Aplicación de la Web Semántica en Biología Molecular

Mikel Egaña Aranguren

mikel.egana.aranguren@gmail.com

http://mikeleganaaranguren.com

http://tinyurl.com/2u6vhqe

Información en Biología Molecular

Life Sciences Semantic Web (LSSW)

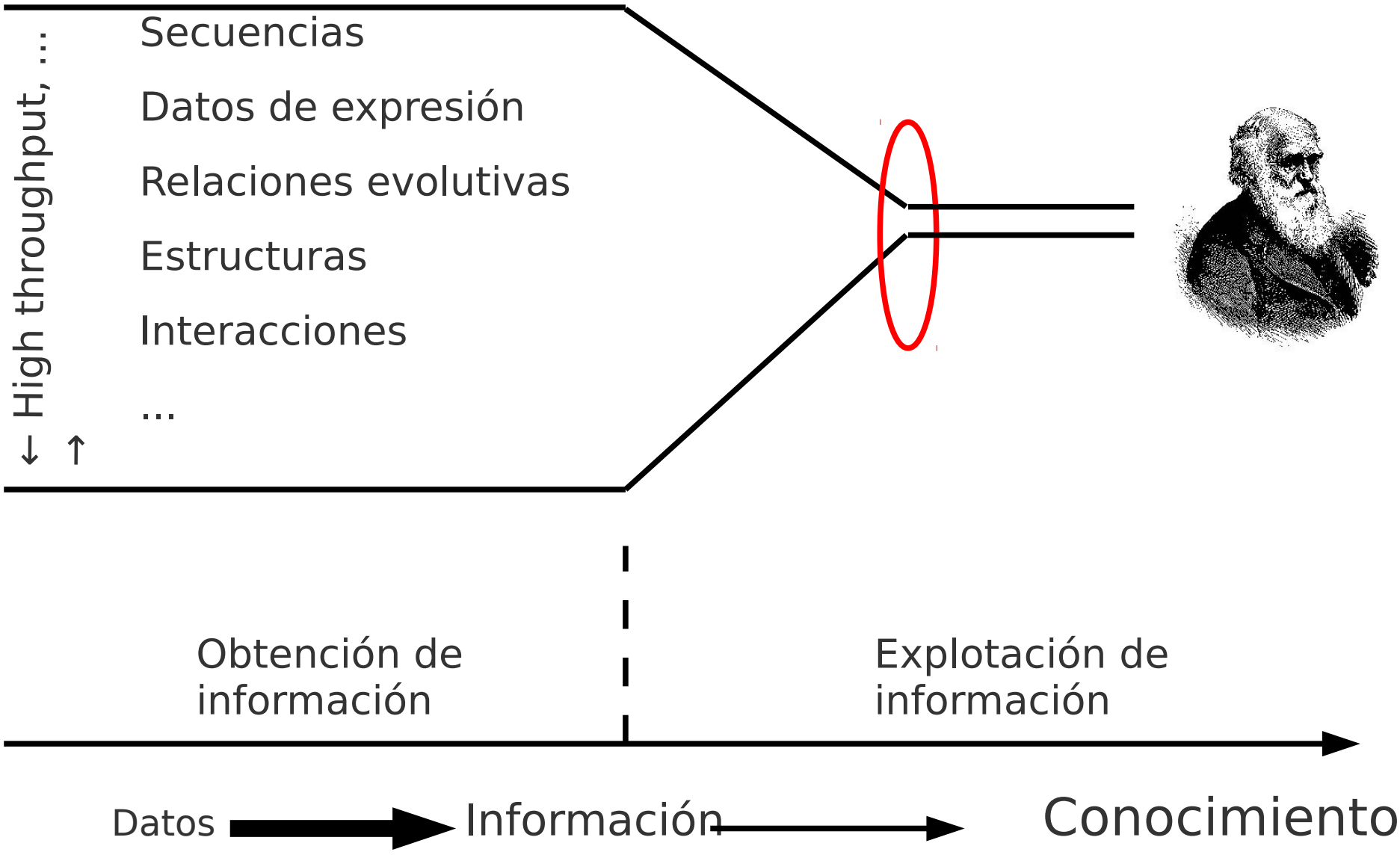
Open Biological and Biomedical Ontologies -OBO- (! LSSW)

¿Por qué no se aplica la LSSW?

Conclusiones

Información en Biología Molecular

Información en Biología Molecular



Información en Biología Molecular

... por la naturaleza misma de la información biológica

- Compleja (¡Y no se puede abstraer!)
- Cambiante
- Producida por muchos agentes diferentes
- Grandes volúmenes
- Crecimiento acelerado (ej. high throughput)

... por cómo hemos representado la información biológica

- Diferentes recursos, diferentes esquemas
- Crisis de identidad
- Para humanos, no para máquinas (Anotaciones, literatura, ...)
- ...

Información en Biología Molecular

"It is quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in a often badly written text and then spend some more millions trying to second guess what the authors really did and found"

Teresa K. Attwood, Douglas B. Kell, Philip McDermott, James Marsh, Steve R. Pettifer, and David Thorne. Calling international rescue: knowledge lost in literature and data landslide! The Biochemical journal, 424(3):317–333, December 2009.

Información en Biología Molecular

Q708Y0 *Saccharomyces cerevisiae*

- Ortólogos de Q708Y0 en *Schizosaccharomyces pombe*
- Si algun ortólogo esta en el núcleo (¡O partes del núcleo!), las proteínas que interaccionan con él por fosforilación
- Obtener los procesos de regulación en los que participan esas proteínas
- ¿Afecta alguno de esos procesos al ciclo celular? ¿Cuál?

Información en Biología Molecular

..... A is part of B ....
....
.....X has part Y ....
M has parts only M and N ...
....
.....F appears only in H .....
....

....
....
... . .PR 2 is located in A ...
....

....
....
PR1 is ortholog of PR2 ...
....

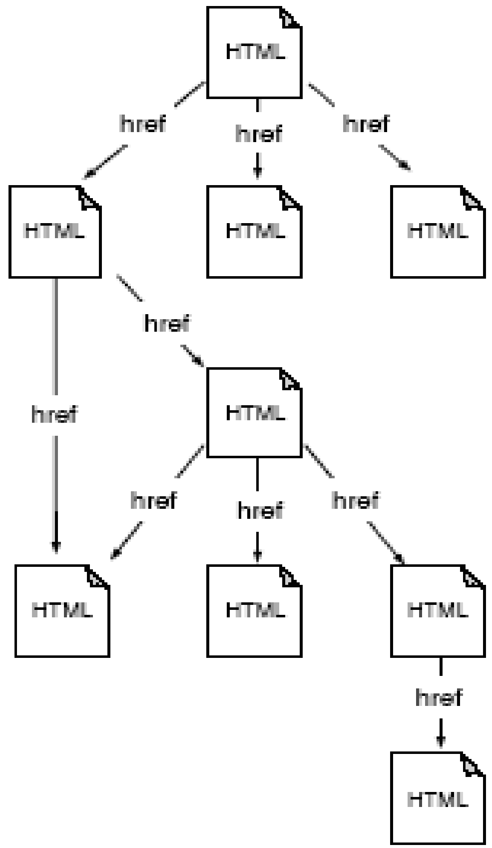
....
....
..... B is part of C ...
....

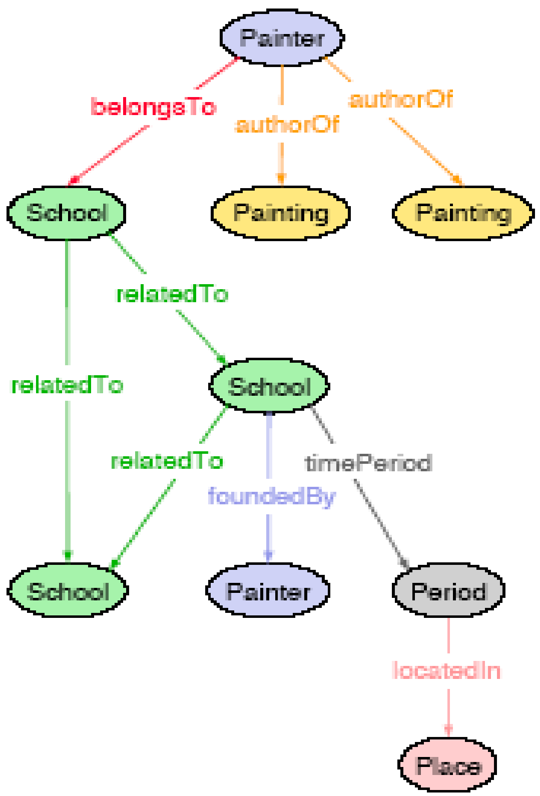
Grandes volúmenes de información desperdigados en la web

Life Sciences Semantic Web

Life Sciences Semantic Web

Web Semántica

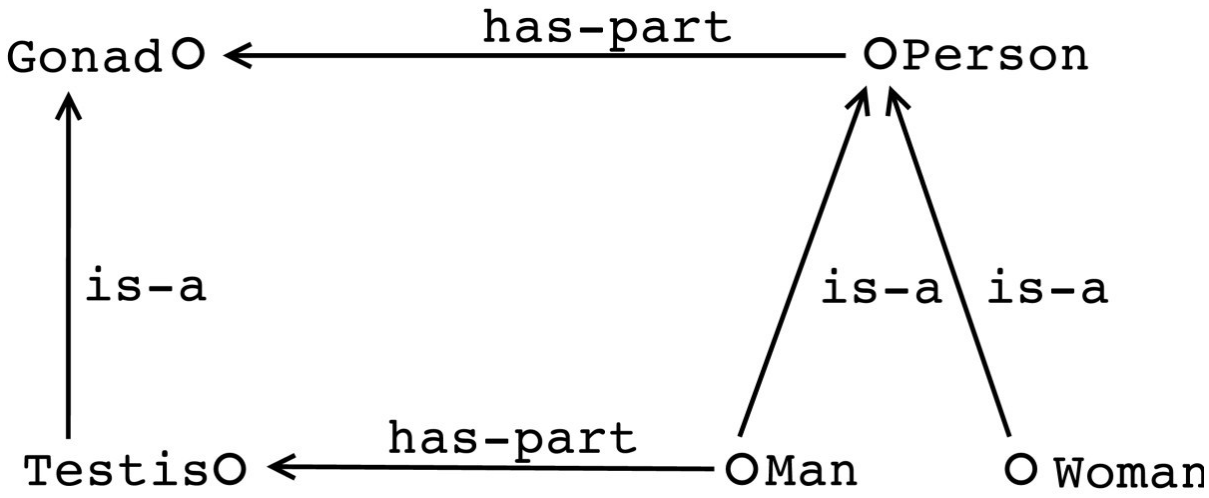




W3C® Semantic Web

Life Sciences Semantic Web

Ontología



Life Sciences Semantic Web

Biología Molecular (Bioinformática)

- Necesidades muy concretas de gestión de información
- Muchos usuarios dispuestos a crear contenido semántico

“Test case” perfecto para la Web Semántica

- W3C Health Care and Life Sciences Interest Group

http://www.w3.org/blog/hcls

Life Sciences Semantic Web

Globally unique and resolvable names for biological entities

Consistent standards for data representation

Consistent standards for knowledge representation

Standard interface definitions for data retrieval and processing

Benjamin M. Good and Mark D. Wilkinson. The life sciences semantic web is full of creeps! Brief Bioinform, 7(3):275-286, September 2006.

