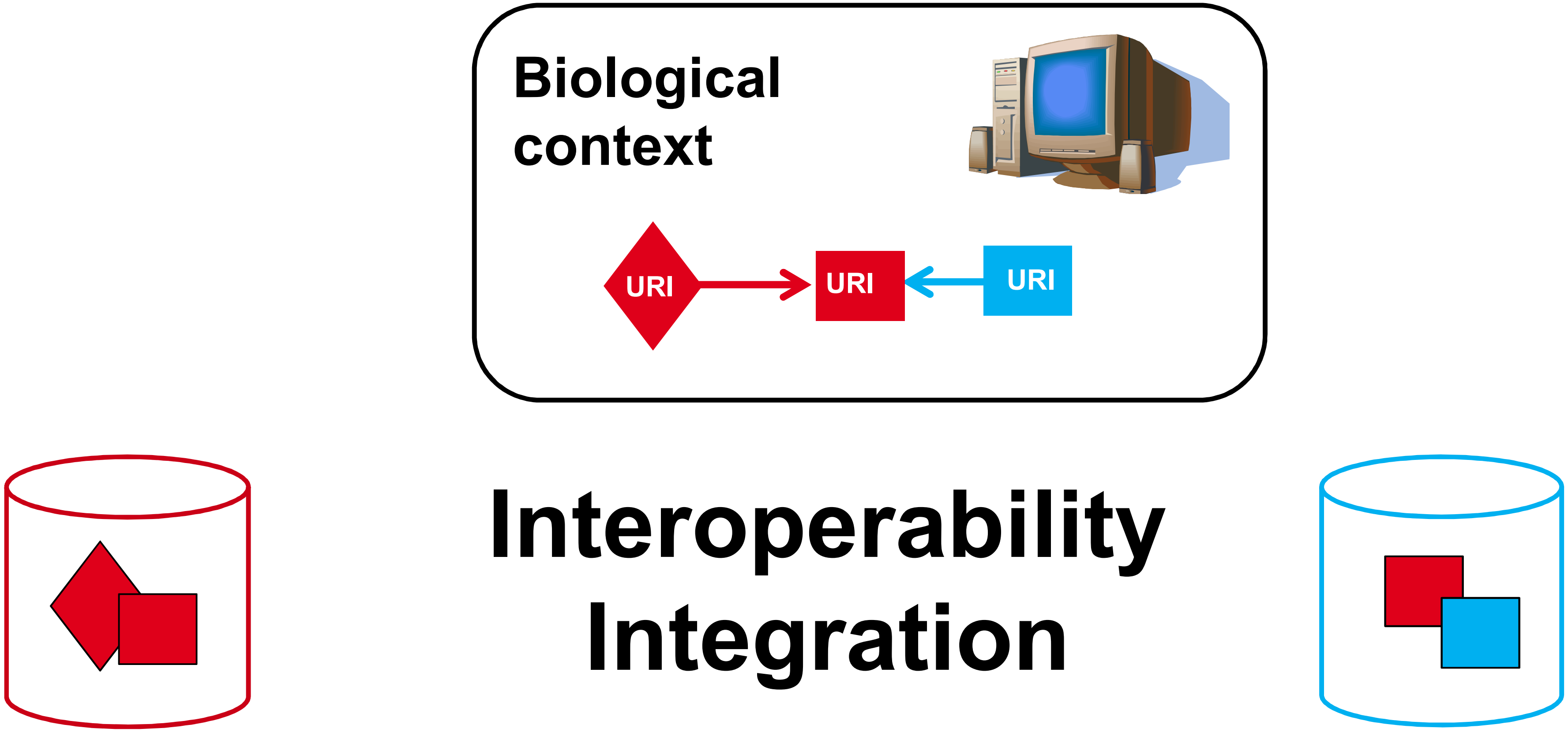
Biological context
→ Analysis

Facts re-use in Biology – Semantic Web

RDF



Facts re-use in Biology – Semantic Web



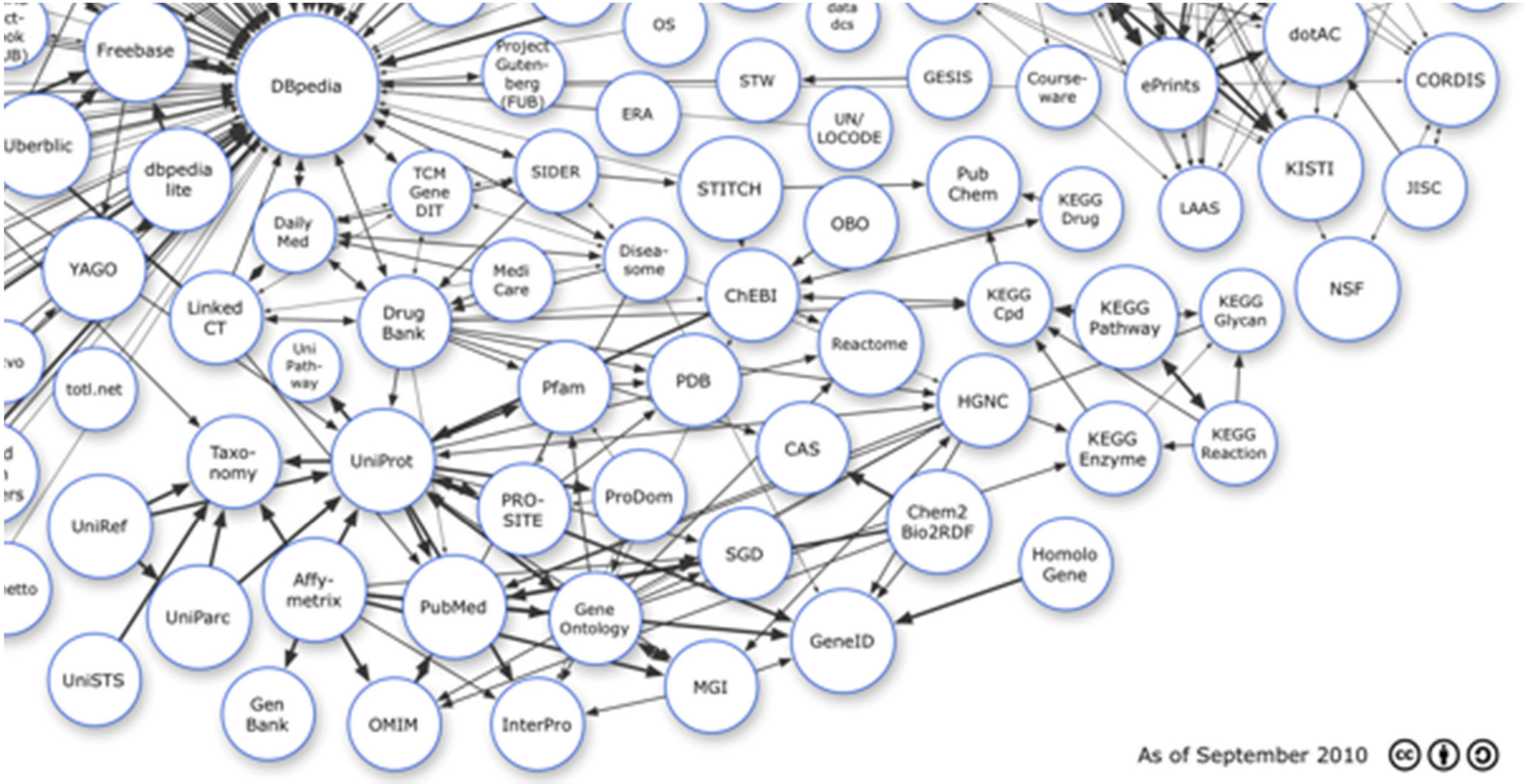
Biological context

Interoperability
Integration

Linked Data Principles (by Tim Berners-Lee)

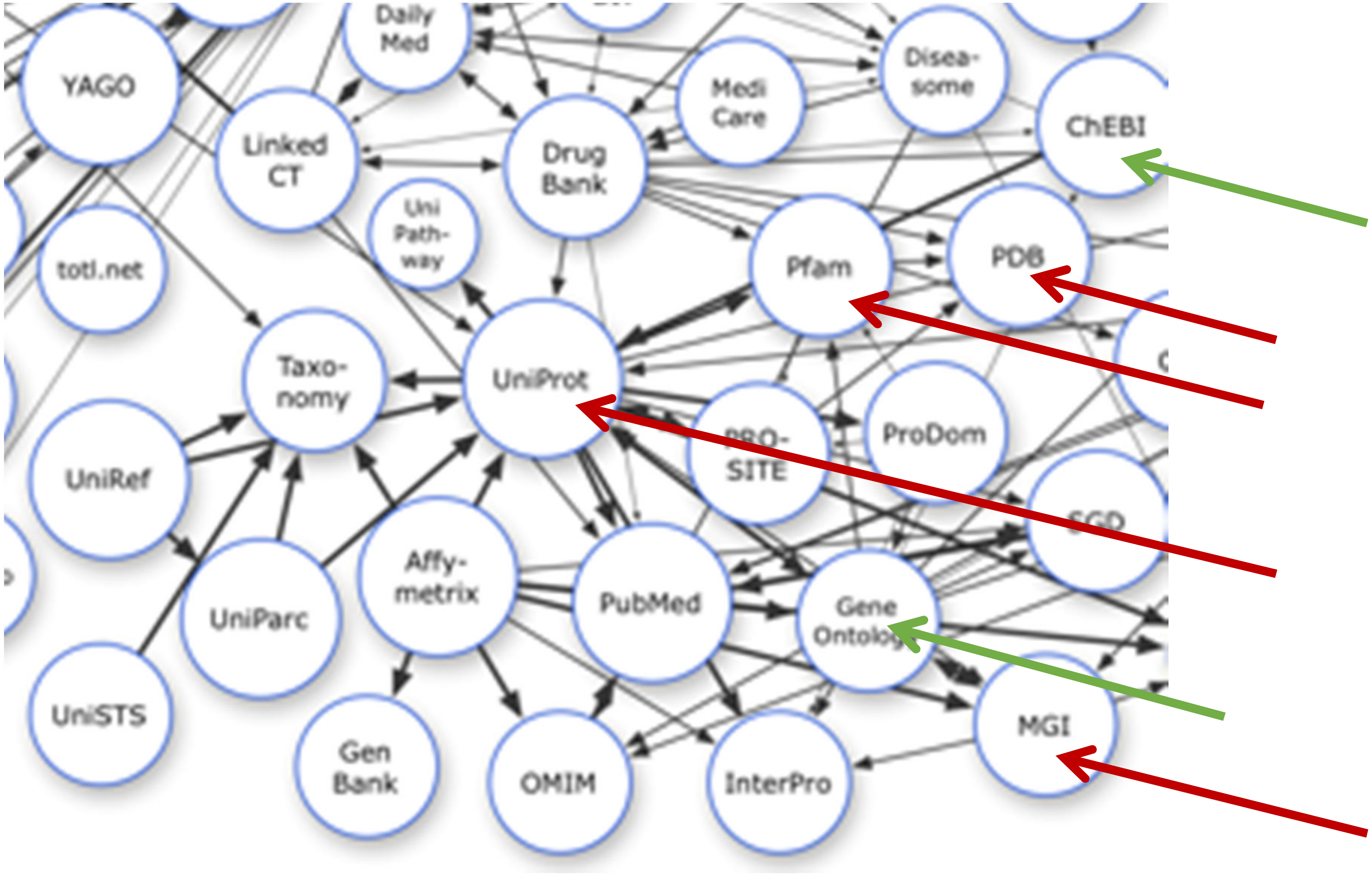
- (P1): use of universal resource identifiers (URIs) to label things or entities
  - e.g.\ for a protein or chemical entity, but also for a database entry, or a patient record or the identification of the patient itself
- (P2): names have to be reachable by their web address (“http://URIs”)
- (P3): the names should lead to useful information, which is given in representation standards (RDF, SPARQL).
- (P4): the links to other URIs should be provided for further discovery

Linked Data, the biological, chemical part

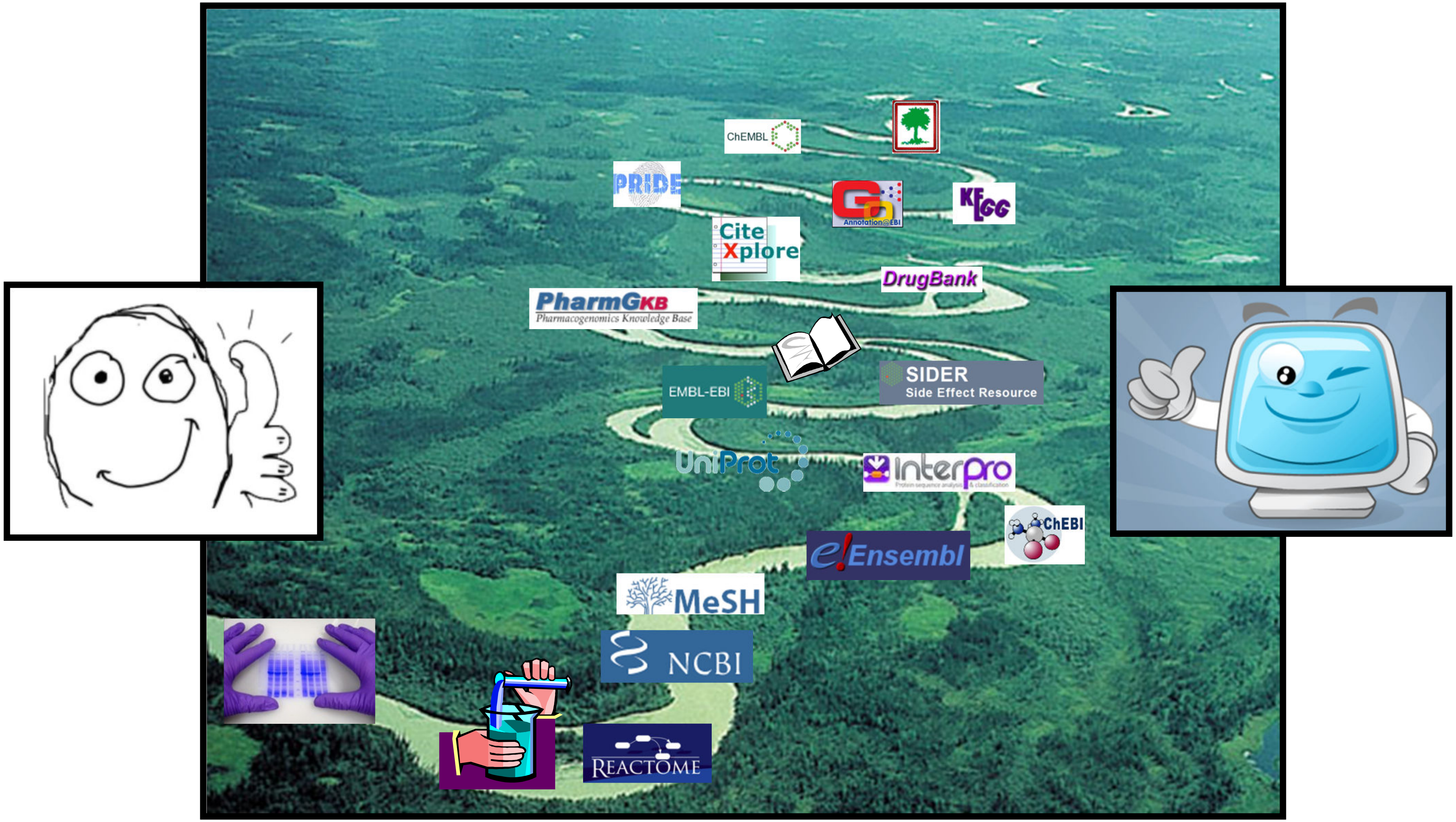


As of September 2010

Different kinds of Data in the Linked Data Cloud



Interoperability and Logic



Source Metaphore: Allemang and Hendler

Semantic Web and Biology

Explicit Structures and definitions (via URIs):

- **Biological Semantics Integration**: Proteins, Genes, Organisms, etc…
- **Format Standards**: RDF
- **Public and accessible**: Web

Semantic Web is just a **method** in Biology.

