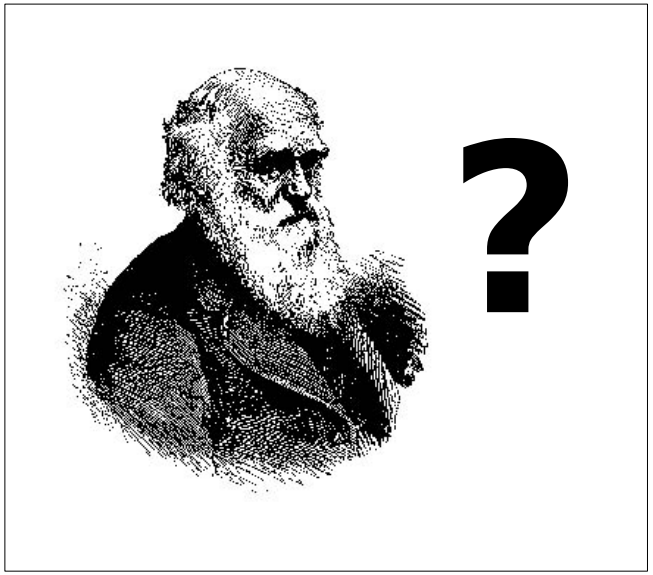
Aplicación de la Web Semántica en Bioinformática

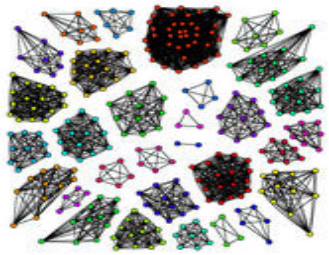
http://mikeleganaaranguren.com/

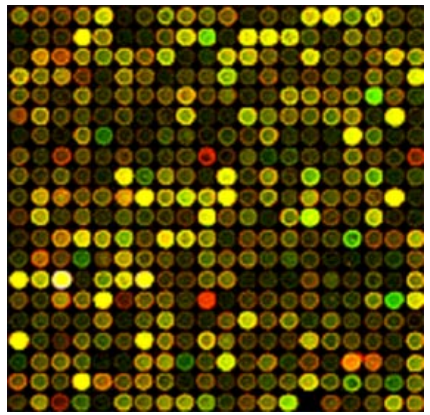
Mikel Egaña Aranguren
mikel.egana.aranguren@gmail.com

School of Computer Science
University of Manchester

El problema ...







IntAct

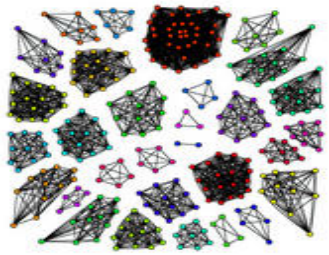
PDB PROTEIN DATA BANK

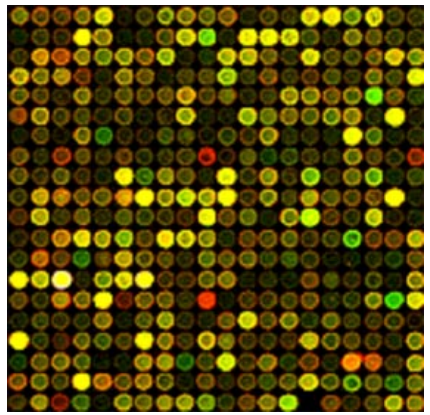
UniProt
the universal protein resource

BioMed CENTRAL

... la solución

?





IntAct

PDB PROTEIN DATA BANK

UniProt the universal protein resource

BioMed CENTRAL

¡Web Semántica!