Life Sciences Semantic Web

Globally unique and resolvable names for biological entities: **LSID(?)**

Consistent standards for data representation: **RDF**

Consistent standards for knowledge representation: **OWL**

Standard interface definitions for data retrieval and processing: Semantic Web Services

Benjamin M. Good and Mark D. Wilkinson. The life sciences semantic web is full of creeps! Brief Bioinform, 7(3):275-286, September 2006.

Life Sciences Semantic Web

LSID (Life Sciences Identifiers): URNs

`urn:lsid:ipni.org:names:30000959-2`

URIs

Bio2RDF (BANFF manifesto)

`http://tinyurl.com/39m9qru`

HCLS IG

`http://www.w3.org/2001/sw/hcls/notes/uris/`

Neurocommons

`http://neurocommons.org/page/URIs`

Shared Names

`http://neurocommons.org/page/Shared_names`

...

Life Sciences Semantic Web

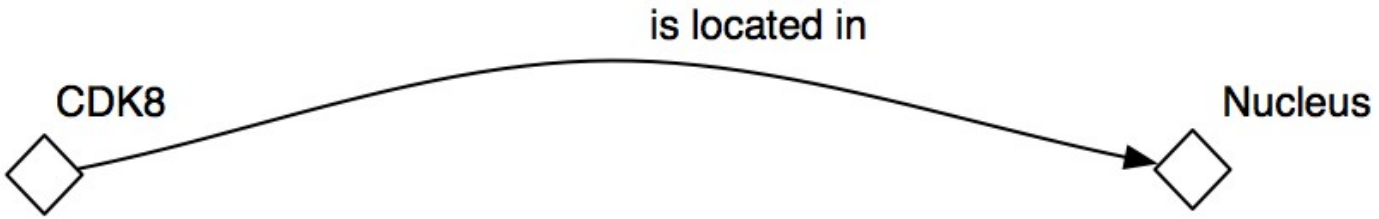
RDF (Resource Description Framework)

- Standard para representar datos en la WS

  `http://www.w3.org/RDF/`

- SPARQL para consultas

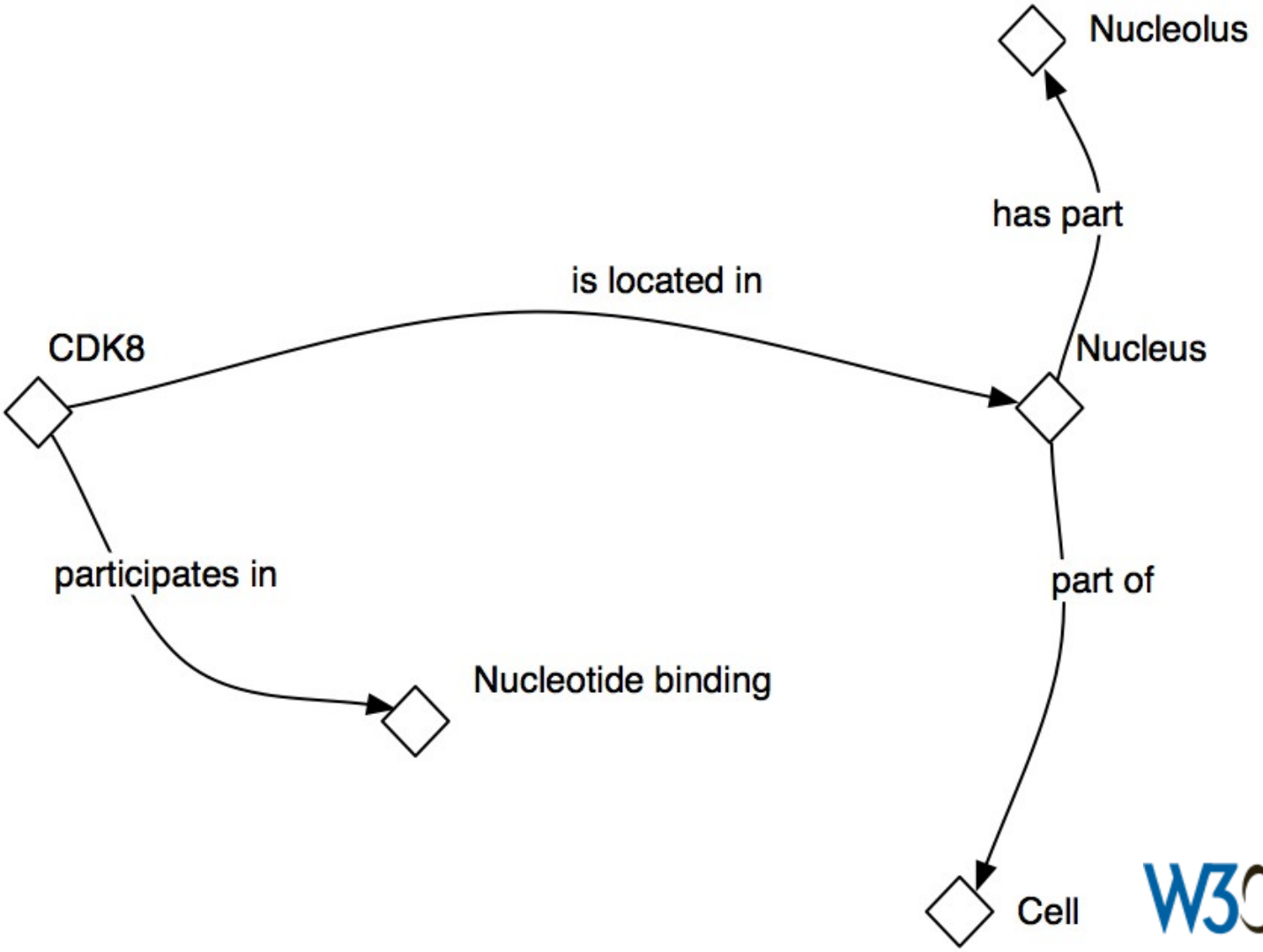
  `http://www.w3.org/TR/rdf-sparql-query/`



W3C Semantic Web

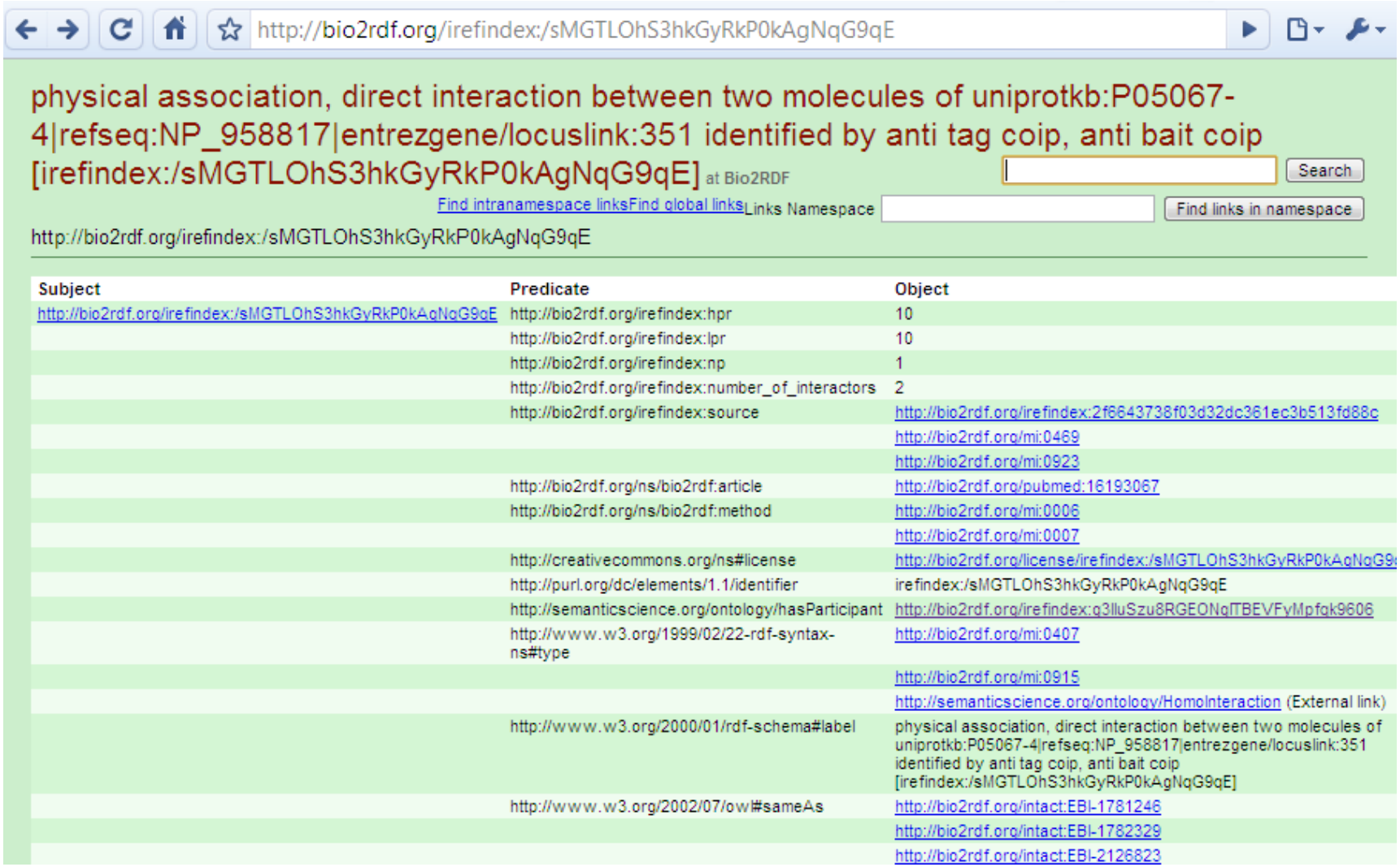
Life Sciences Semantic Web

RDF (Resource Description Framework)