Use it to answer a **biomedical question**!

What have we done? The SESL project

Problematic:

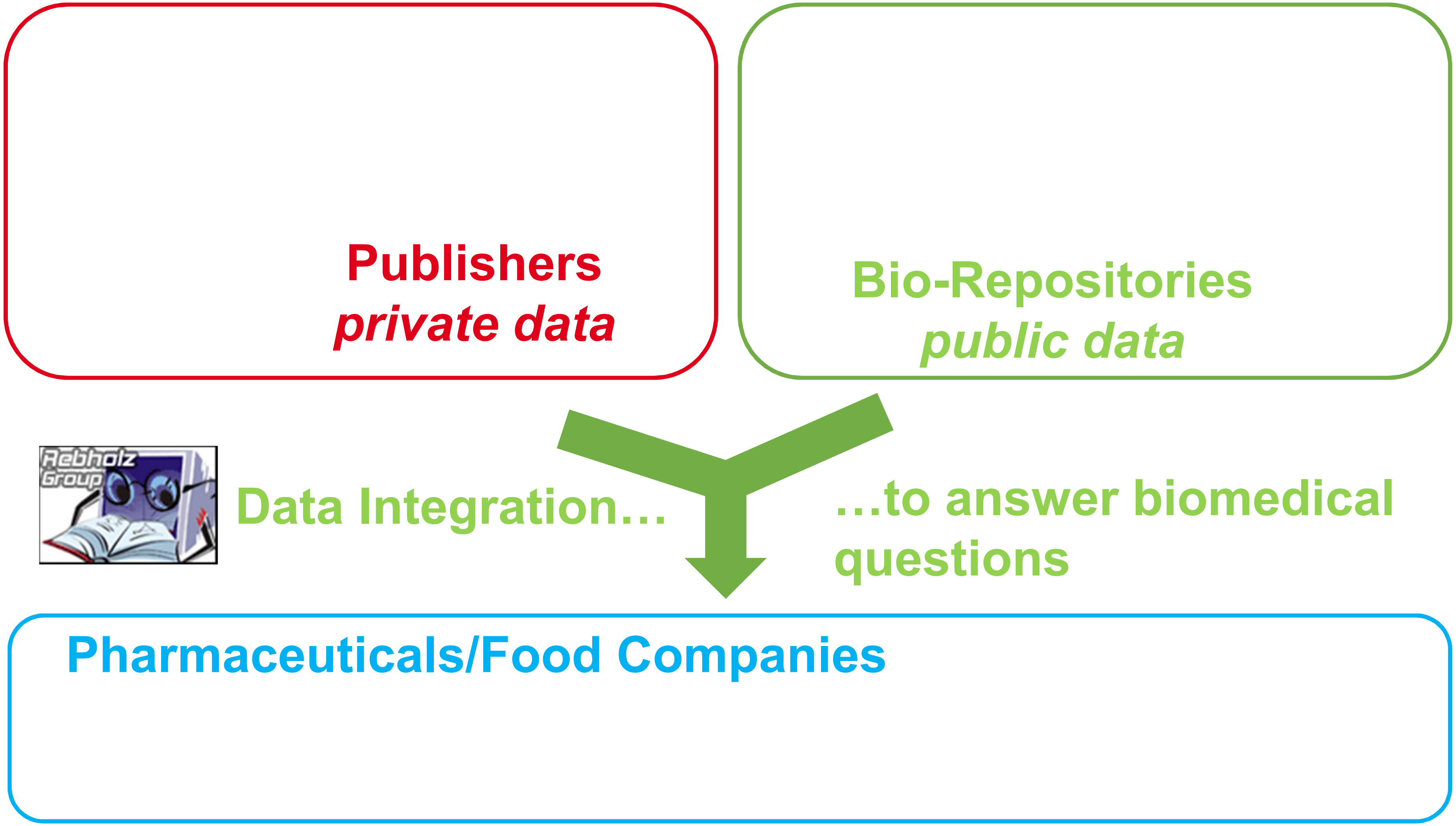
What evidence is available for gene-disease relations?

- What causes of a disease do we know?
- How does the gene/protein function?
- Which process is linked to the gene/protein?
- What hidden knowledge can we produce?

What are the questions that we want to answer through data integration?

- How does the gene/protein function?
  Which process is linked to the gene/protein?
  - Lookup in UniProtKb / BenBank
  - BUT ALSO: use the data from the literature
  - AND ALSO: use indirect data, i.e. protein activities in ChEMBL
  - Integrate: UniProtKb, literature and ChEMBL
- What causes of a disease do we know
  - Lookup in OMIM, MGI, possibly UniProtKb
  - Lookup in all data resources at the same time
  - Find the function / process / phenotype / expression levels that is shared between a gene and a disease
  - Integrate OMIM, MGI, UniProtKb, ArrayExpress / GeneAtlas, possibly GWAS databases, Decipher, …

What have we done? The SESL project



Publishers
private data

Bio-Repositories
public data

Data Integration…

…to answer biomedical questions

Pharmaceuticals/Food Companies

What have we done? The SESL project

RS•C
ROYAL SOCIETY OF CHEMISTRY

npg
nature publishing group

ELSEVIER

OXFORD
UNIVERSITY PRESS

Publishers
private data

ARRAYEXPRESS

OMIM
Online Mendelian Inheritance in Man

WIKIPEDIA

UniProt

Bio-Repositories
public data

Data Integration…

…to answer biomedical questions

Pharmaceuticals/Food Companies

gsk GlaxoSmithKline

AstraZeneca

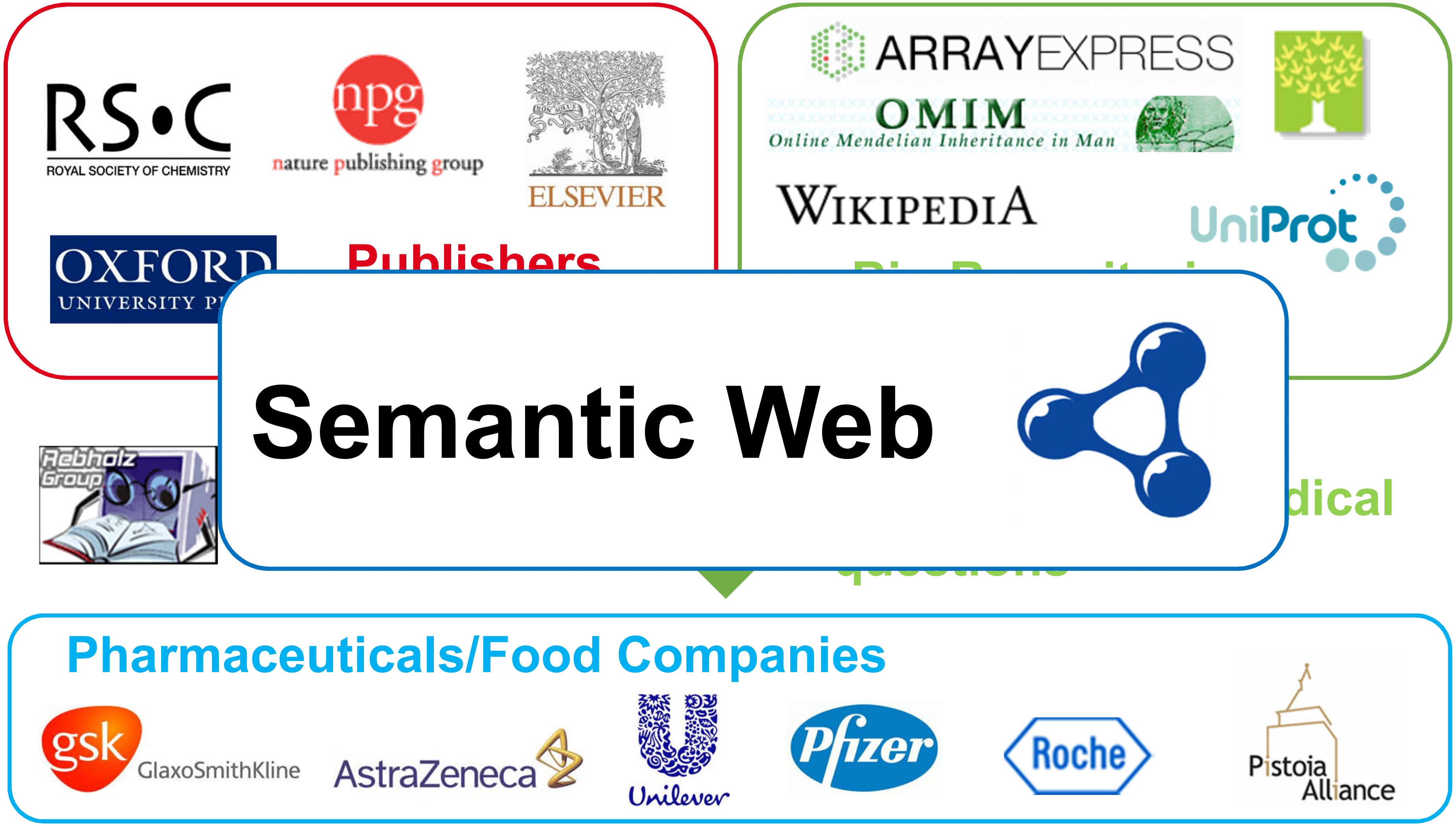
Unilever

Pfizer

Roche

Pistoia
Alliance

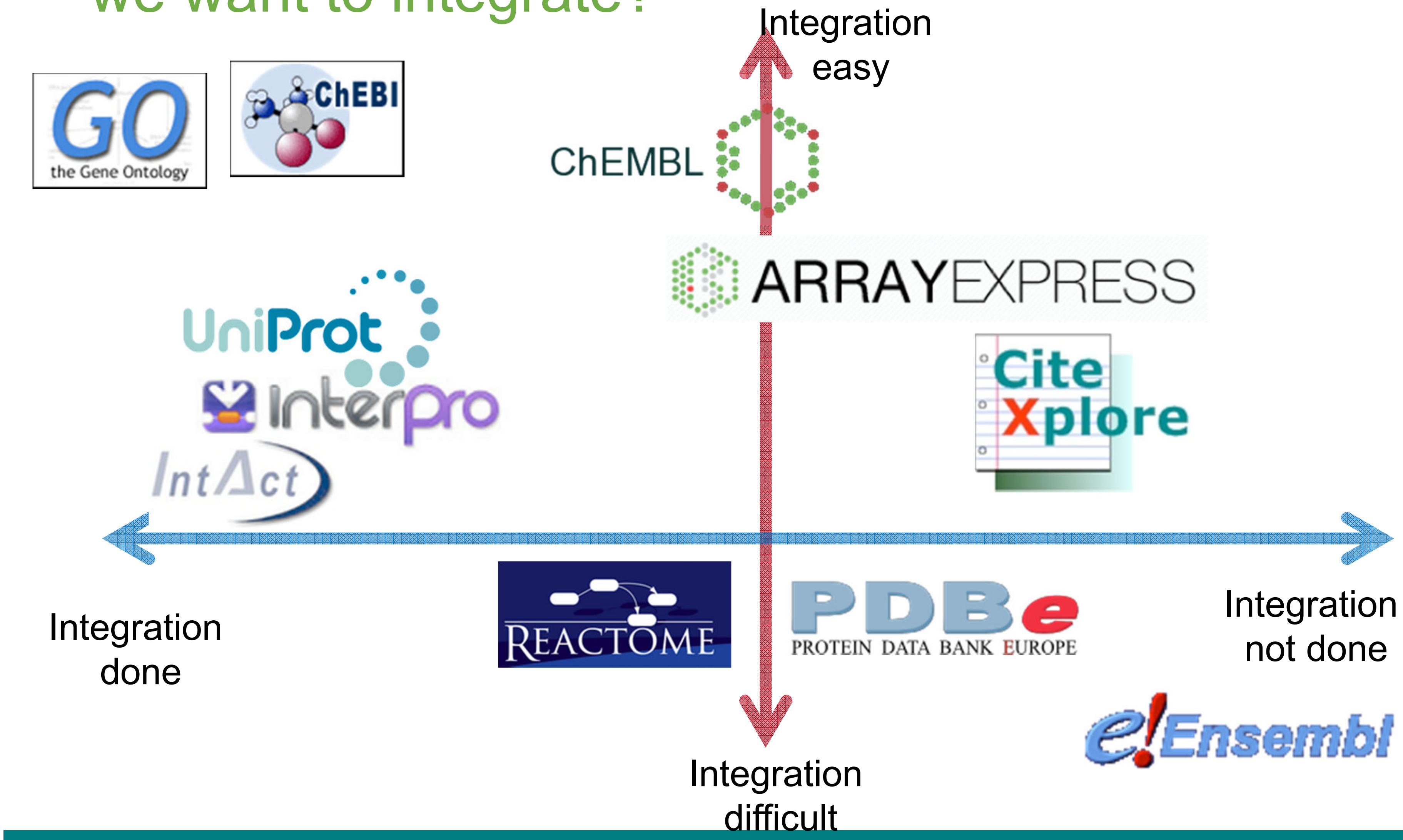
What have we done? The SESL project



How do we approach the data integration

- Gather the data and/or gather the access to the data from the data resource (possibly anywhere in the Web)
- Work out the relations between the entities concepts
    - Within the data resource: explicite & implite links
    - Across data resources: again explicite & implite links
- *Build the ontologies to do the data integration, use the ontologies as the data schema*
- Query across the SPARQL endpoints
- *Use reasoning across the data resources*
    - *Consistency analysis: intra- and inter-database analysis*
    - *Inference of unseen evidence across data resources*

What are the bioinformatics data resources that we want to integrate?



Integration easy

Integration done

Integration not done

Integration difficult

Data in RDF – How to build a house?