Bioinformática actual

La Web Semántica

Implantación de la Web Semántica en la bioinformática actual

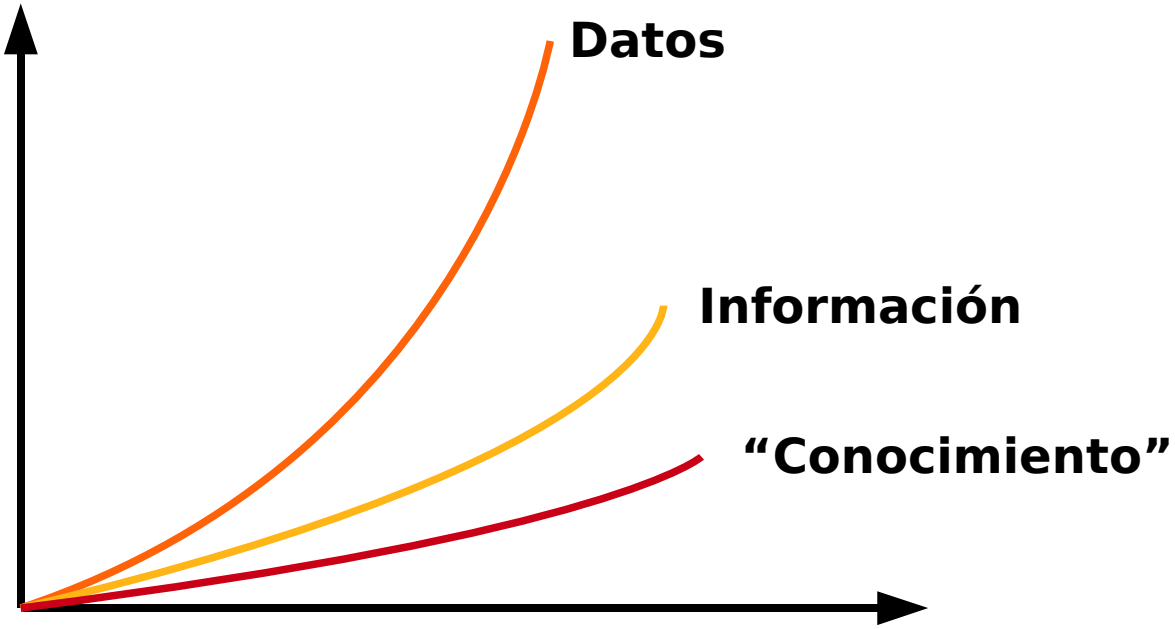
Bioinformática actual: información vs conocimiento

De la biología molecular a la bioinformática

Datos
Secuencias, estructuras, microarrays ...

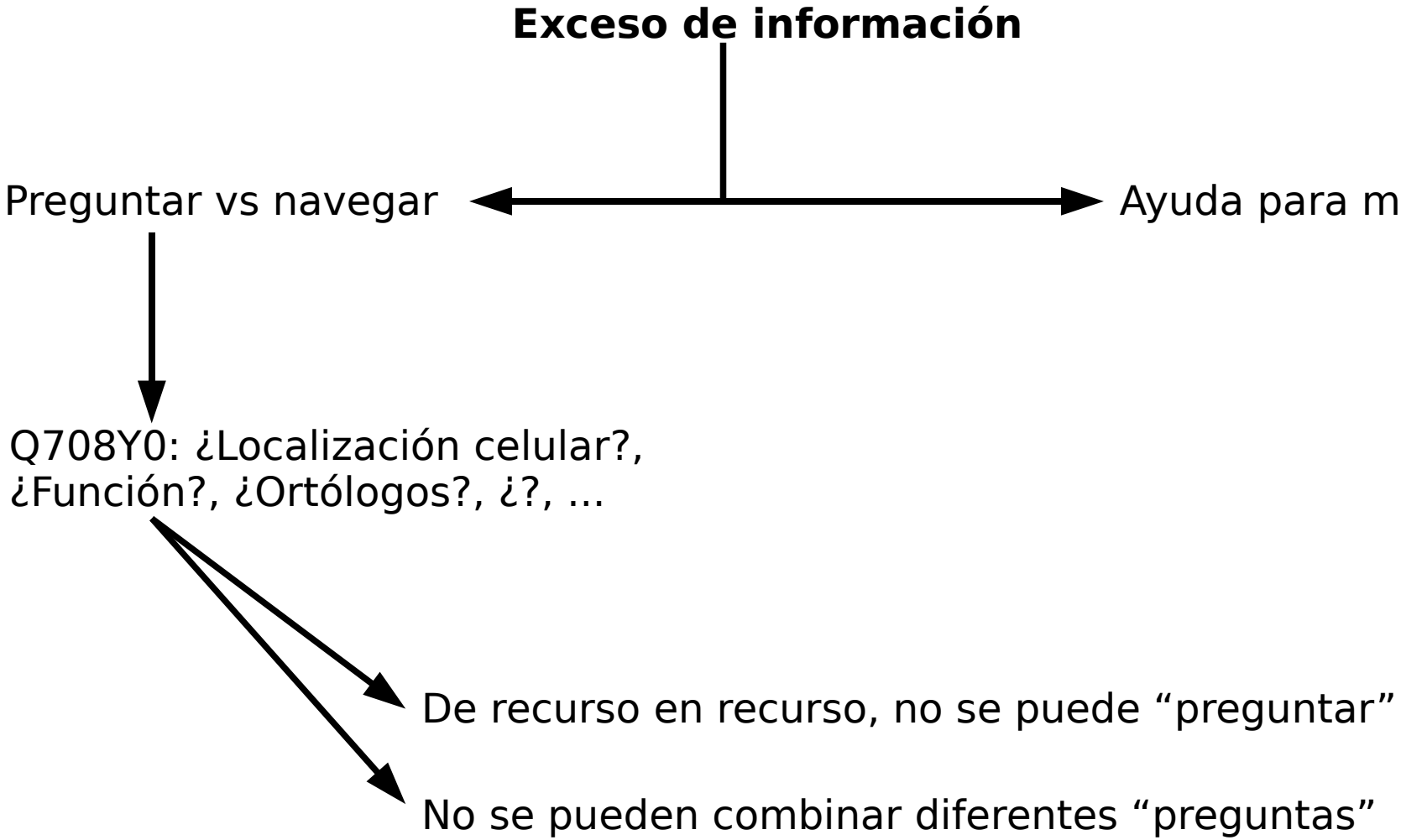
Información
Anotaciones: función, interacciones, localización celular, relaciones evolutivas, ...