Life Sciences Semantic Web

Bio2RDF `http://bio2rdf.org/`

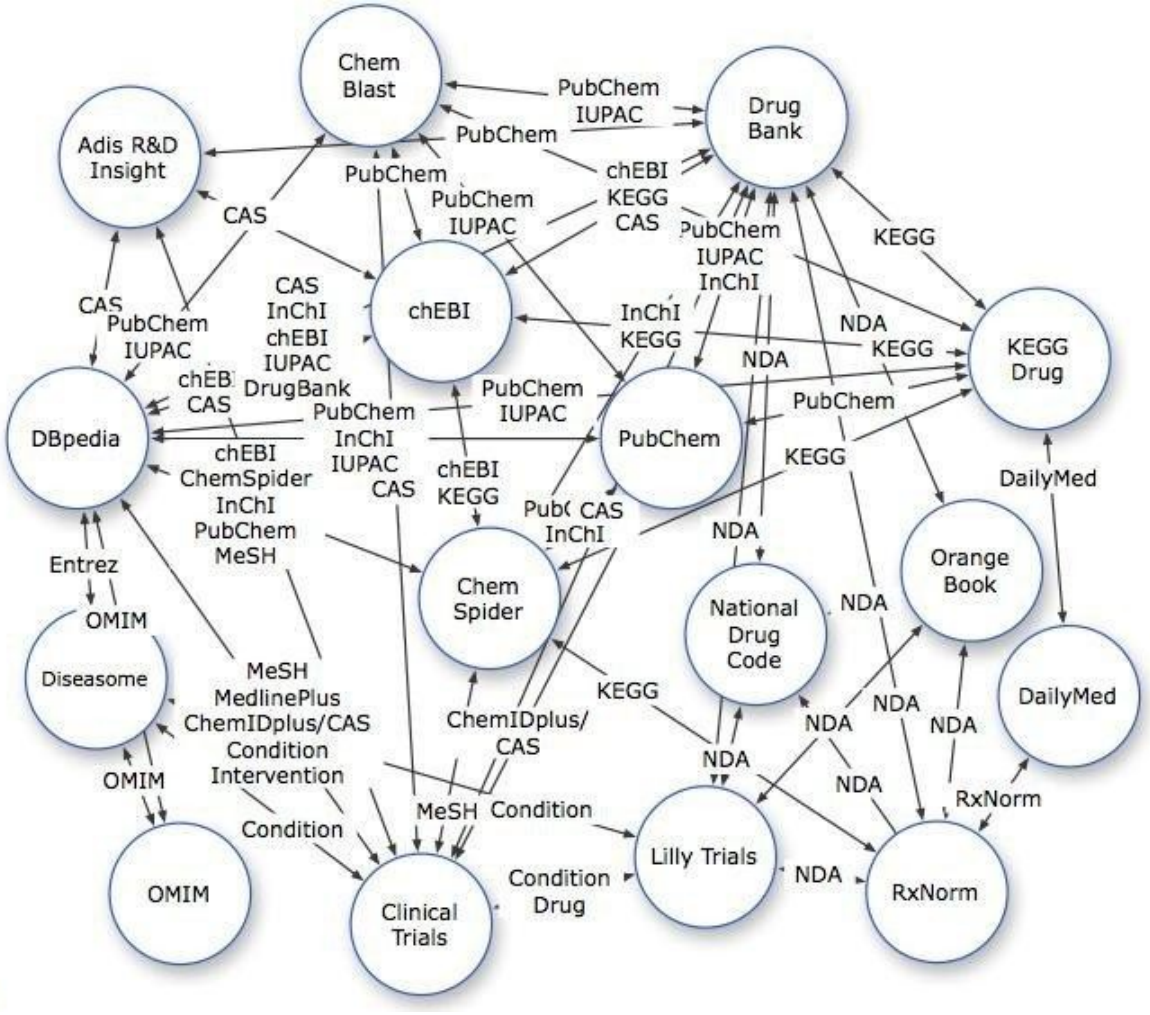


Subject	Predicate	Object
http://bio2rdf.org/irefindex:/sMGTLOhS3hkGyRkP0kAgNqG9qE	http://bio2rdf.org/irefindex:hpr	10
	http://bio2rdf.org/irefindex:lpr	10
	http://bio2rdf.org/irefindex:np	1
	http://bio2rdf.org/irefindex:number_of_interactors	2
	http://bio2rdf.org/irefindex:source	http://bio2rdf.org/irefindex:2f6643738f03d32dc361ec3b513fd88c
		http://bio2rdf.org/mi:0469
		http://bio2rdf.org/mi:0923
	http://bio2rdf.org/ns/bio2rdf:article	http://bio2rdf.org/pubmed:16193067
	http://bio2rdf.org/ns/bio2rdf:method	http://bio2rdf.org/mi:0006
		http://bio2rdf.org/mi:0007
	http://creativecommons.org/ns#license	http://bio2rdf.org/license/irefindex:/sMGTLOhS3hkGyRkP0kAgNqG9
	http://purl.org/dc/elements/1.1/identifier	irefindex:/sMGTLOhS3hkGyRkP0kAgNqG9qE
	http://semanticscience.org/ontology/hasParticipant	http://bio2rdf.org/irefindex:q3lluSzu8RGEONqITBEVFyMpfqk9606
	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/mi:0407
		http://bio2rdf.org/mi:0915
		http://semanticscience.org/ontology/HomoInteraction (External link)
	http://www.w3.org/2000/01/rdf-schema#label	physical association, direct interaction between two molecules of uniprotkb:P05067-4|refseq:NP_958817|entrezgene/locuslink:351 identified by anti tag coip, anti bait coip [irefindex:/sMGTLOhS3hkGyRkP0kAgNqG9qE]
	http://www.w3.org/2002/07/owl#sameAs	http://bio2rdf.org/intact:EBI-1781246
		http://bio2rdf.org/intact:EBI-1782329
		http://bio2rdf.org/intact:EBI-2126823

Life Sciences Semantic Web

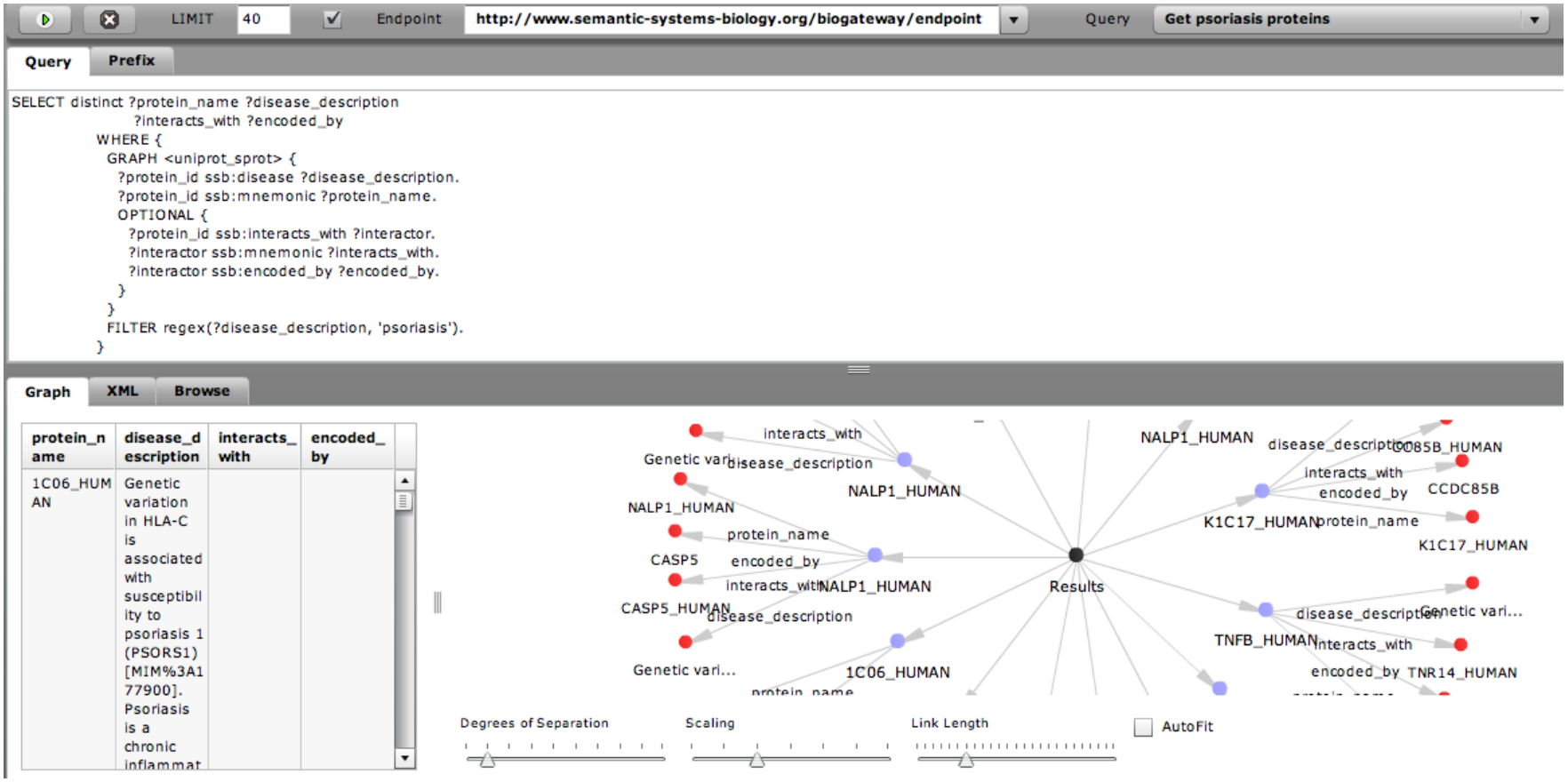
Linking Open Drug Data (LODD) http://esw.w3.org/HCLSIG/LODD

Ganador del triplify challenge http://triplify.org/Challenge/2009



Life Sciences Semantic Web

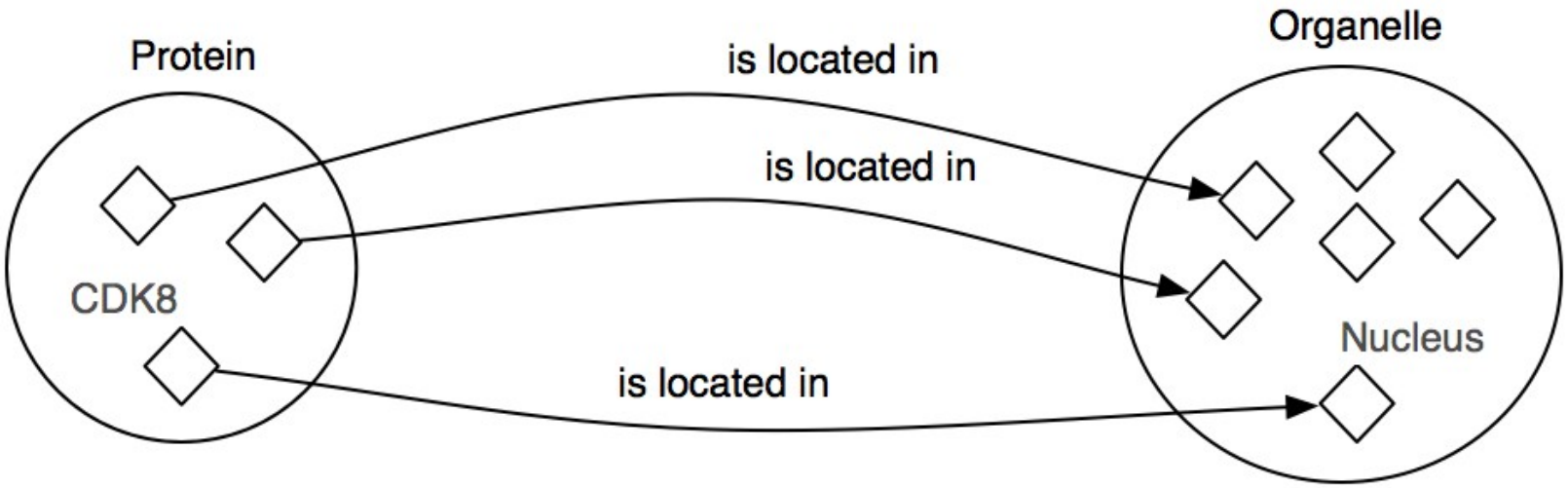
BioGateway http://www.semantic-systems-biology.org/biogateway