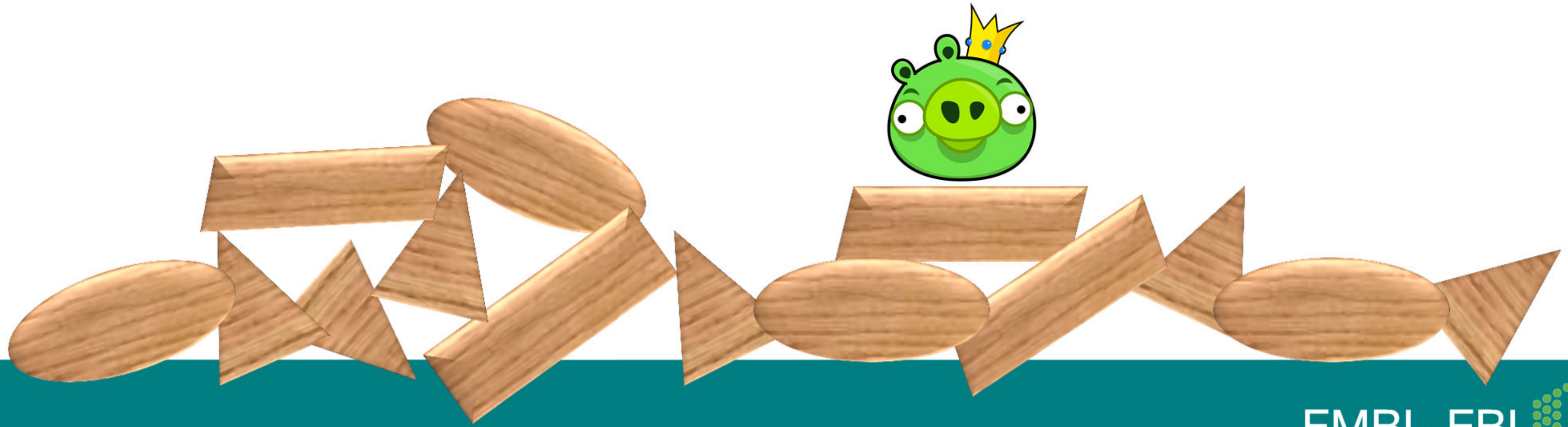
Provider	Format	Shape of the bricks
ELSEVIER	Raw Text	
npg nature publishing group	Raw Text	
ARRAYEXPRESS	XML	
UniProt	RDF	

Data in RDF – How to build a house

Put together:

This doesn't work
You need to have a standard and convenient shape

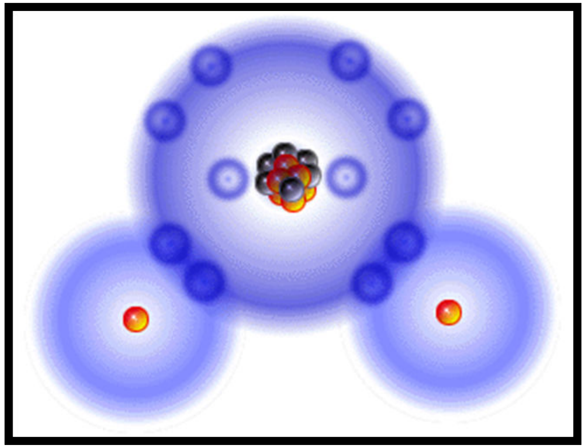
→ **RDF (or RDFS or OWL)**

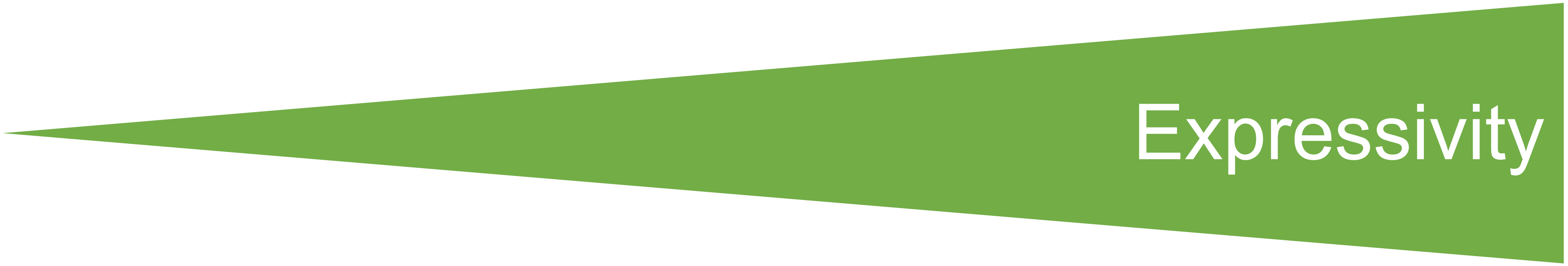


Data in RDF

H_2O

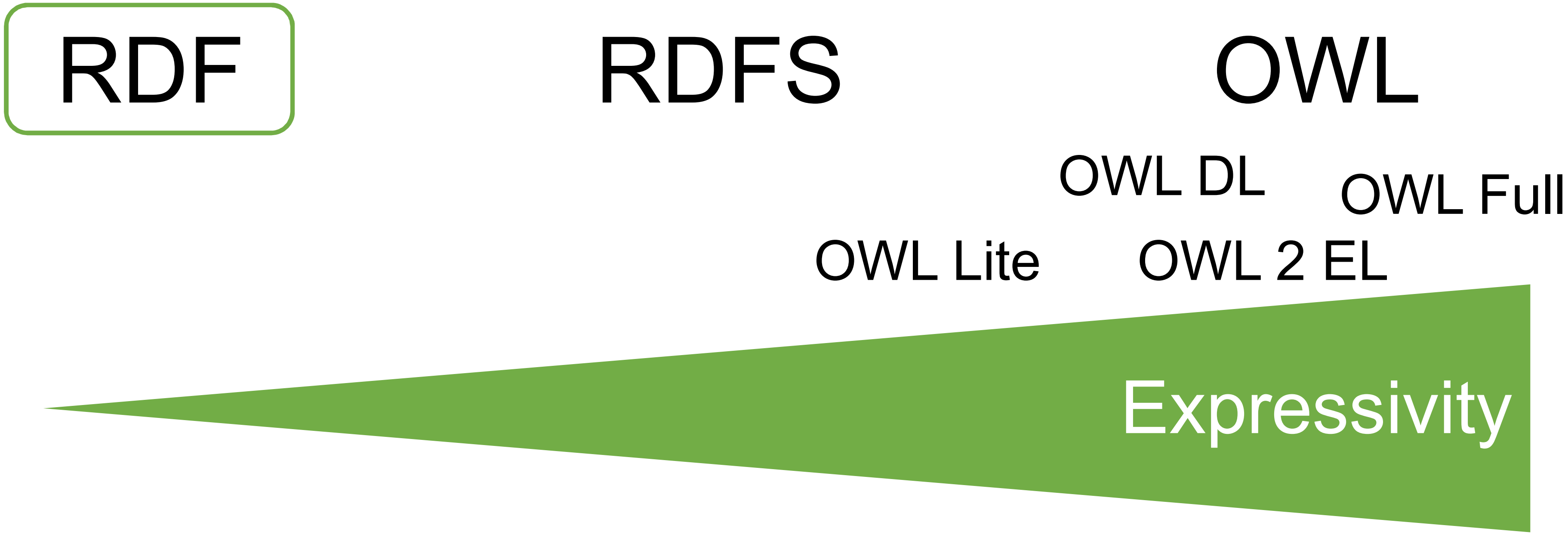
H—O—H





Source Metaphore: Allemang and Hendler

Data in RDF



RDF

RDFS

OWL

OWL DL

OWL Full

OWL Lite

OWL 2 EL

Expressivity

Source Metaphore: Allemang and Hendler

EMBL-EBI

Data in RDF: Adv. / Disadv. Tabular format

Advantages

- Intuitive implementation
- Expansion to the right: more attributes
- Expansion to the bottom: more data
- Combining data across tables through shared keys

Disadvantages

- Semantics between key entry an columns is only implicit
- Change management tends to be costly
- Semantics is only local, i.e. not global, intuitively global

http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

Freie Universität Berlin

Data in RDF – Tabular format to RDF

OMIM:

OMIM_id has_name name;

OMIM_id has_association gene;

UMLS:

UMLS_id has_name name;

UMLS_id same_as mesh_id;

UMLS_id purl URI;

UMLS_id has_type type;

Data in RDF – XML to RDF

ArrayExpress GXA:

Restful API query for Experimental Factor Ontology ids or gene accessions:

experiment has_expression (condition, gene, up/down, pvalue);

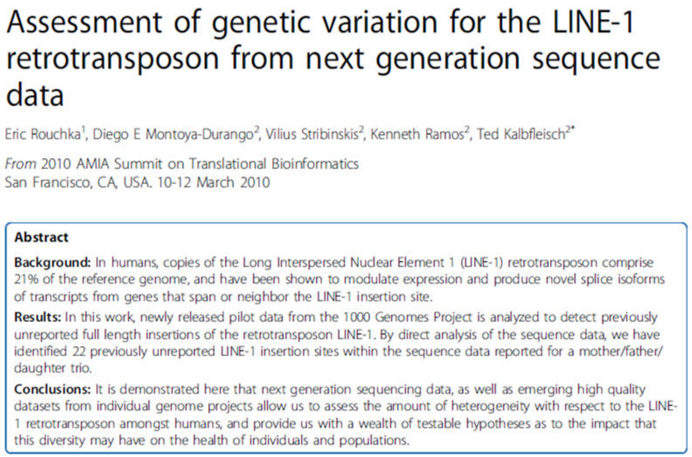
Nice way:

XML → XSLT (mapping) → RDF

Data in RDF – Raw text to RDF

The Challenge

Raw Text



Assessment of genetic variation for the LINE-1 retrotransposon from next generation sequence data

Eric Rouchka[1], Diego E Montoya-Durango[2], Vilius Stribinskis[2], Kenneth Ramos[2], Ted Kalbfleisch[2*]

From 2010 AMIA Summit on Translational Bioinformatics
San Francisco, CA, USA. 10-12 March 2010

Abstract

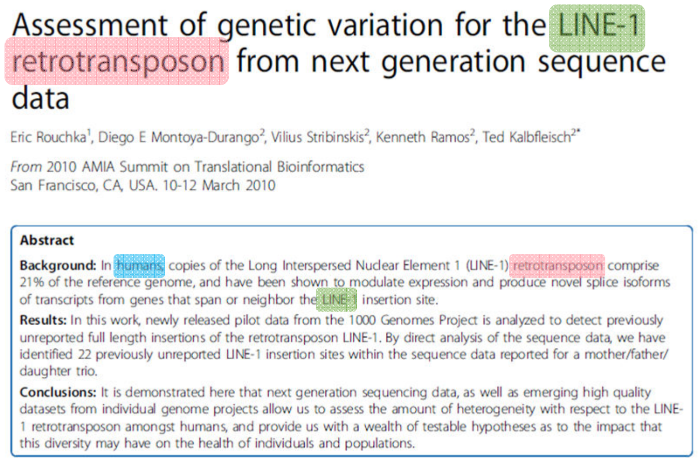
Background: In humans, copies of the Long Interspersed Nuclear Element 1 (LINE-1) retrotransposon comprise 21% of the reference genome, and have been shown to modulate expression and produce novel splice isoforms of transcripts from genes that span or neighbor the LINE-1 insertion site.

Results: In this work, newly released pilot data from the 1000 Genomes Project is analyzed to detect previously unreported full length insertions of the retrotransposon LINE-1. By direct analysis of the sequence data, we have identified 22 previously unreported LINE-1 insertion sites within the sequence data reported for a mother/father/daughter trio.

Conclusions: It is demonstrated here that next generation sequencing data, as well as emerging high quality datasets from individual genome projects allow us to assess the amount of heterogeneity with respect to the LINE-1 retrotransposon amongst humans, and provide us with a wealth of testable hypotheses as to the impact that this diversity may have on the health of individuals and populations.

Text-Mining

Annotated Text



Assessment of genetic variation for the LINE-1 retrotransposon from next generation sequence data

Eric Rouchka[1], Diego E Montoya-Durango[2], Vilius Stribinskis[2], Kenneth Ramos[2], Ted Kalbfleisch[2*]

From 2010 AMIA Summit on Translational Bioinformatics
San Francisco, CA, USA. 10-12 March 2010

Abstract

Background: In humans, copies of the Long Interspersed Nuclear Element 1 (LINE-1) retrotransposon comprise 21% of the reference genome, and have been shown to modulate expression and produce novel splice isoforms of transcripts from genes that span or neighbor the LINE-1 insertion site.

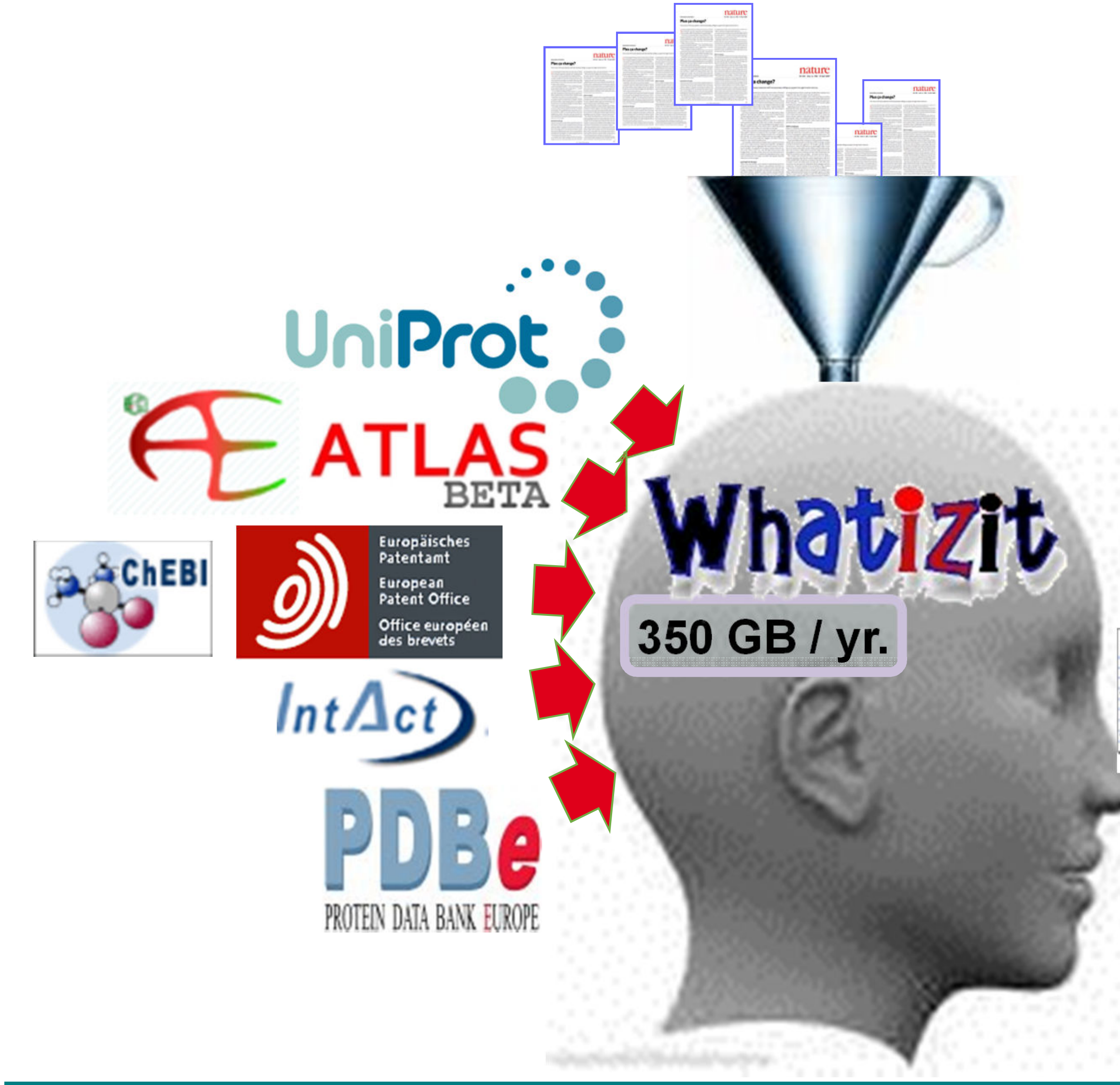
Results: In this work, newly released pilot data from the 1000 Genomes Project is analyzed to detect previously unreported full length insertions of the retrotransposon LINE-1. By direct analysis of the sequence data, we have identified 22 previously unreported LINE-1 insertion sites within the sequence data reported for a mother/father/daughter trio.