“Conocimiento”
¿Cuál es la función y la localización
de los ortólogos provenientes de *A. thaliana* de mi proteina?



Problemas de la bioinformática actual

Miller CJ, Attwood TK. Bioinformatics goes back to the future.
Nat Rev Mol Cell Biol. 2003 Feb;4(2):157-62.



Anotaciones

UniProtKB/Swiss-Prot entry Q708Y0

Submit update
Quick BlastP search
Entry history

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**EBF2_ARATH**
Primary accession number	**Q708Y0**
Secondary accession numbers	None
Integrated into Swiss-Prot on	January 23, 2007
Sequence was last modified on	July 5, 2004 (Sequence version 1)
Annotations were last modified on	April 29, 2008 (Entry version 41)
Name and origin of the protein	
Protein name	**EIN3-binding F-box protein 2**
Synonyms	None
Gene name	**Name: EBF2** OrderedLocusNames: At5g25350 ORFNames: F18G18.90
From	Arabidopsis thaliana (Mouse-ear cress) [TaxID: 3702]
Taxonomy	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; eurosids II; Brassicales; Brassicaceae; Arabidopsis.
Protein existence	1: Evidence at protein level;

Anotaciones

Comments

- *FUNCTION*: Component of SCF(EBF1) E3 ubiquitin ligase complexes, which may mediate the ubiquitination and subsequent proteasomal degradation of target proteins (probably including EIN3 and EIL1). Regulator of the ethylene signaling cascade by modulating the stability of EIN3 and EIL1 proteins.
- *PATHWAY*: Protein degradation; protein ubiquitination.
- *SUBUNIT*: Part of a SCF (SKP1-cullin-F-box) protein ligase complex. Interacts with CUL1, SKP1A/ASK1, SKP1B/ASK2, EIN3, and EIL1.
- *INTERACTION*:
  P43291:ASK1; NbExp=1; IntAct=EBI-593623, EBI-401164;
  Q94AH6:CUL1; NbExp=1; IntAct=EBI-593623, EBI-532411;
  O24606:EIN3; NbExp=1; IntAct=EBI-593623, EBI-593576;
- *SUBCELLULAR LOCATION*: Nucleus.
- *TISSUE SPECIFICITY*: Ubiquitous.
- *INDUCTION*: EIN3-dependent induction by ethylene.
- *SIMILARITY*: Contains 1 F-box domain.
- *SIMILARITY*: Contains 16 LRR (leucine-rich) repeats.
- *WEB RESOURCE*: Name=PlantsUBQ; Note=A functional genomics database for the ubiquitin/26S proteasome proteolytic pathway in plants; URL="http://plantsubq.genomics.purdue.edu/";.

Copyright

Cross-references

Sequence databases	
EMBL	AJ609239; CAE75865.1; -; Genomic_RNA. [EMBL / GenBank / DDBJ] [CoDingSequence] AY485830; AAR27072.1; -; mRNA. [EMBL / GenBank / DDBJ] [CoDingSequence] AC006258; -; NOT_ANNOTATED_CDS; Genomic_DNA. [EMBL / GenBank / DDBJ] AK227858; BAE99835.1; -; mRNA. [EMBL / GenBank / DDBJ] [CoDingSequence]
RefSeq	NP_197917.1; -.
UniGene	At.19865
3D structure databases	
ModBase	Q708Y0.
Protein-protein interaction databases	
IntAct	Q708Y0; -.
Organism-specific databases	
TAIR	At5g25350; -.
Ontologies	
GO	GO:0005515; Molecular function: protein binding (*inferred from physical interaction from IntAct*). QuickGo view.
Family and domain databases	
InterPro	IPR001810; F-box. IPR001611; LRR. IPR006553; LRR_cys_sub-typ. Graphical view of domain structure.
Pfam	PF00646; F-box; 1. PF00560; LRR_1; 3. Pfam graphical view of domain structure.
SMART	SM00256; FBOX; 1. SM00367; LRR_CC; 1. SMART graphical view of domain structure.
PROSITE	PS50181; FBOX; FALSE_NEG. PROSITE graphical view of domain structure (profiles).
BLOCKS	Q708Y0.
Genome annotation databases	
GeneID	832607; -.
GenomeReviews	BA000015_GR; AT5G25350.