Erick Antezana, Ward Blondé, Mikel Egaña, Alistair Rutherford, Robert Stevens, Bernard De Baets, Vladimir Mironov, Martin Kuiper. BioGateway a semantic systems biology tool for the life sciences. BMC bioinformatics 2009, 10(Suppl 10):S11

Life Sciences Semantic Web

OWL (Web Ontology Language)



Life Sciences Semantic Web

- OWL (Web Ontology Language)
  - Autodescriptivo (self-descriptive)
    - datos + esquema en el mismo idioma
  - OWA (Open World Assumption)
    - el conocimiento en biología molecular es necesariamente incompleto
  - Semantica monotónica
  - ~~UNA (Unique Name Assumption)~~
    - diferentes entradas en diferentes recursos se refieren a la misma entidad

Life Sciences Semantic Web

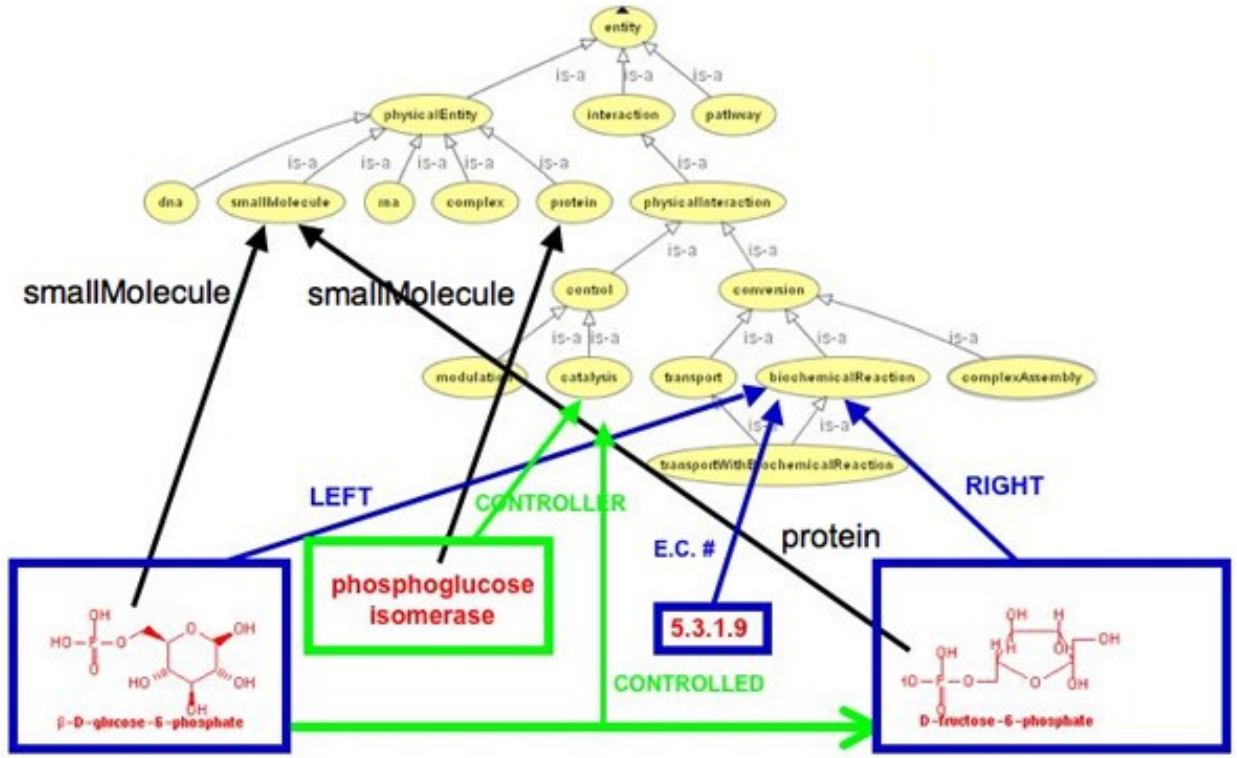
- OWL (Web Ontology Language)
  - Inferencia (“Reasoning”) completa y “eficiente”
    - Consultas
    - Clase-subclase
    - Individuo → clase
    - Consistencia
  - URIs para entidades
  - Editores (Protégé, TopBraid composer, …)
  - APIs (OWL API, …)
  - Razonadores (Pellet, FaCT++, Racer, …)

Life Sciences Semantic Web

- Uso de OWL
  - Vocabulario común
    - almacenar/integrar/reusar conocimiento
  - Inferir conocimiento no evidente
  - Clasificación de información
  - Consultas expresivas
  - Generación de hipótesis
  - Consistencia de la información
  - Al representar un dominio el razonador nos dice las contradicciones que cometemos: nos obliga a definirnos
  - Mantenimiento de conocimiento

Life Sciences Semantic Web

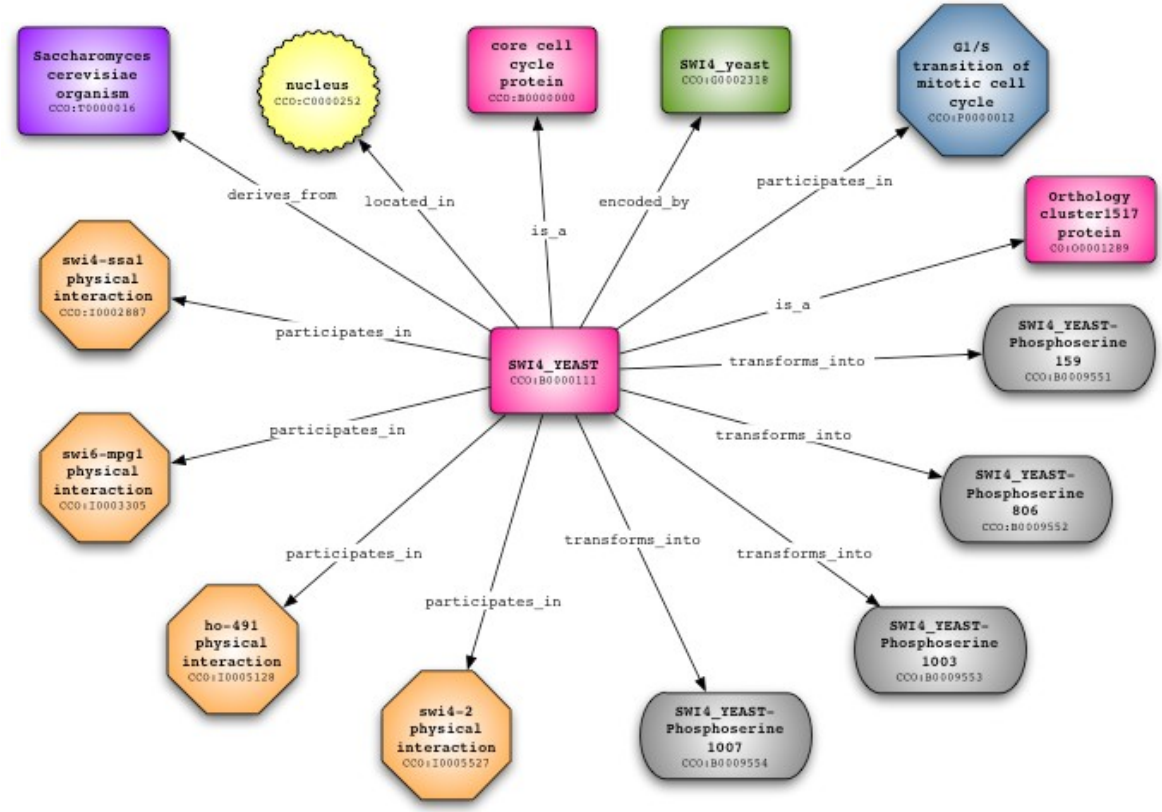
BioPAX `http://www.biopax.org/`



Life Sciences Semantic Web

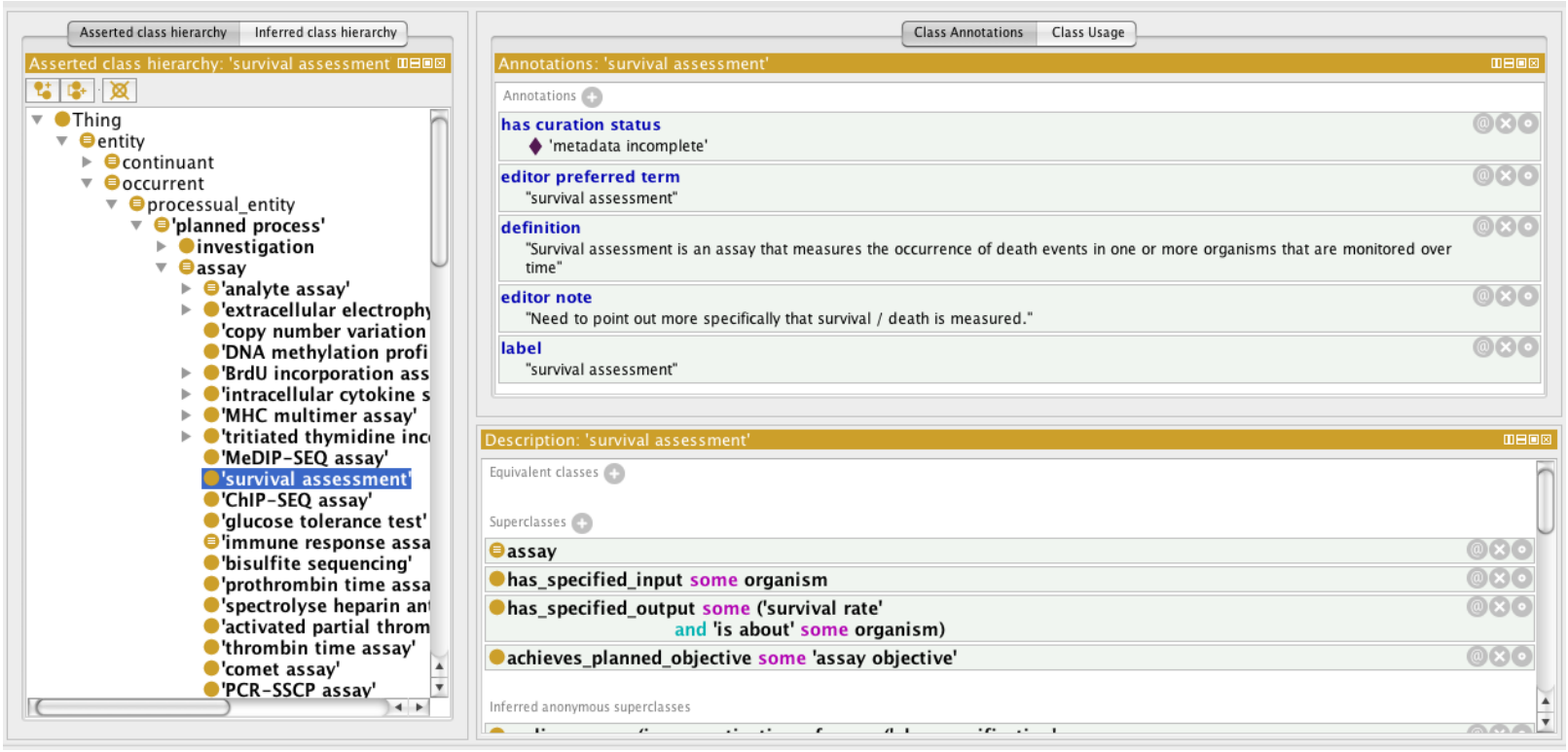
Cell Cycle Ontology http://www.semantic-systems-biology.org/cco