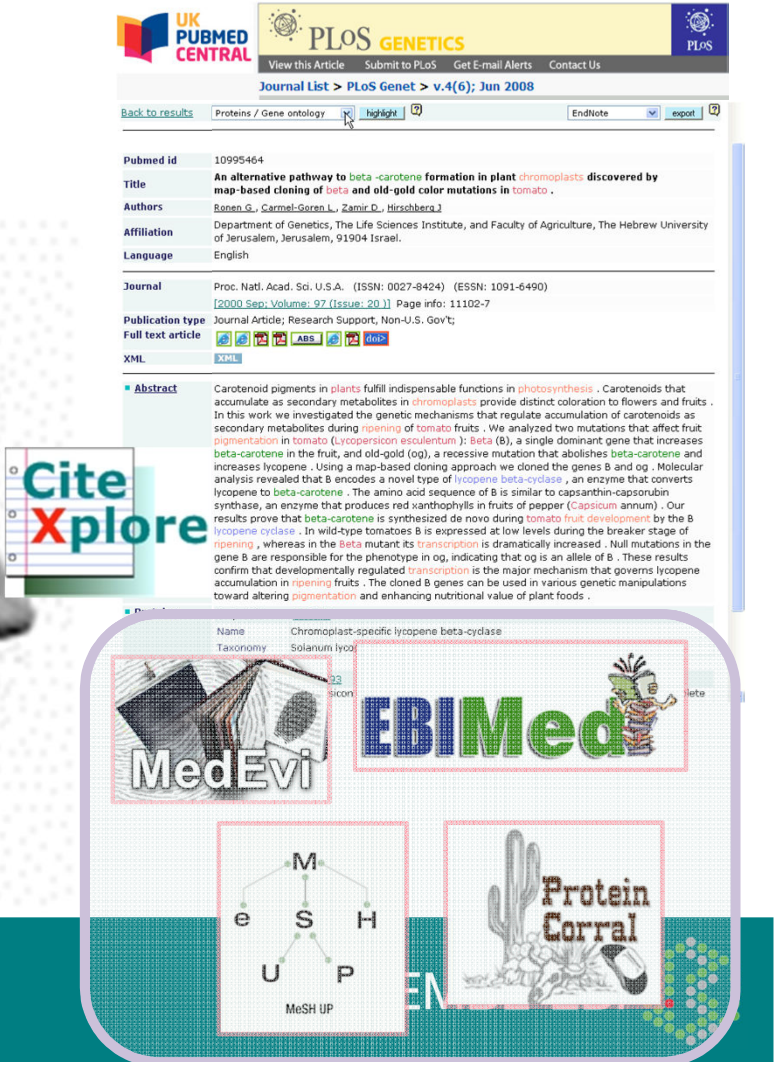
Conclusions: It is demonstrated here that next generation sequencing data, as well as emerging high quality datasets from individual genome projects allow us to assess the amount of heterogeneity with respect to the LINE-1 retrotransposon amongst humans, and provide us with a wealth of testable hypotheses as to the impact that this diversity may have on the health of individuals and populations.

RDF

Integrating biomedical literature and data

Rebholz-Schuhmann, D., et al. Text Processing through Web Services: Calling Whatizit. Bioinformatics 24, no. 2 (2008): 296-98.





Journal List > PLoS Genet > v.4(6); Jun 2008

Pubmed id	10995464
Title	An alternative pathway to beta -carotene formation in plant chromoplasts discovered by map-based cloning of beta and old-gold color mutations in tomato .
Authors	Ronen G , Carmel-Goren L , Zamir D , Hirschberg J
Affiliation	Department of Genetics, The Life Sciences Institute, and Faculty of Agriculture, The Hebrew University of Jerusalem, Jerusalem, 91904 Israel.
Language	English
Journal	Proc. Natl. Acad. Sci. U.S.A. (ISSN: 0027-8424) (ESSN: 1091-6490) [2000 Sep; Volume: 97 (Issue: 20)] Page info: 11102-7
Publication type	Journal Article; Research Support, Non-U.S. Gov't;
Full text article	
XML	XML
Abstract	Carotenoid pigments in plants fulfill indispensable functions in photosynthesis . Carotenoids that accumulate as secondary metabolites in chromoplasts provide distinct coloration to flowers and fruits . In this work we investigated the genetic mechanisms that regulate accumulation of carotenoids as secondary metabolites during ripening of tomato fruits . We analyzed two mutations that affect fruit pigmentation in tomato (Lycopersicon esculentum): Beta (B), a single dominant gene that increases beta-carotene in the fruit, and old-gold (og), a recessive mutation that abolishes beta-carotene and increases lycopene . Using a map-based cloning approach we cloned the genes B and og . Molecular analysis revealed that B encodes a novel type of lycopene beta-cyclase , an enzyme that converts lycopene to beta-carotene . The amino acid sequence of B is similar to capsanthin-capsorubin synthase, an enzyme that produces red xanthophylls in fruits of pepper (Capsicum annum) . Our results prove that beta-carotene is synthesized de novo during tomato fruit development by the B lycopene cyclase . In wild-type tomatoes B is expressed at low levels during the breaker stage of ripening , whereas in the Beta mutant its transcription is dramatically increased . Null mutations in the gene B are responsible for the phenotype in og, indicating that og is an allele of B . These results confirm that developmentally regulated transcription is the major mechanism that governs lycopene accumulation in ripening fruits . The cloned B genes can be used in various genetic manipulations toward altering pigmentation and enhancing nutritional value of plant foods .

Name Chromoplast-specific lycopene beta-cyclase

Taxonomy Solanum lyco

Semantic Web approach: Triples from Text

- Represent everything in Triples, long collections of triples
- Subject – Predicate – Object
- “John loves Mary”
  “The aortic valve is part of the heart”
  “Tamoxifen binds to the estrogen receptor”
  “Retinoblastoma is located in the eye”
- Formatting:

John | Mary	[In the love or relationship database]
John | loves | Mary	[In a simple relational database]
John | love | Mary	[In a normalised relational database]
:John :love :Mary	[In a very simple RDF representation]

Data in RDF – Raw text to RDF

Literature – Meta Level - XML documents:

document_id has_paragraphs paragraph_id;
has_title “title”; has authors (“author”, “author”);
has_DOI doi; has_metadata metadata .

Literature – Annotation Level - XML annotations:

document_id has_sentence “sentence”;
part_of paragraph_id;
has_annotation (**type**, **URI**, frequency) .

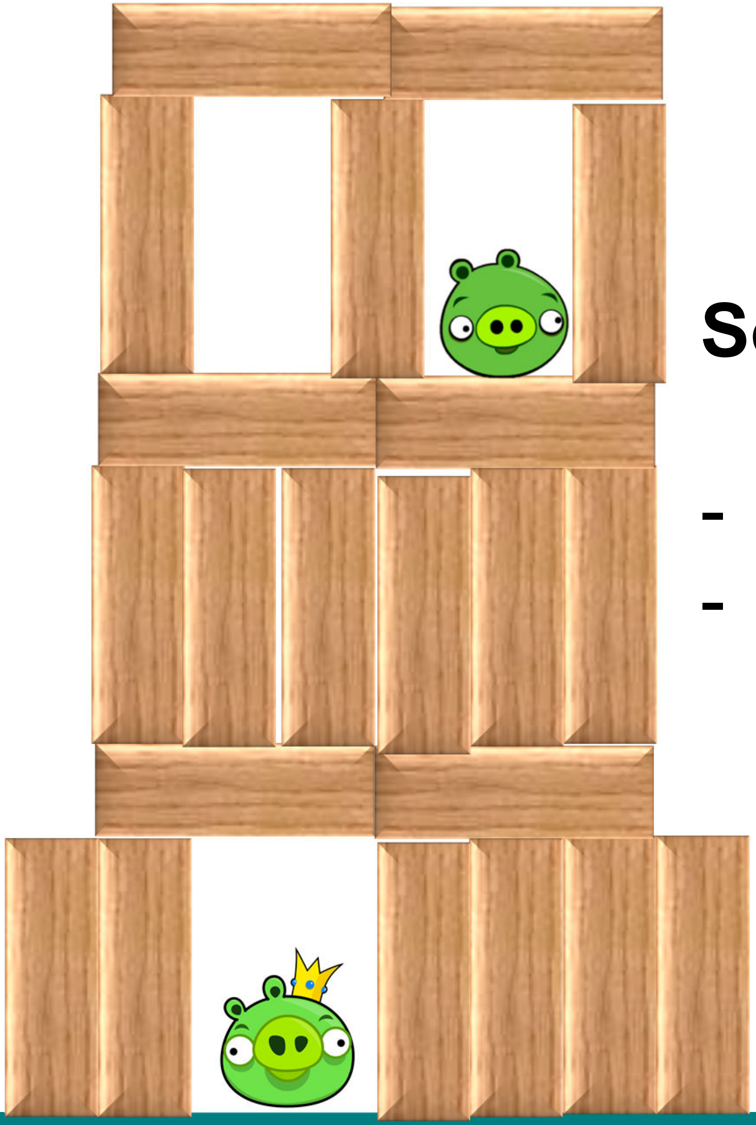
Data in RDF – How to build a house

Provider	Format	Shape of the bricks
ELSEVIER	RDF	
npg nature publishing group	RDF	
ARRAYEXPRESS	RDF	
UniProt Was already in RDF!!	RDF	

Data in RDF – How to build a house

Put together:

This work!

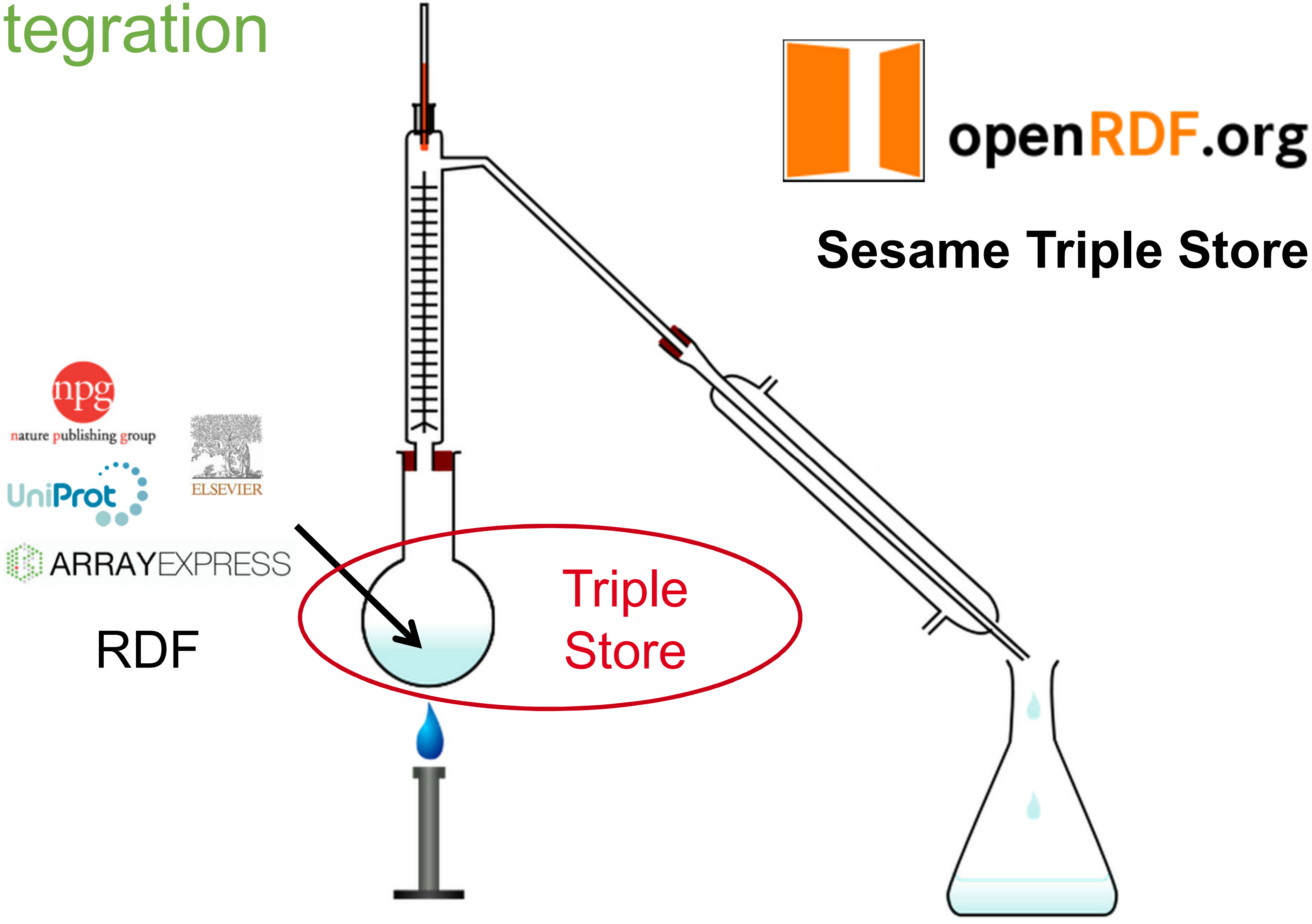


Solid Integration:

- URIs
- Triples

EMBL-EBI

Integration



Sesame Triple Store

What have we done? The SESL project

RS•C ROYAL SOCIETY OF CHEMISTRY

npg nature publishing group

ELSEVIER

OXFORD UNIVERSITY P

Publishers

ARRAYEXPRESS

OMIM Online Mendelian Inheritance in Man

WIKIPEDIA

UniProt

Bio Repositories

Rebholz Group

Semantic Web

dical

questions

Pharmaceuticals/Food Companies

gsk GlaxoSmithKline

AstraZeneca

Unilever

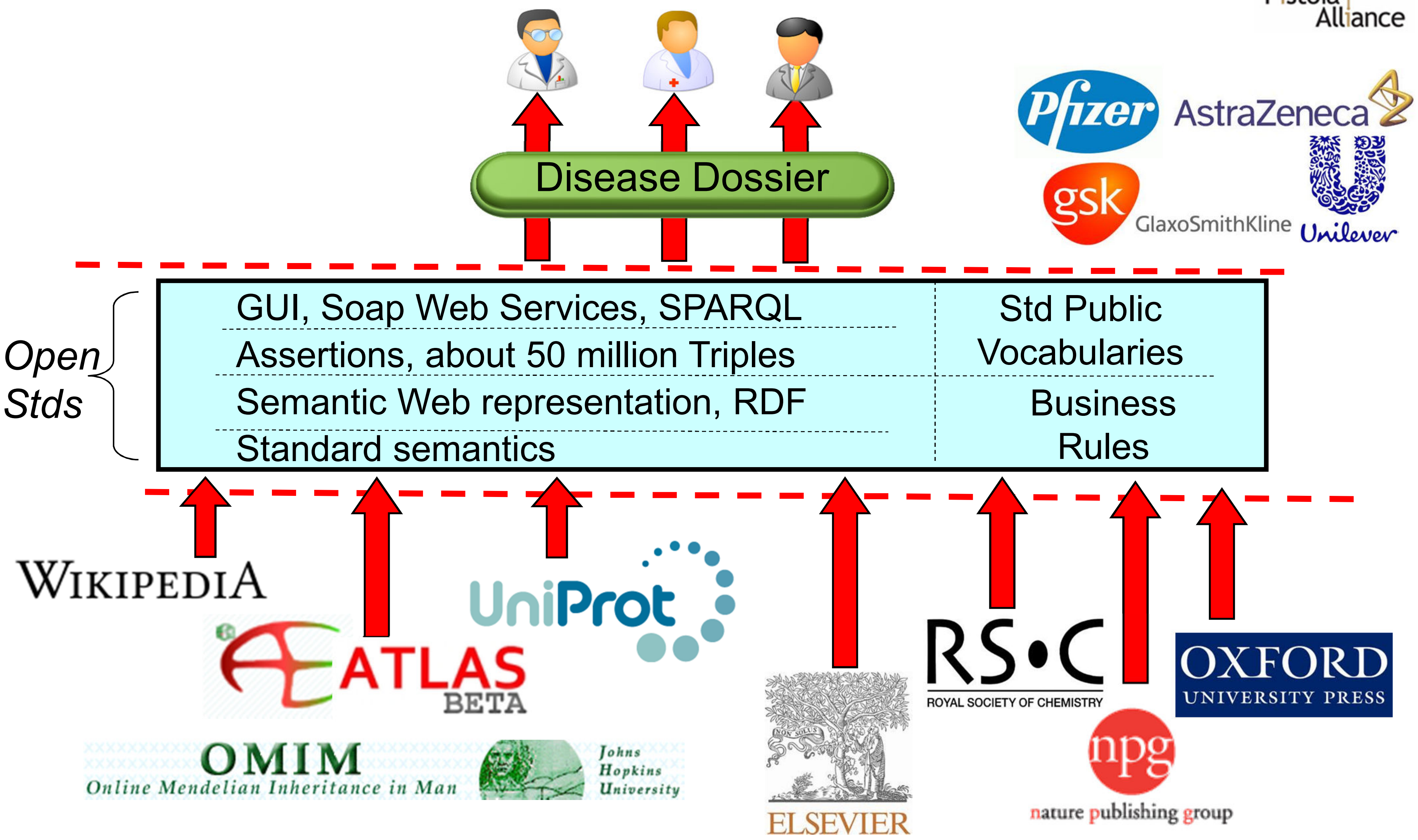
Pfizer

Roche

Pistoia Alliance

EMBL-EBI

SESL .. The brokering of knowledge



Disease Dossier

Open Stds

GUI, Soap Web Services, SPARQL

Assertions, about 50 million Triples

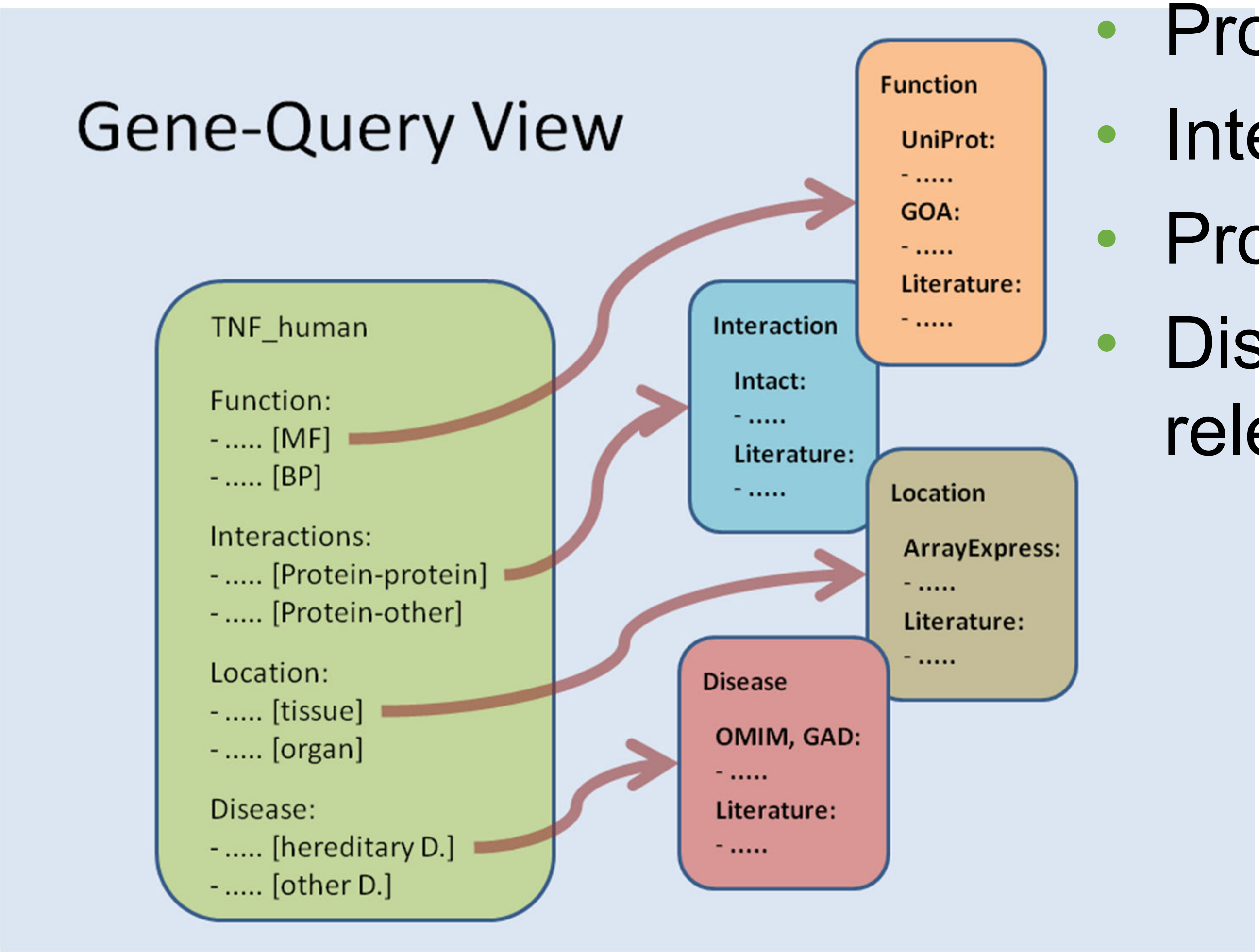
Semantic Web representation, RDF

Standard semantics

Std Public Vocabularies

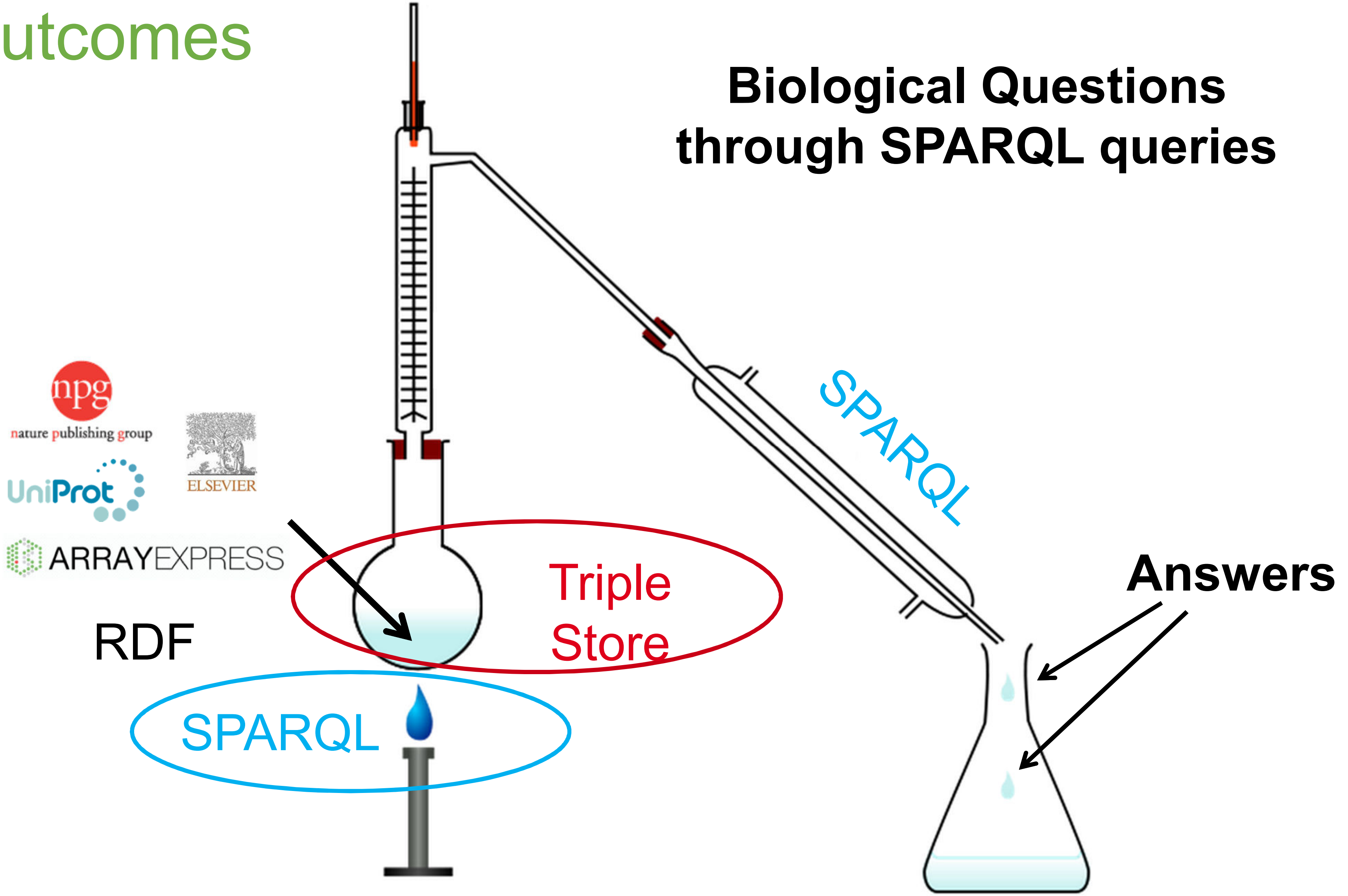
Business Rules

Querying a gene for the dossier

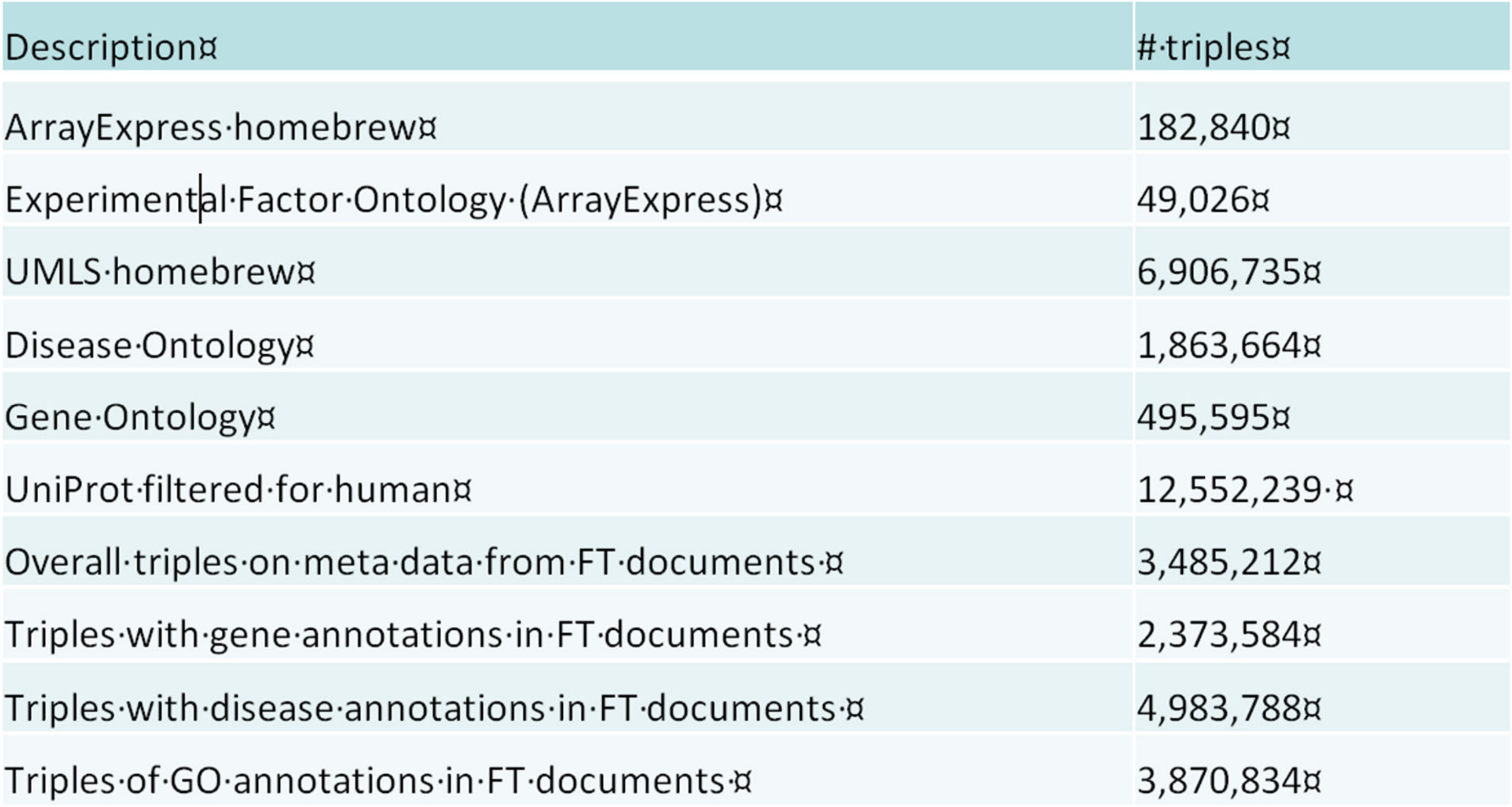


- Protein Function
- Interactions
- Protein location
- Disease relevance

Outcomes


Biological Questions through SPARQL queries
npg
nature publishing group
ELSEVIER
UniProt
ARRAYEXPRESS
RDF
Triple Store
SPARQL
SPARQL
Answers