Erick Antezana, Mikel Egaña, Ward Blondé, Aitzol Illarramendi, Iñaki Bilbao, Bernard De Baets, Robert Stevens, Vladimir Mironov, and Martin Kuiper. The cell cycle ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome Biology, 10(5):R58+, 2009



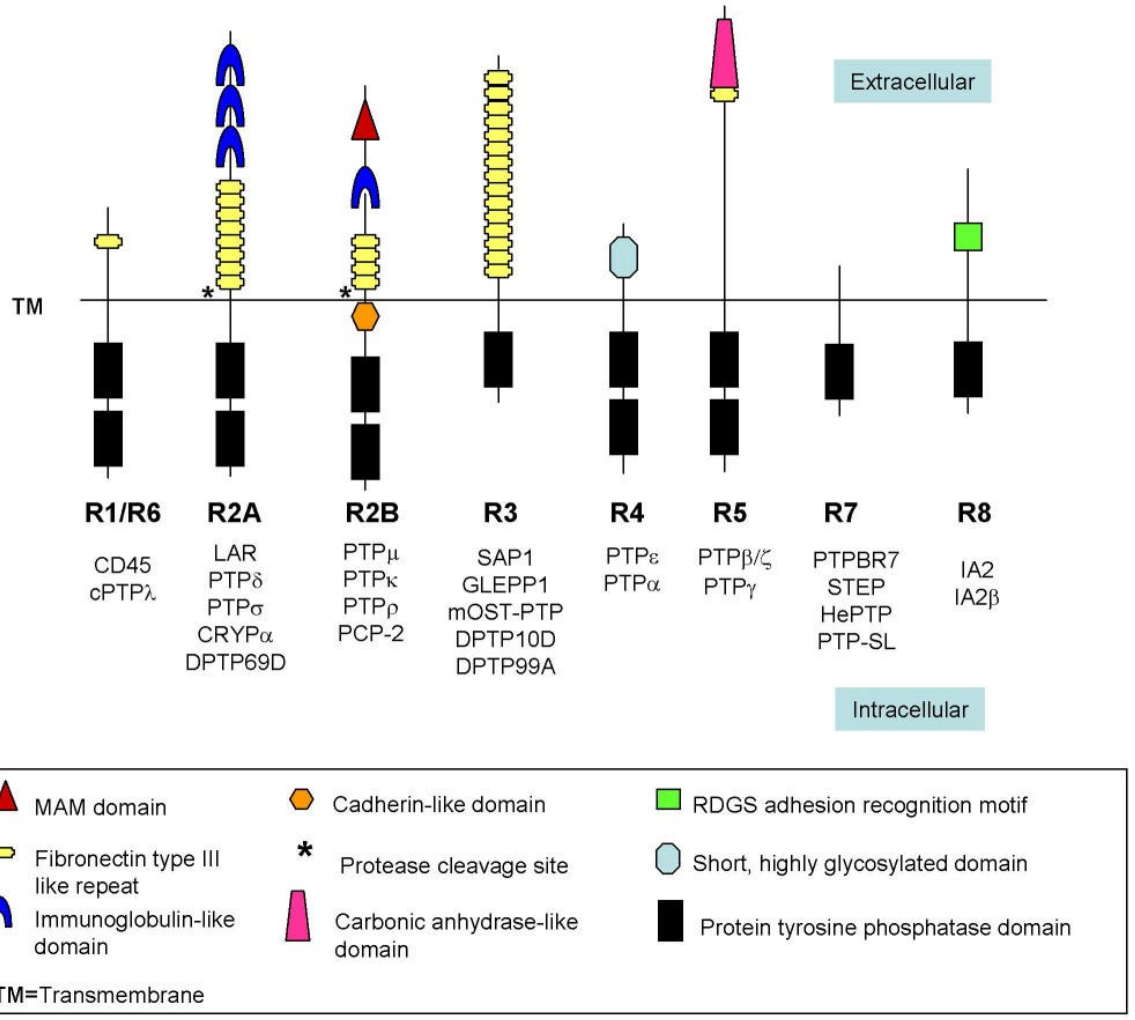
Life Sciences Semantic Web

OBI `http://obi-ontology.org`



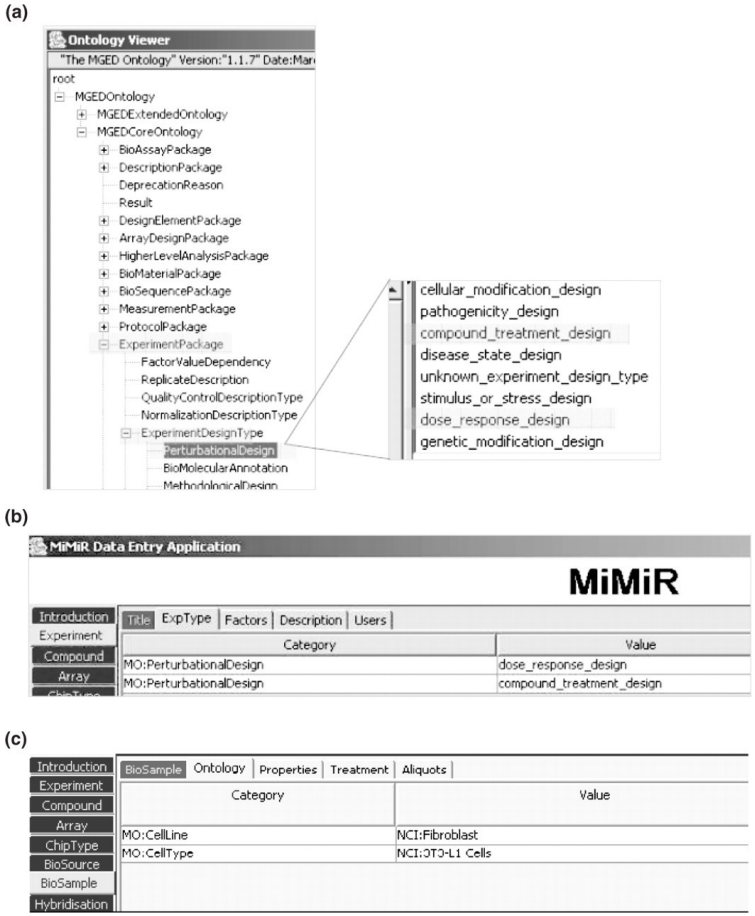
Life Sciences Semantic Web

PhosphaBase http://www.bioinf.manchester.ac.uk/phosphabase/



Life Sciences Semantic Web

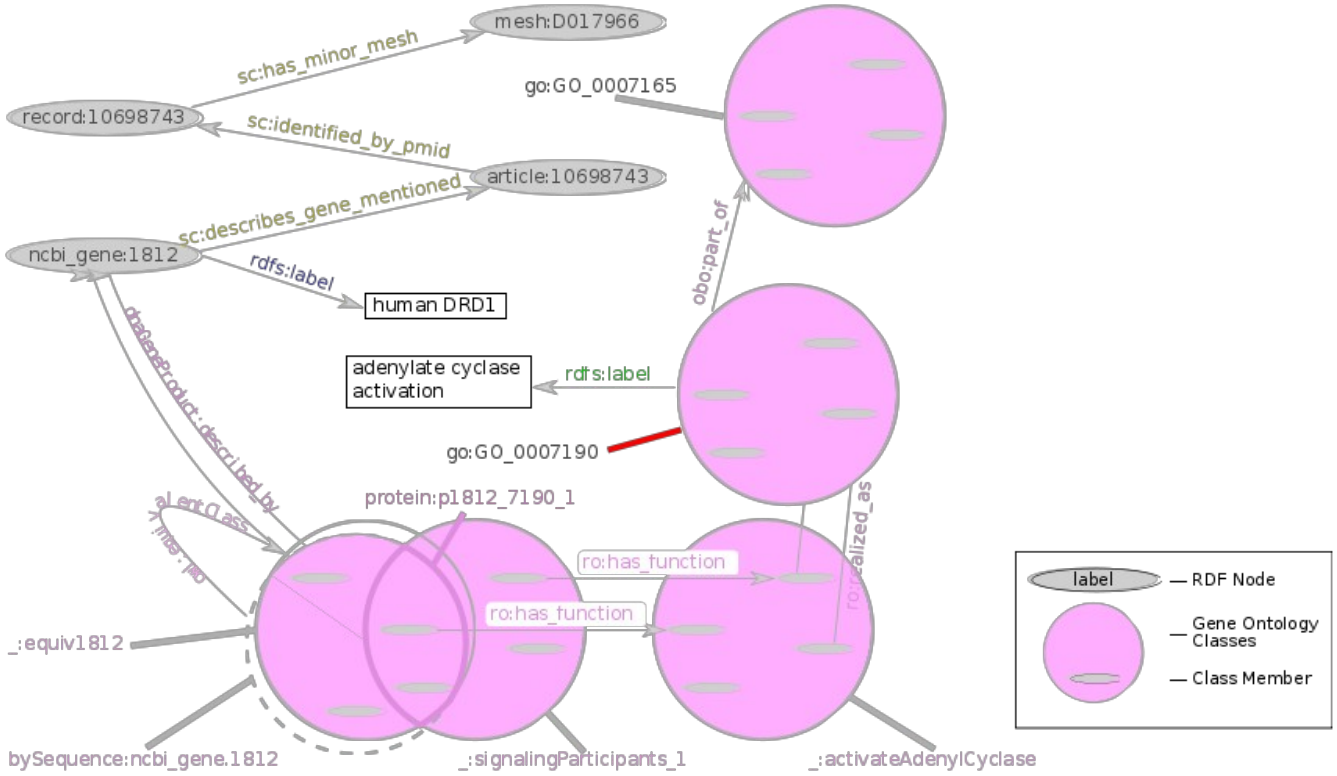
MGED `http://mged.sourceforge.net/`



Life Sciences Semantic Web

A prototype KB for the Life Sciences

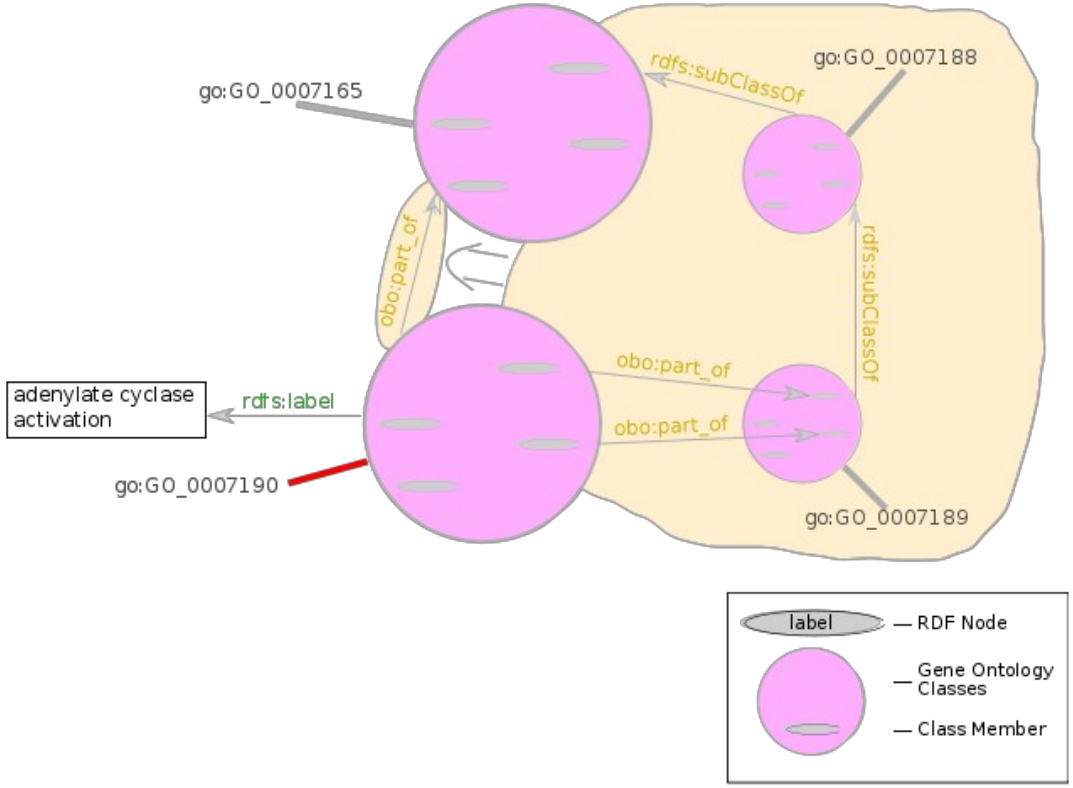
http://www.w3.org/TR/hcls-kb/



Life Sciences Semantic Web

A prototype KB for the Life Sciences

http://www.w3.org/TR/hcls-kb/



Open Biological and Biomedical Ontologies (OBO)

Open Biological and Biomedical Ontologies

OBO Foundry http://www.obofoundry.org/

- Open
- Common shared syntax
- Unique identifier space
- Versions
- Delineated content
- Definitions
- OBO Relation Ontology
- Well documented
- Users
- Collaboratively

Open Biological and Biomedical Ontologies

OBO Foundry

OBO Foundry ontologies

Title	Domain	Prefix	File	Last changed
Biological process	biological process	GO	gene_ontology_edit.obo	2010/08/29
Cellular component	anatomy	GO	gene_ontology_edit.obo	2010/08/29
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo	2010/08/03
Molecular function	biological function	GO	gene_ontology_edit.obo	2010/08/29
Phenotypic quality	phenotype	PATO	quality.obo	2010/08/19
PRotein Ontology (PRO)	proteins	PRO	pro.obo	2010/08/20
Xenopus anatomy and development	anatomy	XAO	xenopus_anatomy.obo	2009/12/02
Zebrafish anatomy and development	anatomy	ZFA	zebrafish_anatomy.obo	2010/08/06

OBO Foundry candidate ontologies and other ontologies of interest

Title	Domain	Prefix	File	Last changed
Amphibian gross anatomy	anatomy	AAO	amphibian_anatomy.obo	2008/06/19
Amphibian taxonomy	anatomy	ATO	amphibian_taxonomy.obo	
Ascomycete phenotype ontology	phenotype	APO	ascomycete_phenotype.obo	2010/05/12

Open Biological and Biomedical Ontologies

Gene Ontology http://geneontology.org/

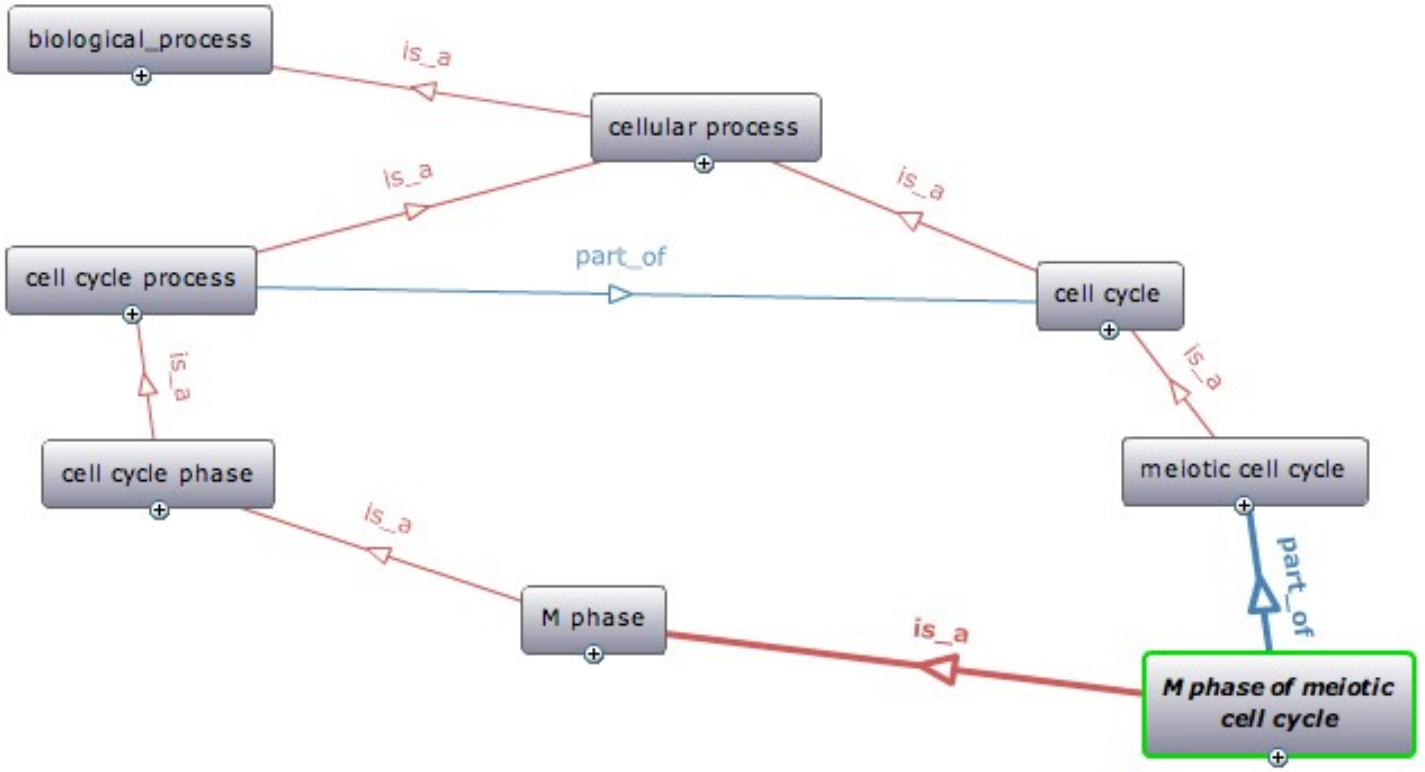
The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. (2000) 25: 25-29

Michael Bada, Robert Stevens, Carole Goble, Yolanda Gil, Michael Ashburner, Judith A. Blake, J. Michael Cherry, Midori Harris, and Suzanna Lewis. A Short Study on the Success of the Gene Ontology . Web Semantics: Science, Services and Agents on the World Wide Web, 1(2):235-240, 2004.

- Vocabulario controlado para describir la función molecular, el componente celular y el proceso biológico de genes (“Gene Products”)

Open Biological and Biomedical Ontologies

Gene Ontology: ~ 32.500 términos en una estructura `is_a`, `part_of`, `regulates` (+,–)

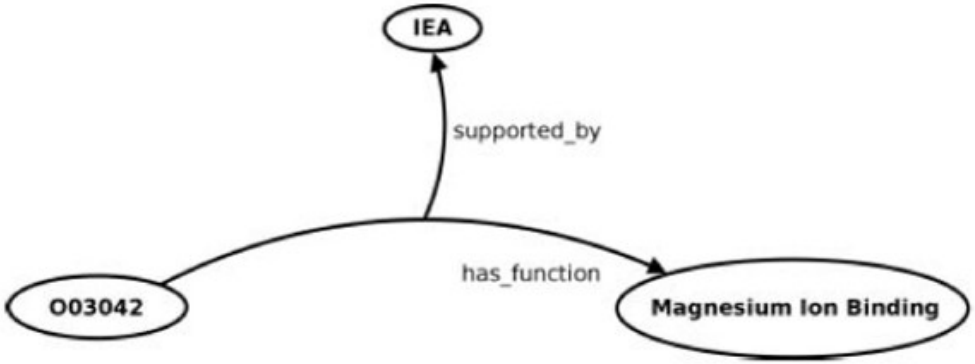


Open Biological and Biomedical Ontologies

Gene Ontology

Integración de recursos (anotaciones)

Gene Ontology Annotation (GOA)



Explotación de la estructura

Ontological analysis of gene expression data: current tools, limitations, and open problems.Bioinformatics. 2005 Sep 15;21(18):3587-95. Epub 2005 Jun 30.

Open Biological and Biomedical Ontologies

Otras ontologías importantes: ChEBI, Cell Type, Sequence Ontology, Phenotype Ontology, UberOntology, …

“Meta Ontologías”

- Basic Formal Ontology (BFO)
- OBO Relation Ontology (RO)

OBO Foundry tiene mucho contenido de relativamente alta calidad, pero …

… la mayoría de las ontologías OBO son muy pobres axiomáticamente

… OBO format

Open Biological and Biomedical Ontologies

OBO Format

- No tiene una definición semántica clara: ¿Qué quieren decir las expresiones en OBO?