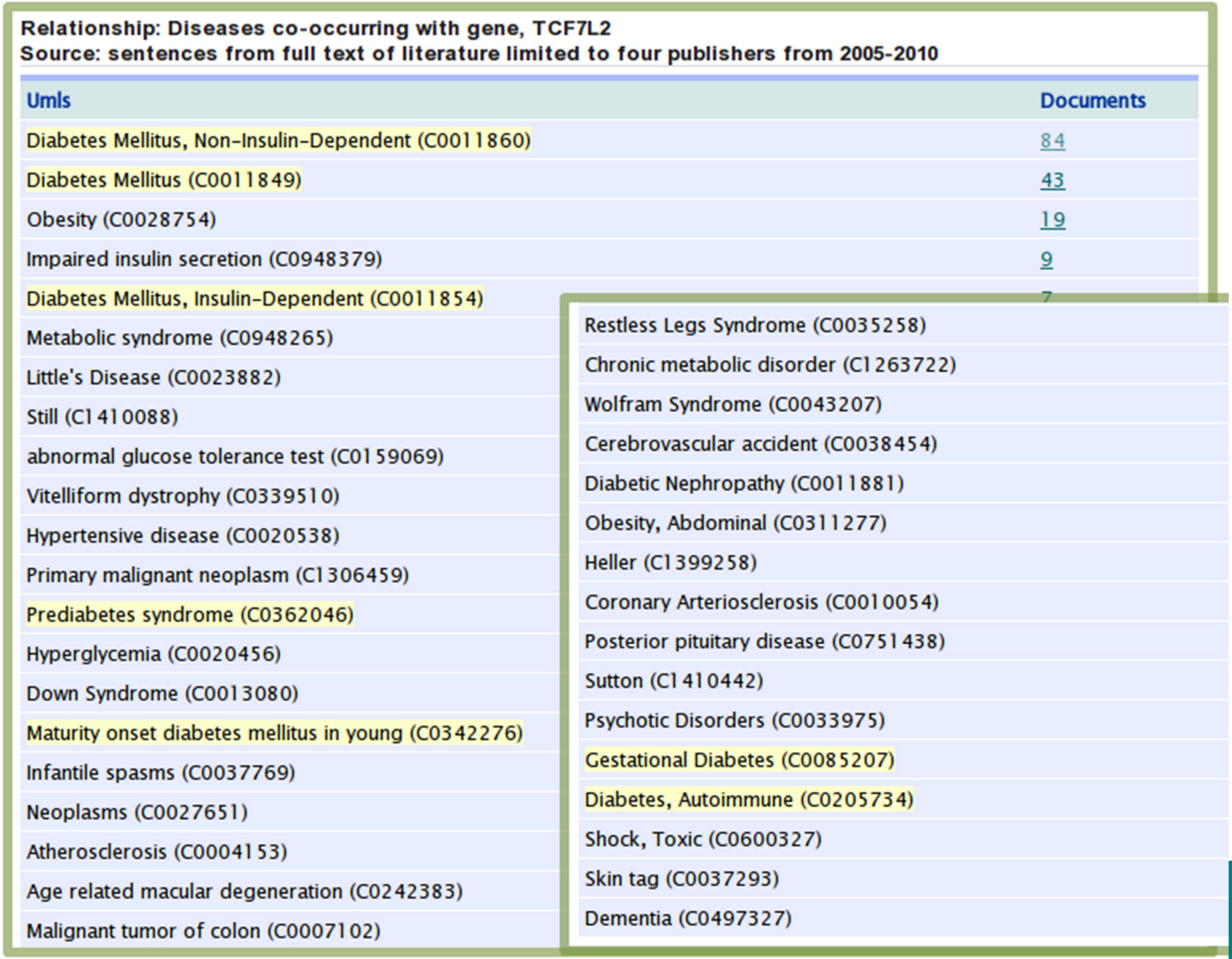
Content of the triple store (1)



Description¤	#·triples¤
ArrayExpress·homebrew¤	182,840¤
Experimental·Factor·Ontology·(ArrayExpress)¤	49,026¤
UMLS·homebrew¤	6,906,735¤
Disease·Ontology¤	1,863,664¤
Gene·Ontology¤	495,595¤
UniProt·filtered·for·human¤	12,552,239·¤
Overall·triples·on·meta·data·from·FT·documents·¤	3,485,212¤
Triples·with·gene·annotations·in·FT·documents·¤	2,373,584¤
Triples·with·disease·annotations·in·FT·documents·¤	4,983,788¤
Triples·of·GO·annotations·in·FT·documents·¤	3,870,834¤

Diseases related to TCF7L2

Pistoia Alliance



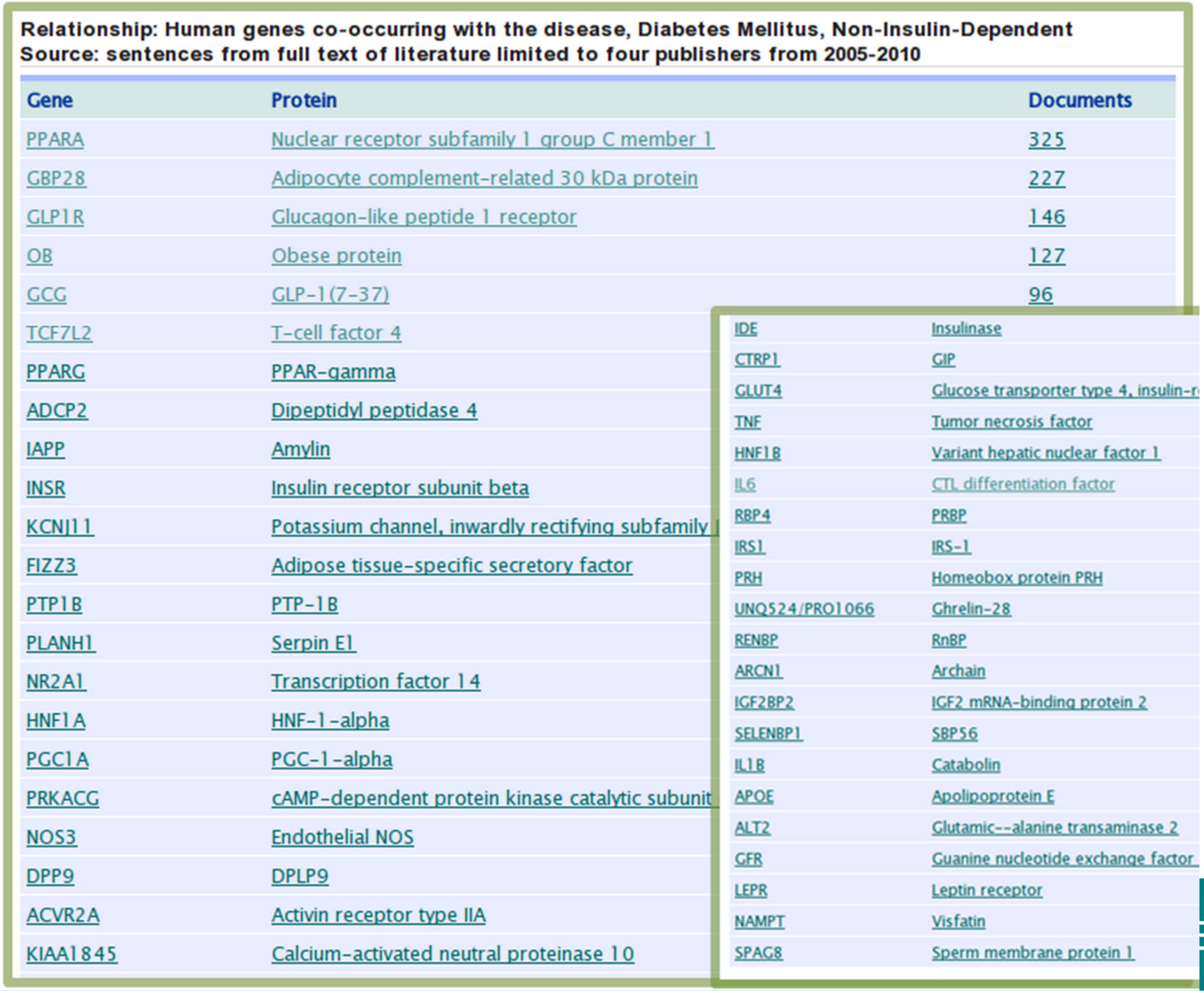
Relationship: Diseases co-occurring with gene, TCF7L2
Source: sentences from full text of literature limited to four publishers from 2005-2010

Umls	Documents
Diabetes Mellitus, Non-Insulin-Dependent (C0011860)	84
Diabetes Mellitus (C0011849)	43
Obesity (C0028754)	19
Impaired insulin secretion (C0948379)	9
Diabetes Mellitus, Insulin-Dependent (C0011854)	7
Metabolic syndrome (C0948265)	
Little's Disease (C0023882)	
Still (C1410088)	
abnormal glucose tolerance test (C0159069)	
Vitelliform dystrophy (C0339510)	
Hypertensive disease (C0020538)	
Primary malignant neoplasm (C1306459)	
Prediabetes syndrome (C0362046)	
Hyperglycemia (C0020456)	
Down Syndrome (C0013080)	
Maturity onset diabetes mellitus in young (C0342276)	
Infantile spasms (C0037769)	
Neoplasms (C0027651)	
Atherosclerosis (C0004153)	
Age related macular degeneration (C0242383)	
Malignant tumor of colon (C0007102)	
Restless Legs Syndrome (C0035258)	
Chronic metabolic disorder (C1263722)	
Wolfram Syndrome (C0043207)	
Cerebrovascular accident (C0038454)	
Diabetic Nephropathy (C0011881)	
Obesity, Abdominal (C0311277)	
Heller (C1399258)	
Coronary Arteriosclerosis (C0010054)	
Posterior pituitary disease (C0751438)	
Sutton (C1410442)	
Psychotic Disorders (C0033975)	
Gestational Diabetes (C0085207)	
Diabetes, Autoimmune (C0205734)	
Shock, Toxic (C0600327)	
Skin tag (C0037293)	
Dementia (C0497327)	

BL-EBI

Genes relevant to DmT2

Pistoia Alliance



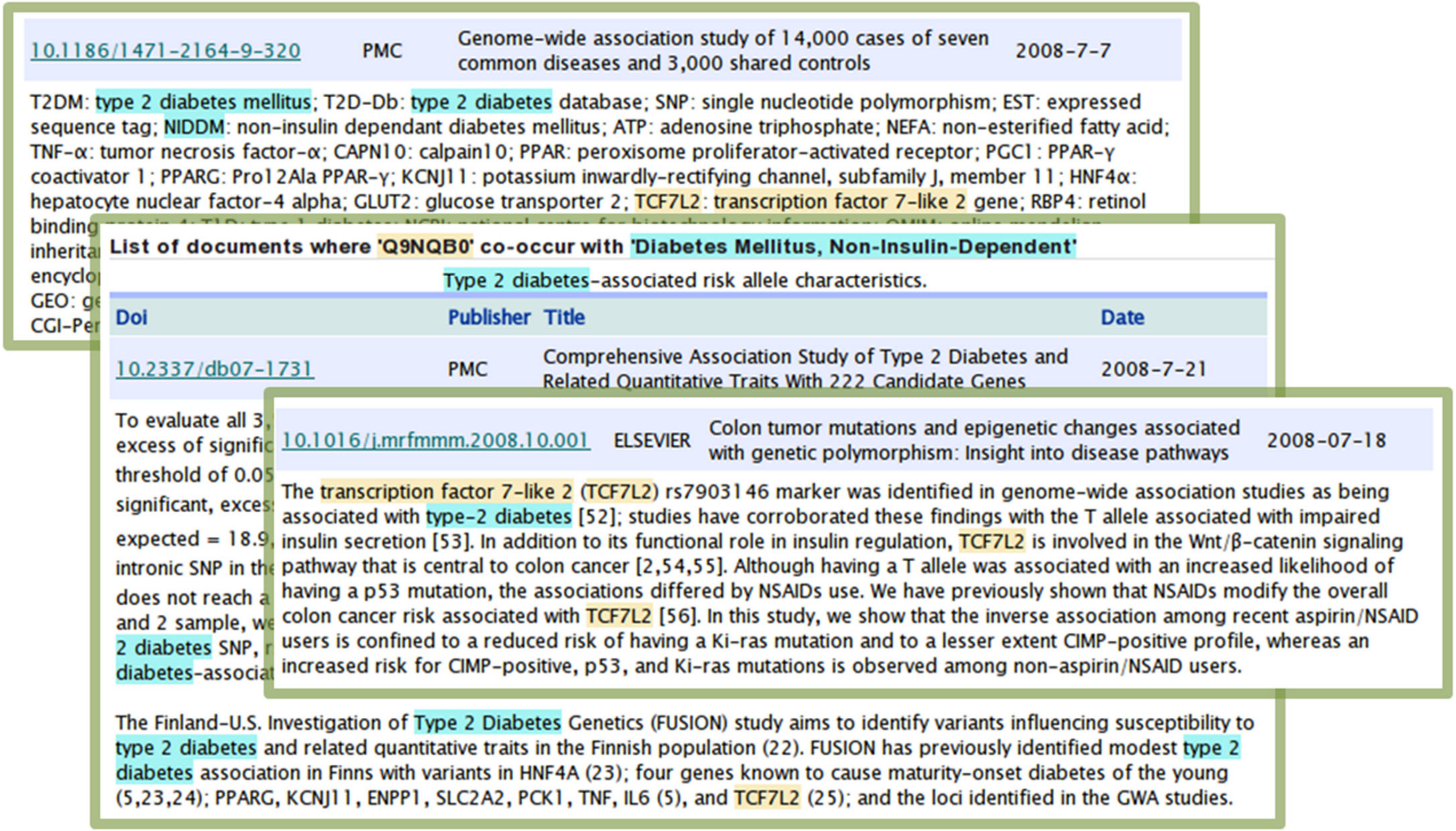
Relationship: Human genes co-occurring with the disease, Diabetes Mellitus, Non-Insulin-Dependent
Source: sentences from full text of literature limited to four publishers from 2005-2010

Gene	Protein	Documents
PPARA	Nuclear receptor subfamily 1 group C member 1	325
GBP28	Adipocyte complement-related 30 kDa protein	227
GLP1R	Glucagon-like peptide 1 receptor	146
OB	Obese protein	127
GCG	GLP-1(7-37)	96
TCF7L2	T-cell factor 4	
PPARG	PPAR-gamma	
ADCP2	Dipeptidyl peptidase 4	
IAPP	Amylin	
INSR	Insulin receptor subunit beta	
KCNJ11	Potassium channel, inwardly rectifying subfamily	
FIZZ3	Adipose tissue-specific secretory factor	
PTP1B	PTP-1B	
PLANH1	Serpin E1	
NR2A1	Transcription factor 14	
HNF1A	HNF-1-alpha	
PGC1A	PGC-1-alpha	
PRKACG	cAMP-dependent protein kinase catalytic subunit	
NOS3	Endothelial NOS	
DPP9	DPLP9	
ACVR2A	Activin receptor type IIA	
KIAA1845	Calcium-activated neutral proteinase 10	

IDE	Insulinase
CTRP1	GIP
GLUT4	Glucose transporter type 4, insulin-r
TNF	Tumor necrosis factor
HNF1B	Variant hepatic nuclear factor 1
IL6	CTL differentiation factor
RBP4	PRBP
IRS1	IRS-1
PRH	Homeobox protein PRH
UNQ524/PRO1066	Ghrelin-28
RENBP	RnBP
ARCN1	Archain
IGF2BP2	IGF2 mRNA-binding protein 2
SELENBP1	SBP56
IL1B	Catabolin
APOE	Apolipoprotein E
ALT2	Glutamic--alanine transaminase 2
GFR	Guanine nucleotide exchange factor
LEPR	Leptin receptor
NAMPT	Visfatin
SPAG8	Sperm membrane protein 1

EMBL-EBI

Literature content



10.1186/1471-2164-9-320	PMC	Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls	2008-7-7

T2DM: type 2 diabetes mellitus; T2D-Db: type 2 diabetes database; SNP: single nucleotide polymorphism; EST: expressed sequence tag; NIDDM: non-insulin dependant diabetes mellitus; ATP: adenosine triphosphate; NEFA: non-esterified fatty acid; TNF-α: tumor necrosis factor-α; CAPN10: calpain10; PPAR: peroxisome proliferator-activated receptor; PGC1: PPAR-γ coactivator 1; PPARG: Pro12Ala PPAR-γ; KCNJ11: potassium inwardly-rectifying channel, subfamily J, member 11; HNF4α: hepatocyte nuclear factor-4 alpha; GLUT2: glucose transporter 2; TCF7L2: transcription factor 7-like 2 gene; RBP4: retinol binding
inherita
encyclo
GEO: ge
CGI-Per

List of documents where 'Q9NQB0' co-occur with 'Diabetes Mellitus, Non-Insulin-Dependent'

Type 2 diabetes-associated risk allele characteristics.