  Mikel Egaña Aranguren, Sean Bechoffer, Phillip Lord, Ulrike Sattler and Robert Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics 2007, 8:57

- Para usar inferencias, traducir a OWL ...

  Christine Golbreich, Matthew Horridge, Ian Horrocks, Boris Motik, and Rob Shearer. OBO and OWL: Leveraging semantic web technologies for the life sciences. ISWC 2007, 4825:169–182, 2007

- ... o usar razonadores “ad hoc”:
  - OBO Edit reasoner
  - OBD-SQL reasoner
  - OBO Language (OBOL)

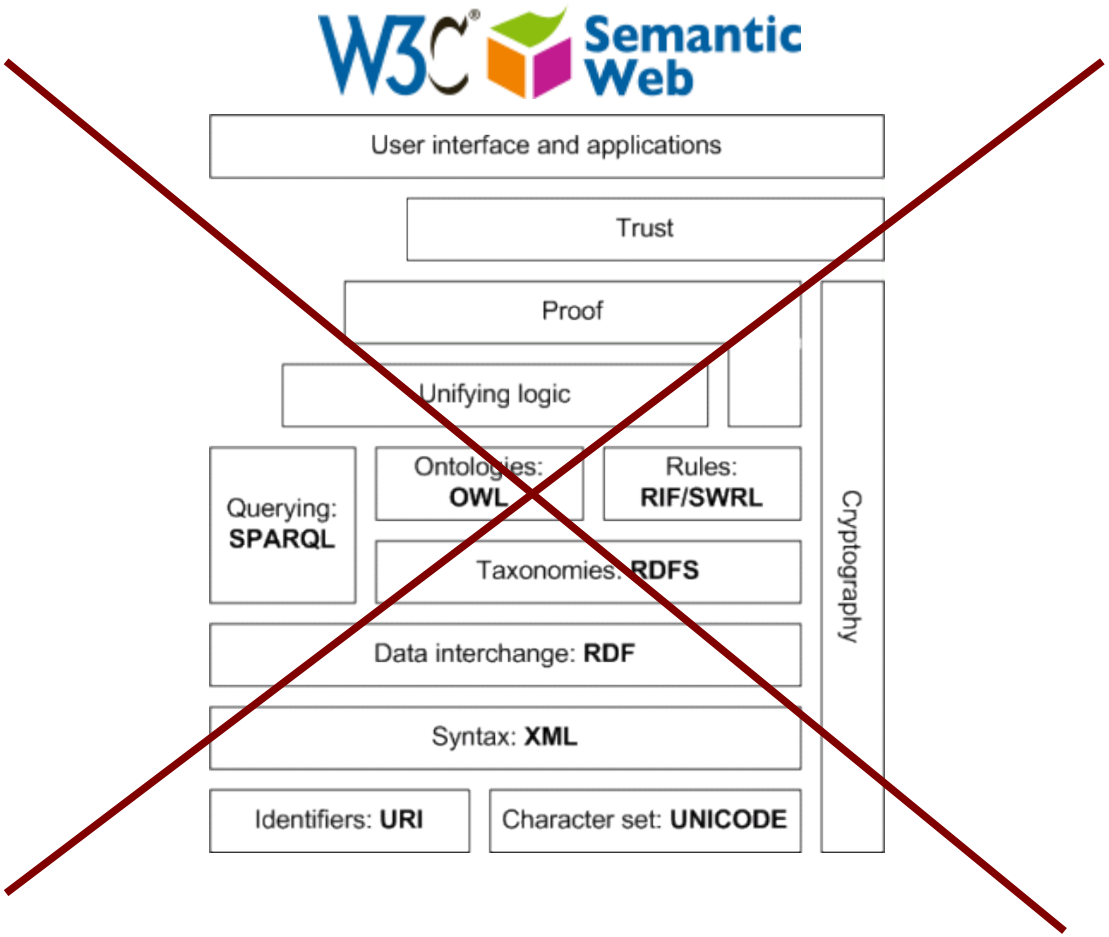
Open Biological and Biomedical Ontologies

OBO Format

```
[Term]
id: CCO:B0002060
name: NEB2_HUMAN
def: "Neurabin-2" [UniProt:Q96SB3]
synonym: "Neurabin-II" EXACT [UniProt:Q96SB3]
xref: UniProt:Q8TCR9
is_a: CCO:B0000000 ! core cell cycle protein
relationship: belongs_to CCO:T0000004 ! Homo sapiens organism
relationship: encoded_by CCO:G0005171 ! PPP1R9B_human
relationship: participates_in CCO:I0006401 ! aah62584-q96sb3 physical interaction
relationship: transforms_into CCO:B0013139 ! NEB2_HUMAN-Phosphoserine15
```

Open Biological and Biomedical Ontologies

OBO format



¿Por qué no se aplica LSSW?

¿Por qué no se aplica LSSW?

Problema general en bioinformática: los biólogos crean sus propias soluciones artesanalmente y con afán de publicar

- Carole Goble, The Seven Deadly Sins of Bioinformatics:

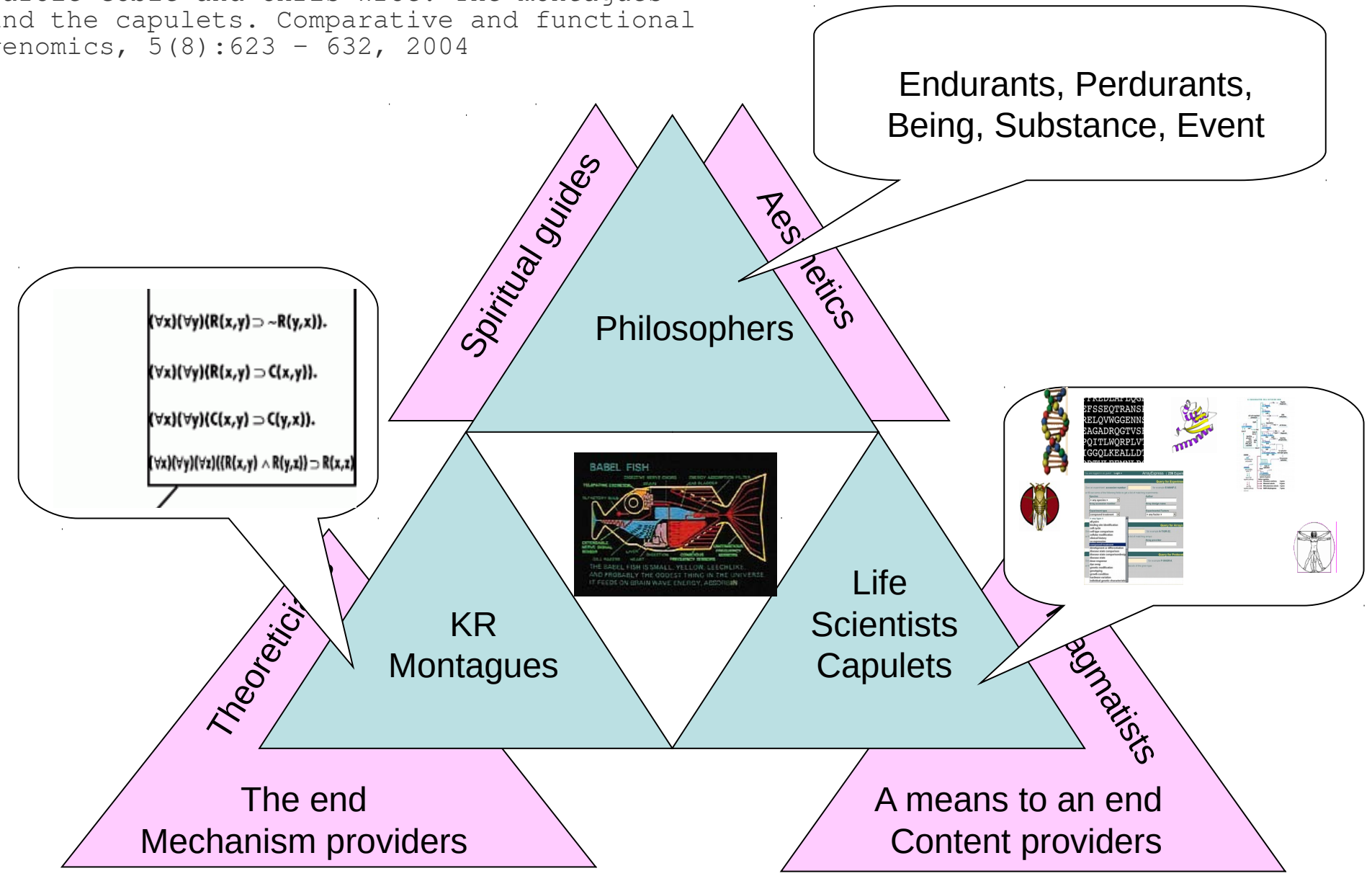
  `http://tinyurl.com/6nvoe4`

OBO es un ejemplo más de “artesanía”

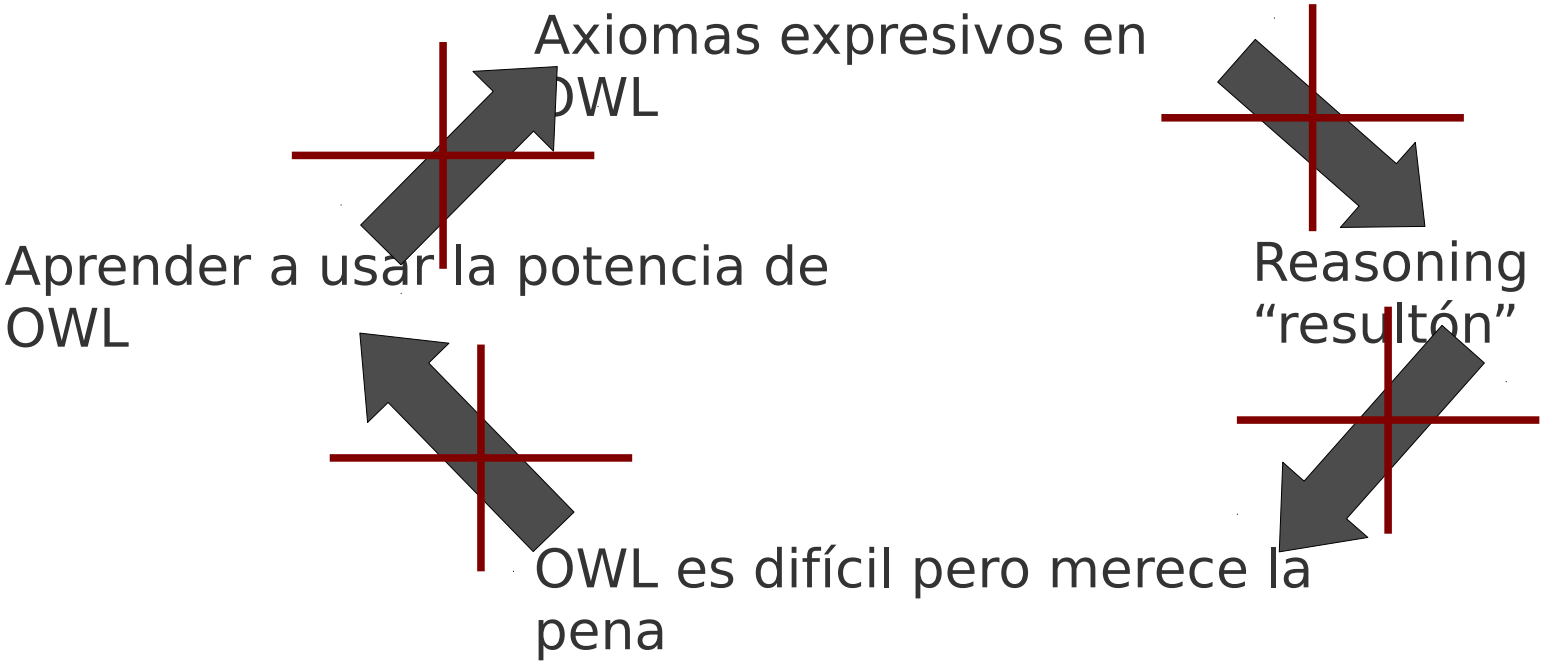
- Cuando surgió GO, OWL no existía (y mucho menos Protégé)
- OWL es anti-intuitivo (ej. OWA, !UNA, …)

¿Por qué no se aplica LSSW?

Carole Goble and Chris Wroe. The montagues and the capulets. Comparative and functional genomics, 5(8):623 – 632, 2004



¿Por qué no se aplica LSSW?



Las ontologías se diseñan para consumo humano (ej. Muchos axiomas "enterrados" en anotaciones en GO)

La prioridad es la integración de recursos, no la representación de conocimiento

¿Por qué no se aplica LSSW?

Ingeniería ontológica vs. ingeniería software

DEXA 2010 (Deusto). Philosophy Goes Information Technology - A Critical Reflection on Ontologies. Nick Falkner, University of Adelaide, Australia.

Conclusiones

Conclusiones

Para que LSSW se implante

- IDs (Shared Names?)
- Bio-ontologías axiomáticamente ricas (más funcionalidades): best practices (Ontology Design Patterns, ...)

  Mikel Egaña, Alan Rector, Robert Stevens, Erick Antezana. Applying Ontology Design Patterns in bio-ontologies. EKAW 2008, LNCS 5268, pp. 7-16

- Bio-ontologías *consensuadas*:
  - Upper Level Ontology
  - RO
- Menos “Realism”: bio-ontologías como medio, no como fin.

  Phillip Lord and Robert Stevens. Adding a little reality to building ontologies for biology. PLoS ONE, 5(9):e12258, 2010.

Conclusiones

Para que LSSW se implante:

- Herramientas
  - ¡OBO se adoptó por OBO-Edit!
- Reasoning eficiente, predecible, y comprensible
- Demostración de beneficios inmediatos de reasoning
  - Consultas
  - Generación de hipótesis
  - Mantenimiento

Conclusiones

Reasoning comprensible

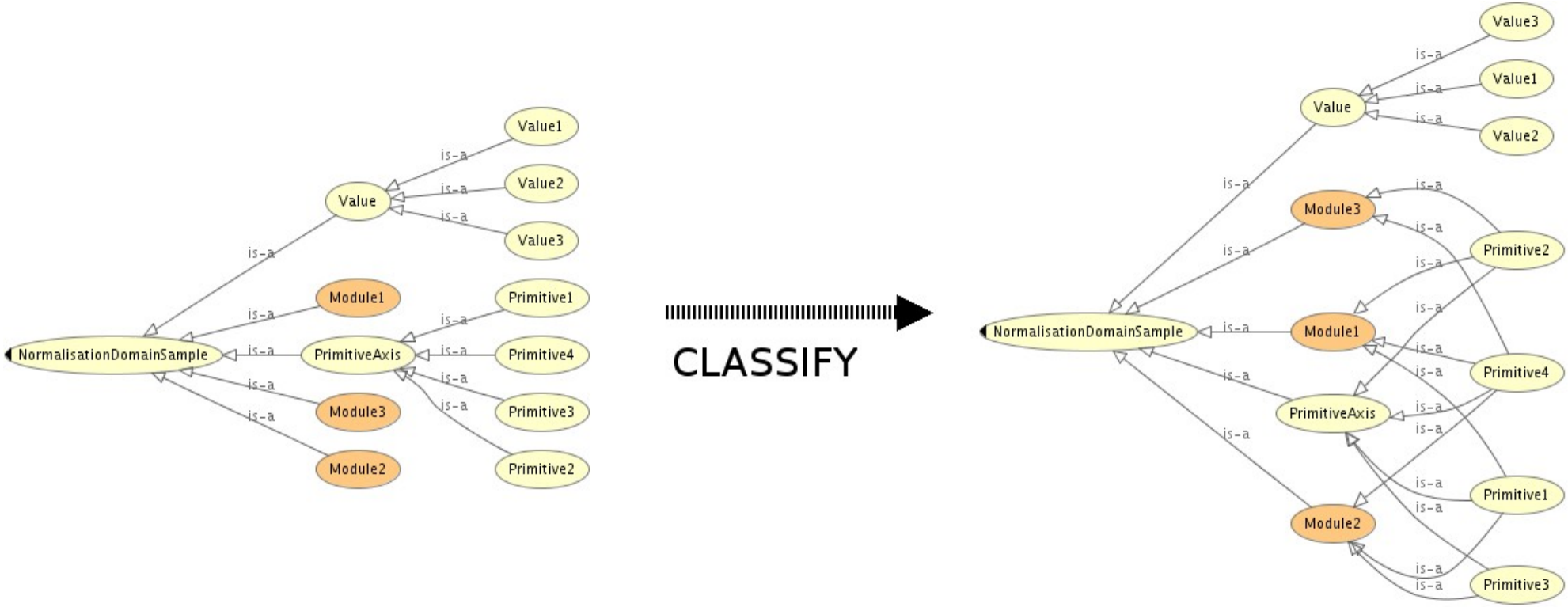
http://owl.cs.manchester.ac.uk/explanation/

	Employee **subClassOf** Person
1)	Employee **subClassOf** worksFor **some** Thing
2)	worksFor **subPropertyOf** memberOf
3)	memberOf **subPropertyOf** **inv** (member)
4)	member **range** Person

Conclusiones

Reasoning para mantenimiento

http://www.gong.manchester.ac.uk/odp/html/Normalisation.html

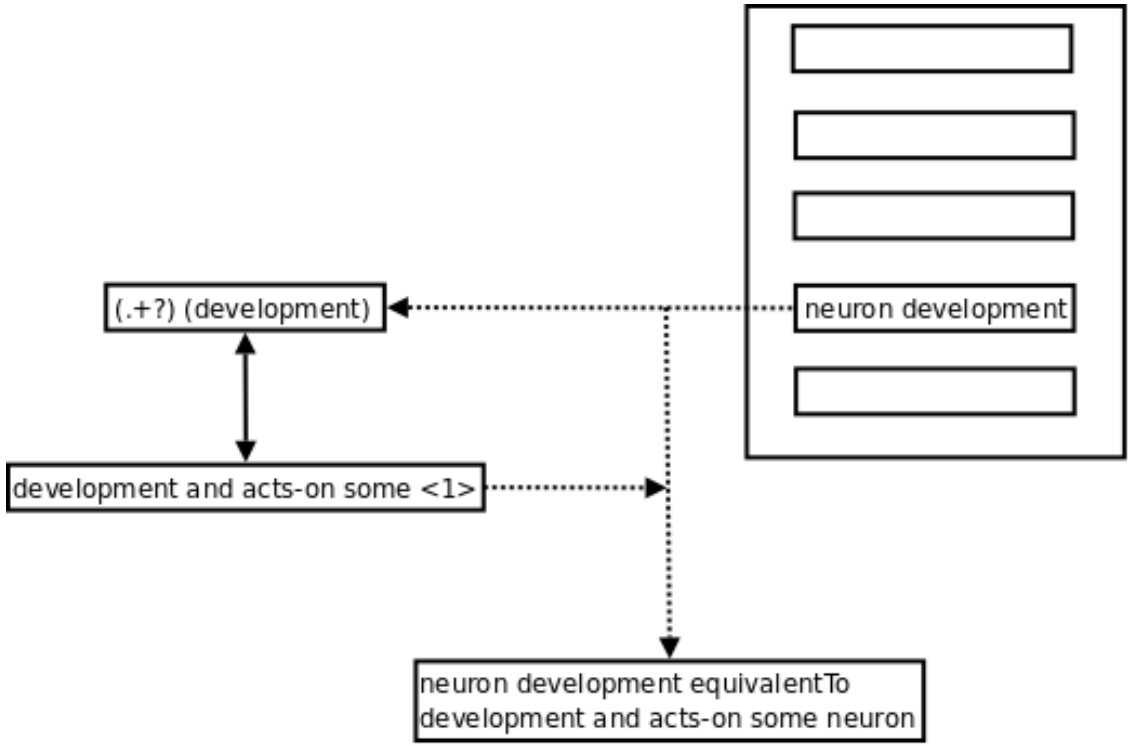


Open Biological and Biomedical Ontologies

Reasoning para mantenimiento y consultas

Gene Ontology Next Generation (GONG): más axiomas con mínimo esfuerzo

Mikel Egaña Aranguren, Chris Wroe, Carole Goble, Robert Stevens. In situ migration of handcrafted ontologies to Reason-able Forms. Data & Knowledge Engineering 2008, 66, 147-162



Conclusiones

Life Sciences Semantic Web:
¿“Killer app” de Semantic Web?

Para saber más ...

Para saber más ...

Bio-ontologies SIG at ISMB (Intelligent Systems for Molecular Biology) http://www.bio-ontologies.org.uk/

SWAT4LS (Semantic Web Applications and Tools for Life Sciences) http://www.swat4ls.org/

ICBO (International Conference on Biomedical Ontology) http://icbo.buffalo.edu/

Para saber más ...

EBI (European Bioinformatics Institute) http://www.ebi.ac.uk/

NCBO (National Center for Biomedical Ontology) http://www.bioontology.org/

Journal of Biomedical Semantics http://www.jbiomedsem.com/

Semantic Systems Biology http://www.semantic-systems-biology.org/

Repositorios:

http://bioportal.bioontology.org/

http://obo.sf.net/

http://www.ebi.ac.uk/ontology-lookup/

Mi trabajo en todo esto (Hasta ahora)

Mi trabajo en todo esto (hasta ahora)

- Hacer SW más fácil a los biólogos
  - Ontology Design Patterns
  - OPPL 2
  - GONG
  - Tutoriales OWL
- Construir recursos que usan SW
  - Cell Cycle Ontology
  - BioGateway
  - OGO

`http://mikeleganaaranguren.wordpress.com/publications/`

Agradecimientos

Inspiración para esta presentación

Michel Dumontier

Jesualdo Tomás Fernández Breis

Carlos Tejo

Erick Antezana

Phil Lord

Robert Stevens