Doi	Publisher	Title	Date
10.2337/db07-1731	PMC	Comprehensive Association Study of Type 2 Diabetes and Related Quantitative Traits With 222 Candidate Genes	2008-7-21

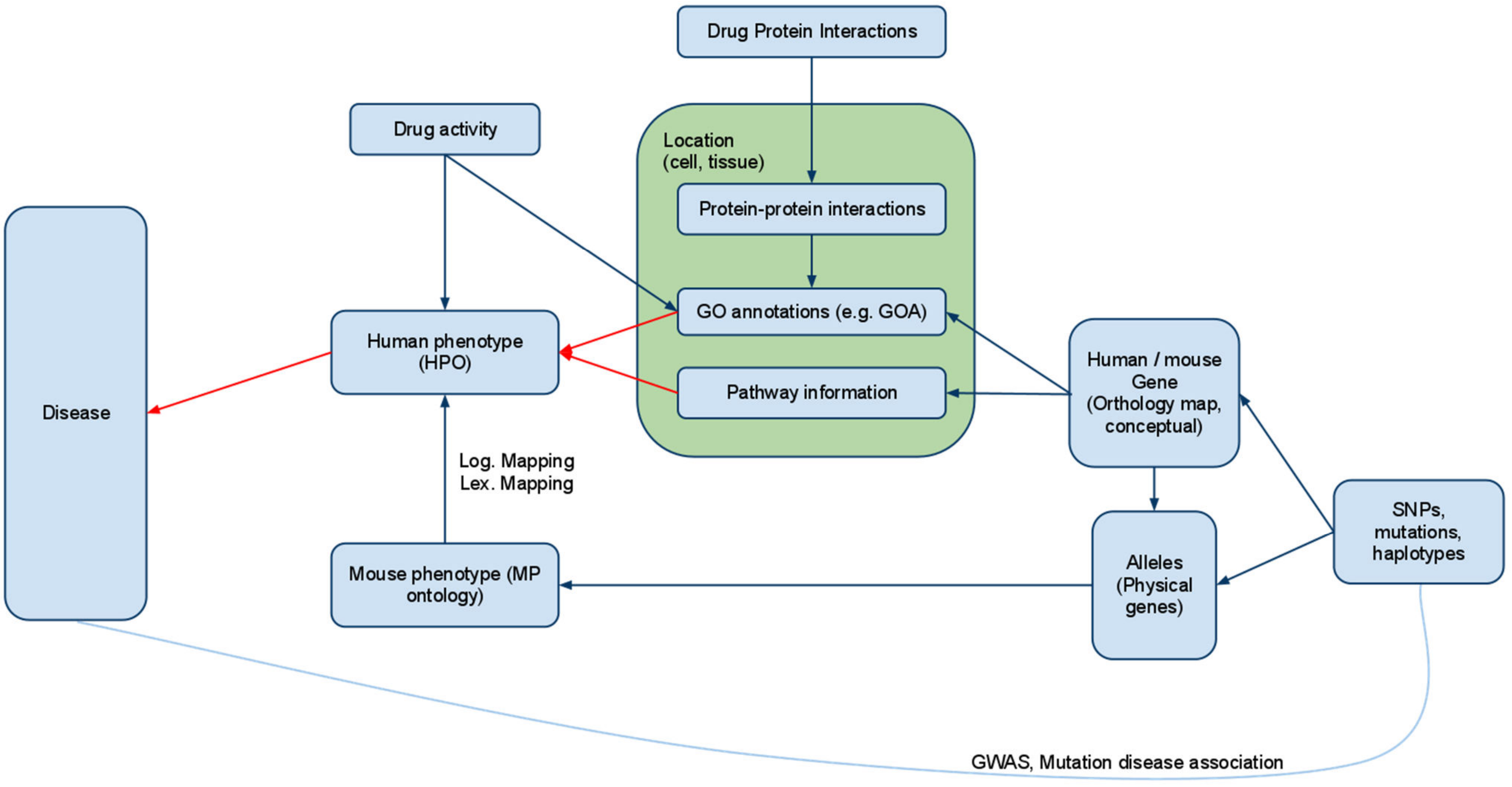
To evaluate all 3
excess of signific
threshold of 0.05
significant, exces
expected = 18.9
intronic SNP in th
does not reach a
and 2 sample, we
2 diabetes SNP,
diabetes-associat

10.1016/j.mrfmmm.2008.10.001	ELSEVIER	Colon tumor mutations and epigenetic changes associated with genetic polymorphism: Insight into disease pathways	2008-07-18

The transcription factor 7-like 2 (TCF7L2) rs7903146 marker was identified in genome-wide association studies as being associated with type-2 diabetes [52]; studies have corroborated these findings with the T allele associated with impaired insulin secretion [53]. In addition to its functional role in insulin regulation, TCF7L2 is involved in the Wnt/β-catenin signaling pathway that is central to colon cancer [2,54,55]. Although having a T allele was associated with an increased likelihood of having a p53 mutation, the associations differed by NSAIDs use. We have previously shown that NSAIDs modify the overall colon cancer risk associated with TCF7L2 [56]. In this study, we show that the inverse association among recent aspirin/NSAID users is confined to a reduced risk of having a Ki-ras mutation and to a lesser extent CIMP-positive profile, whereas an increased risk for CIMP-positive, p53, and Ki-ras mutations is observed among non-aspirin/NSAID users.

The Finland-U.S. Investigation of Type 2 Diabetes Genetics (FUSION) study aims to identify variants influencing susceptibility to type 2 diabetes and related quantitative traits in the Finnish population (22). FUSION has previously identified modest type 2 diabetes association in Finns with variants in HNF4A (23); four genes known to cause maturity-onset diabetes of the young (5,23,24); PPARG, KCNJ11, ENPP1, SLC2A2, PCK1, TNF, IL6 (5), and TCF7L2 (25); and the loci identified in the GWA studies.

G2D: semantic links (Reprise)



What next? Web Ontology Language (OWL)

Reasoning:

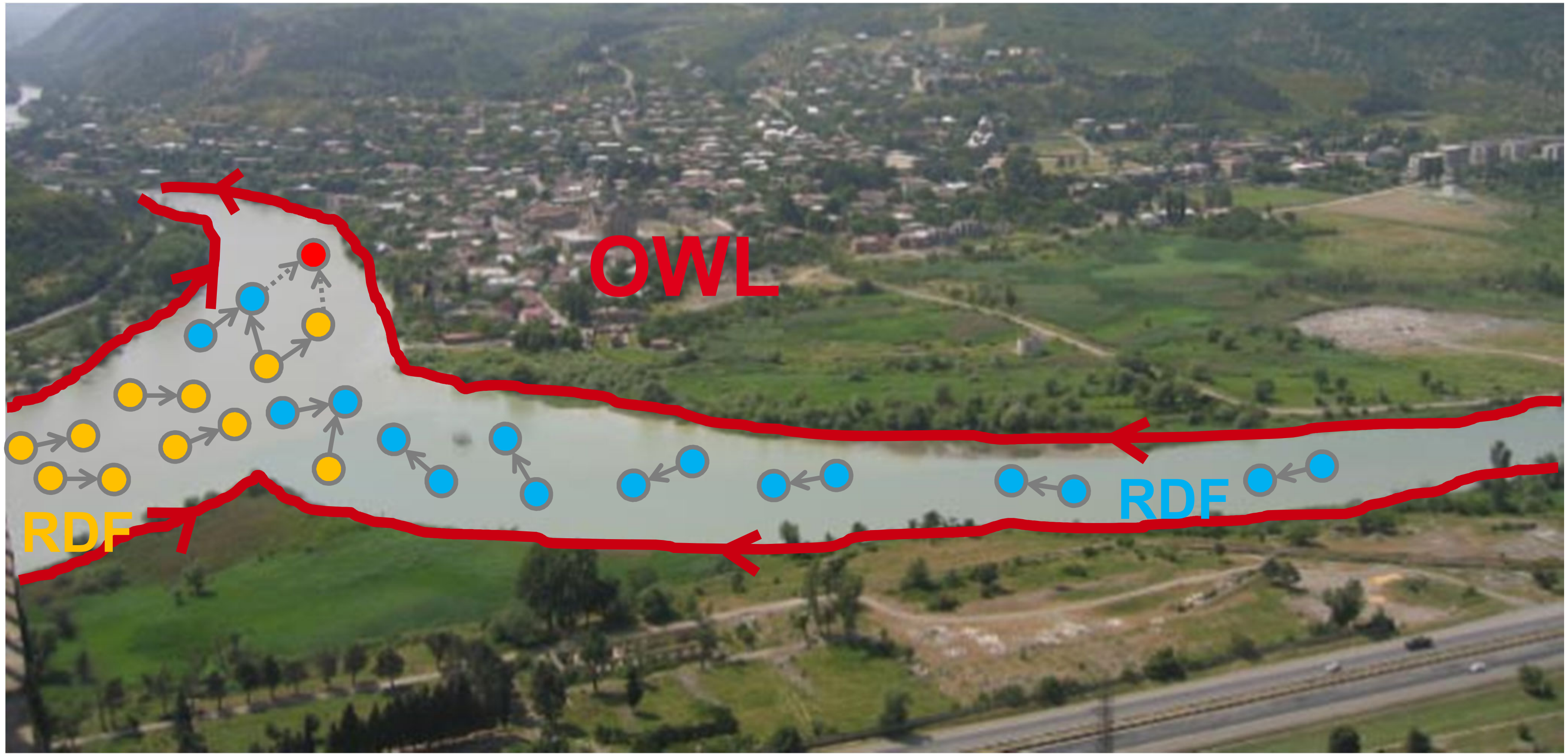
- Extension of RDF
- Properties could have extra logics
  - Transitivity
  - Symmetry
  - Classes – Subclasses
  - Exclusion
  - …

A resoner is able to interpret these rules:

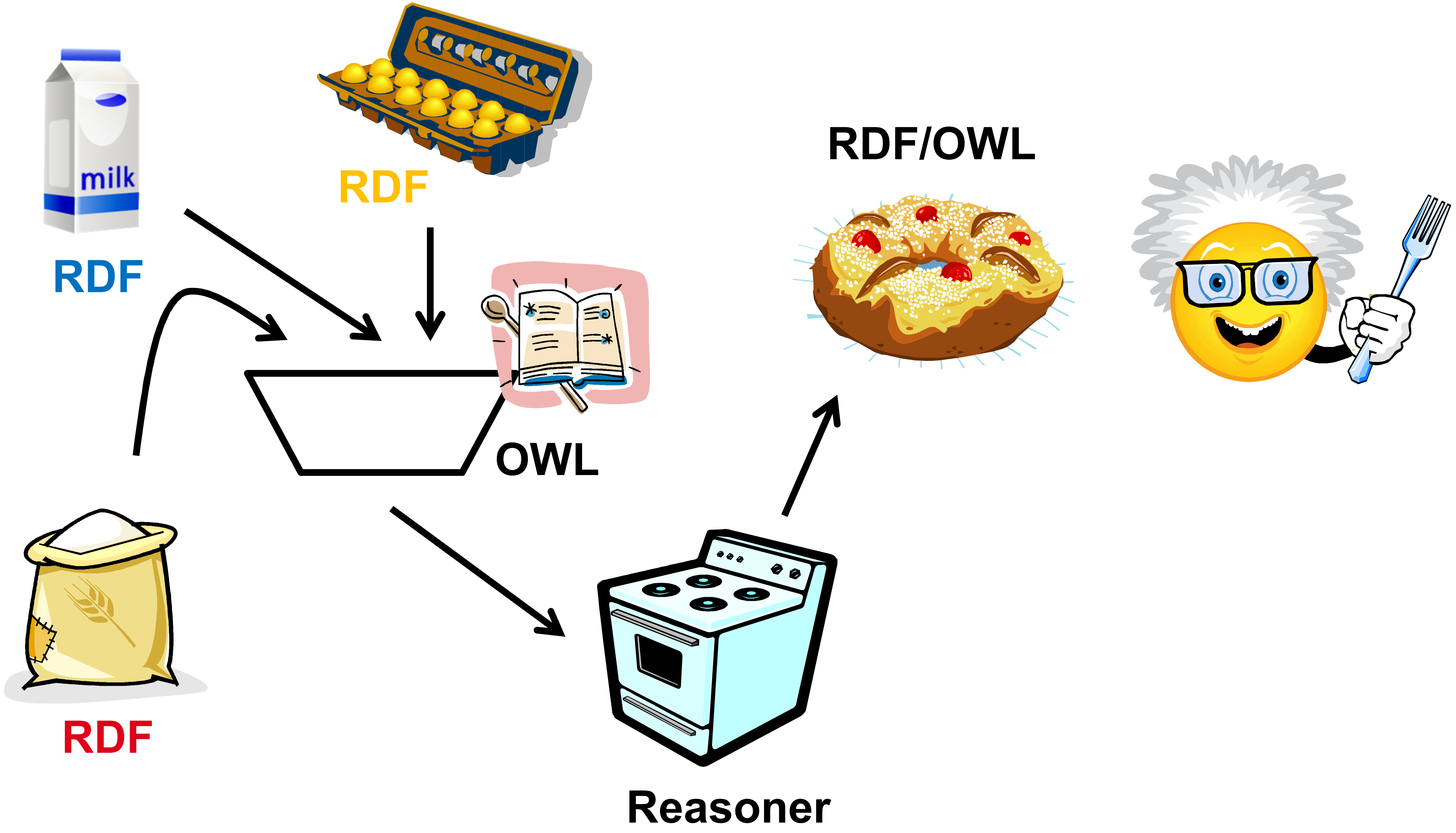
→ Consistent integration of different datasets in RDF

→ Knowledge discovery

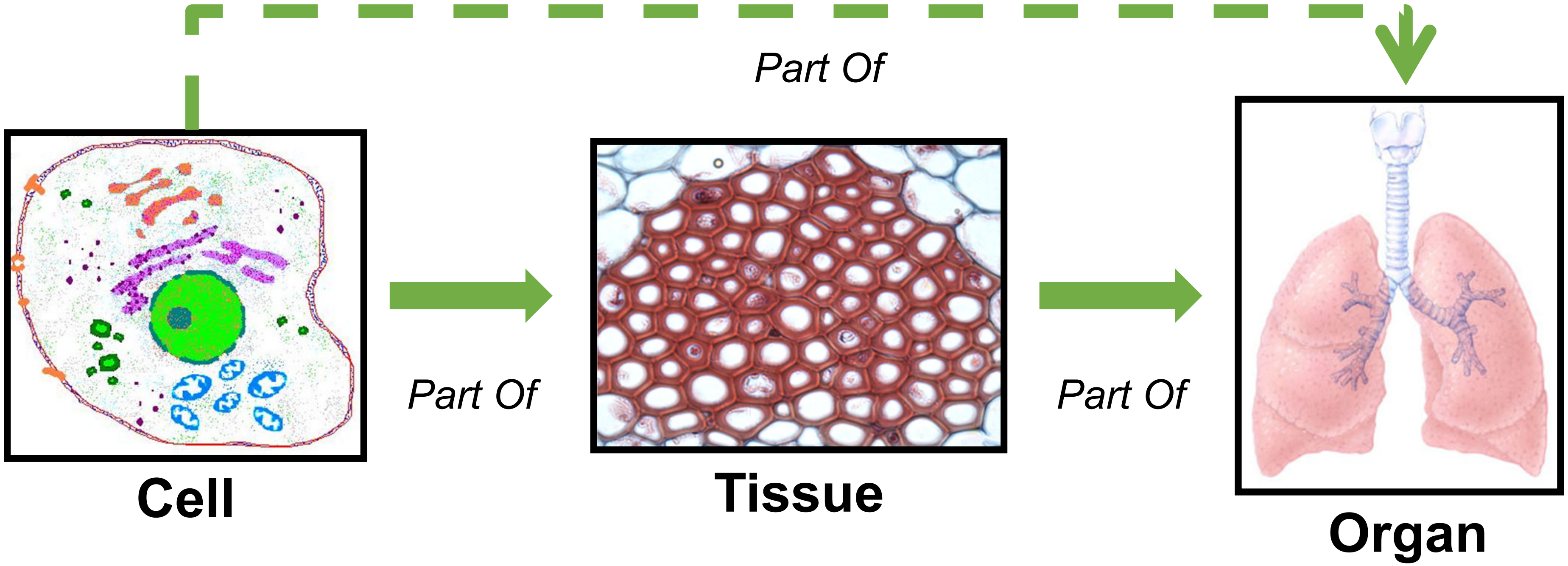
Web Ontology Language (OWL)



Web Ontology Language (OWL)



Transitive Property



Part Of

Part Of

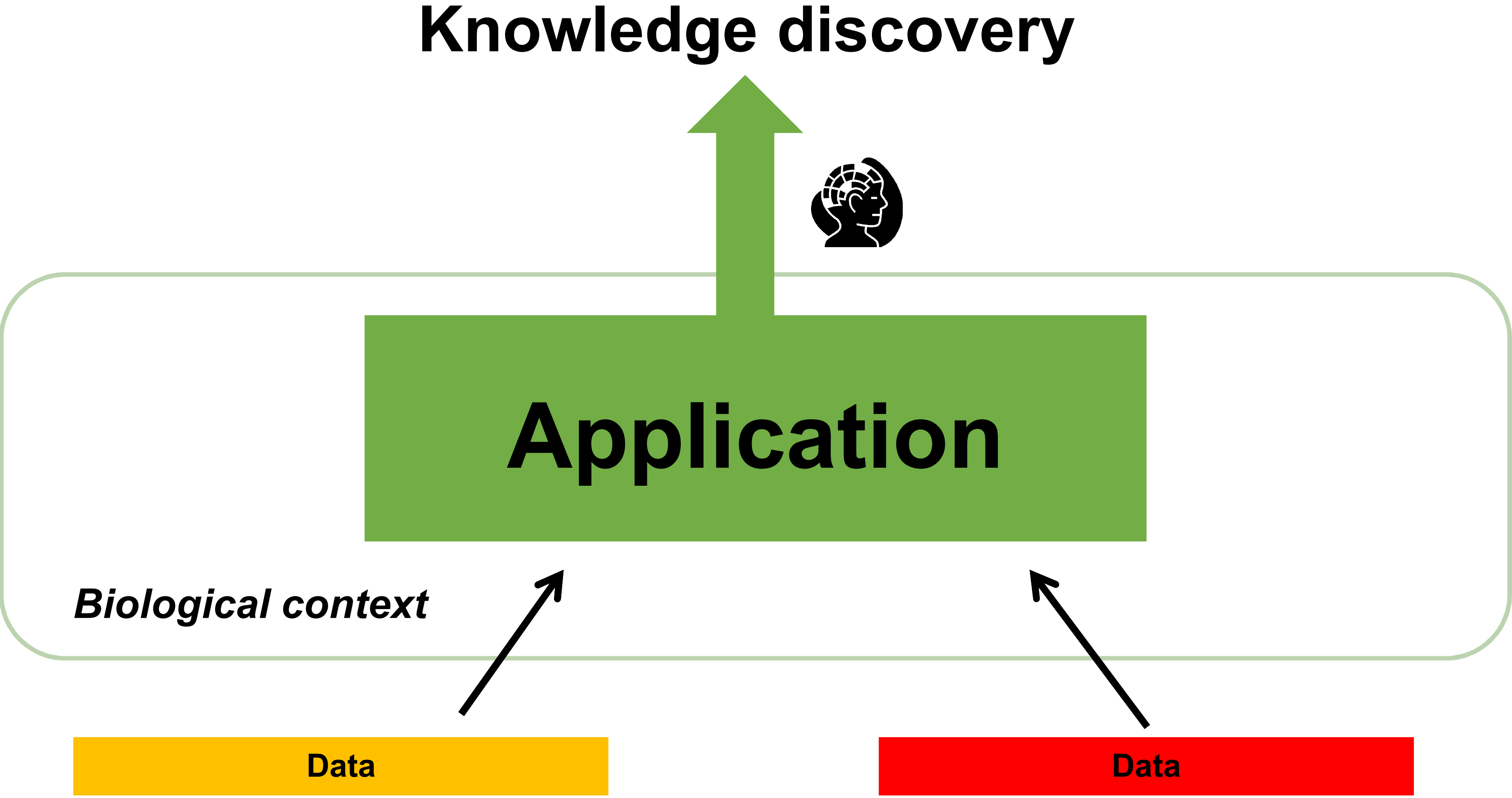
Part Of

Cell

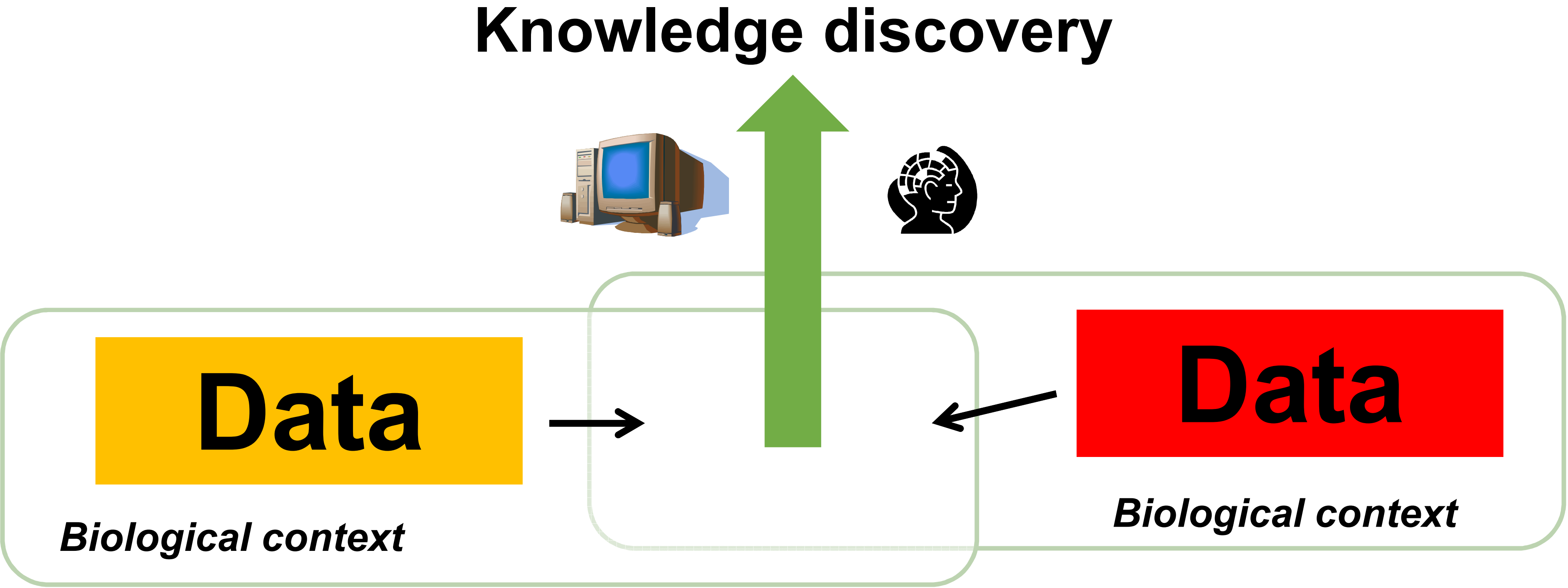
Tissue

Organ

Classical integration in Life Science

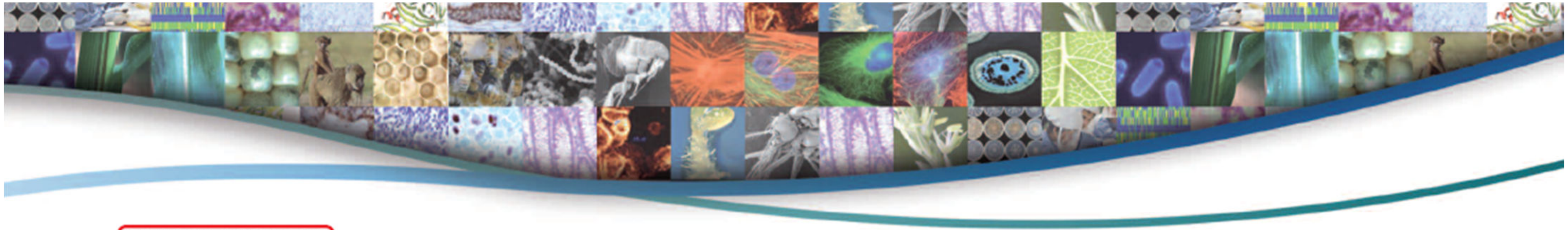


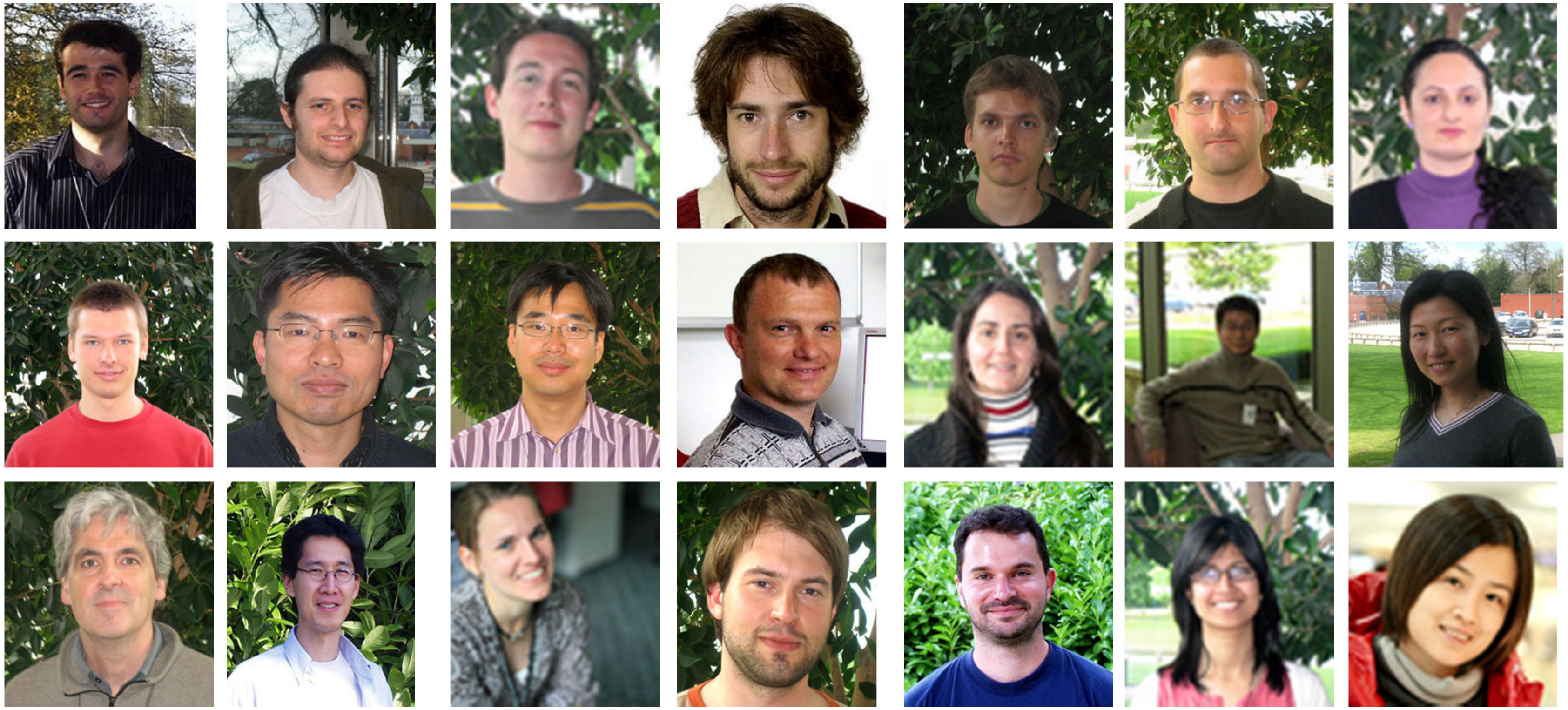
Semantic Web Integration



Final Conclusions

- Content from the scientific literature can be processed to produce facts in RDF representation (triples)
- The integration of all data can be achieved in such a way that:
  - The user gets a fully integrated body of data
  - The underlying resources can be distributed
  - The volume can be large
  - The resources are delivered from different providers





Delphine Bas, Abhishek Dixit, Arun Gupta, Rohit Rexa, Darius Sulskus, Michele Mattioni, Friteyre Caroline, Bailif Alexandre, Electra Tapanari, Elisabet Casanova, Romain Tertieux, Alejandro Pironti, Ewa Stocka, Francisco Couto, Joerg Hakenberg, Dolf Trieschnigg, Andra Waagmeester, Pinar Yildirim, Samira Jaeger, Senay Kafkas, Tiago Grego, David Campos, Ernesto Jimeno Ruiz, Yann Guilbaud, Adam Bernard, Samuel Croset, Sylvain Gaudan, Chen Li, Anika Oellrich, Ying Yan, Kevin Nagel, Ruth Lowering, Alex Griegspoor, Robert Hoehndorf, Jung-Jae Kim, Maria Liakata, Miguel Arregui, Christoph Grabmueller, Silvestras Kavaliauskas, Jee-Hyub Kim, Harald Kirsch, Vivian Lee, Ian Lewin, Menaka Narayasamy, Piotr Pezik, Shyamasree Saha, Antonio J Yepes,