La Web Semántica

Web actual

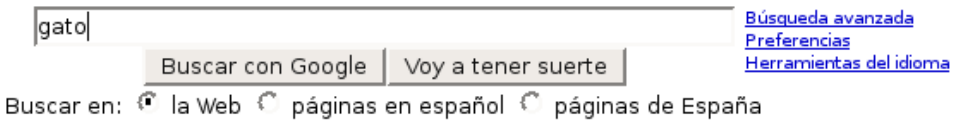
gato

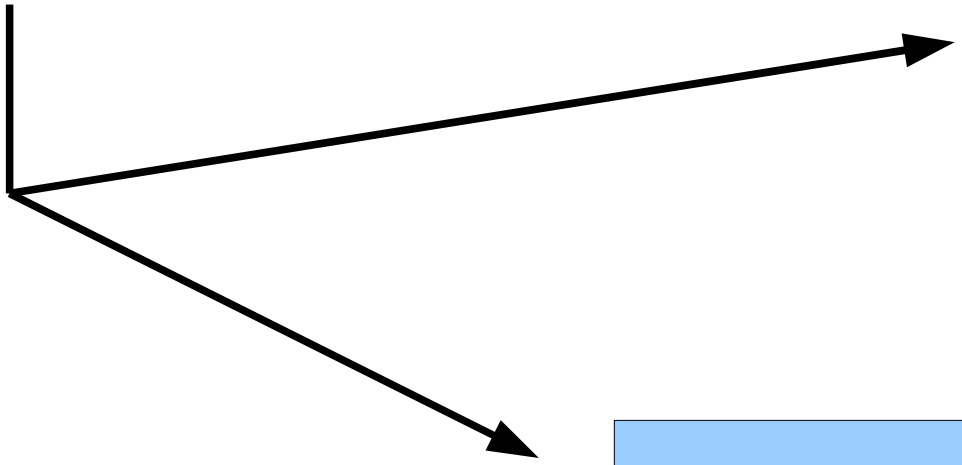
Buscar con Google | Voy a tener suerte

Búsqueda avanzada
Preferencias
Herramientas del idioma

Buscar en: la Web páginas en español páginas de España

Web actual





bla bla bla
bla bla bla
bla gato bla
bla bla bla
bla bla bla

bla gato bla
bla bla bla
bla bla bla
bla bla bla
bla bla bla

Web actual

gato

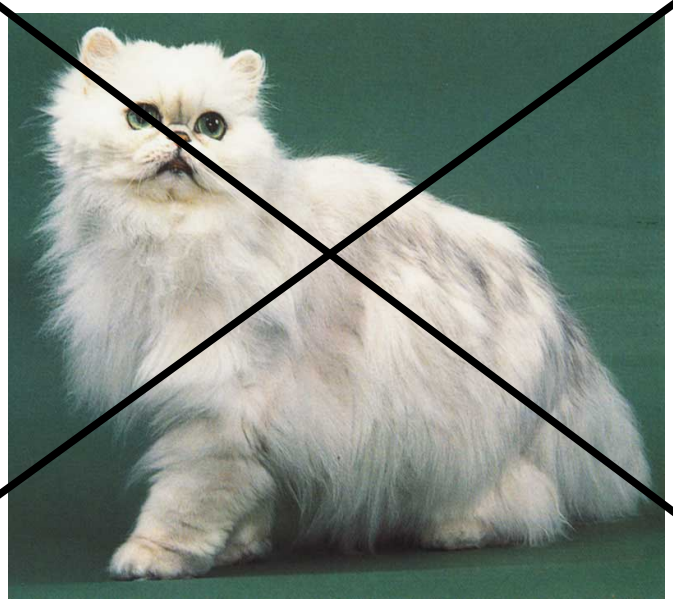
Buscar con Google | Voy a tener suerte

Búsqueda avanzada
Preferencias
Herramientas del idioma

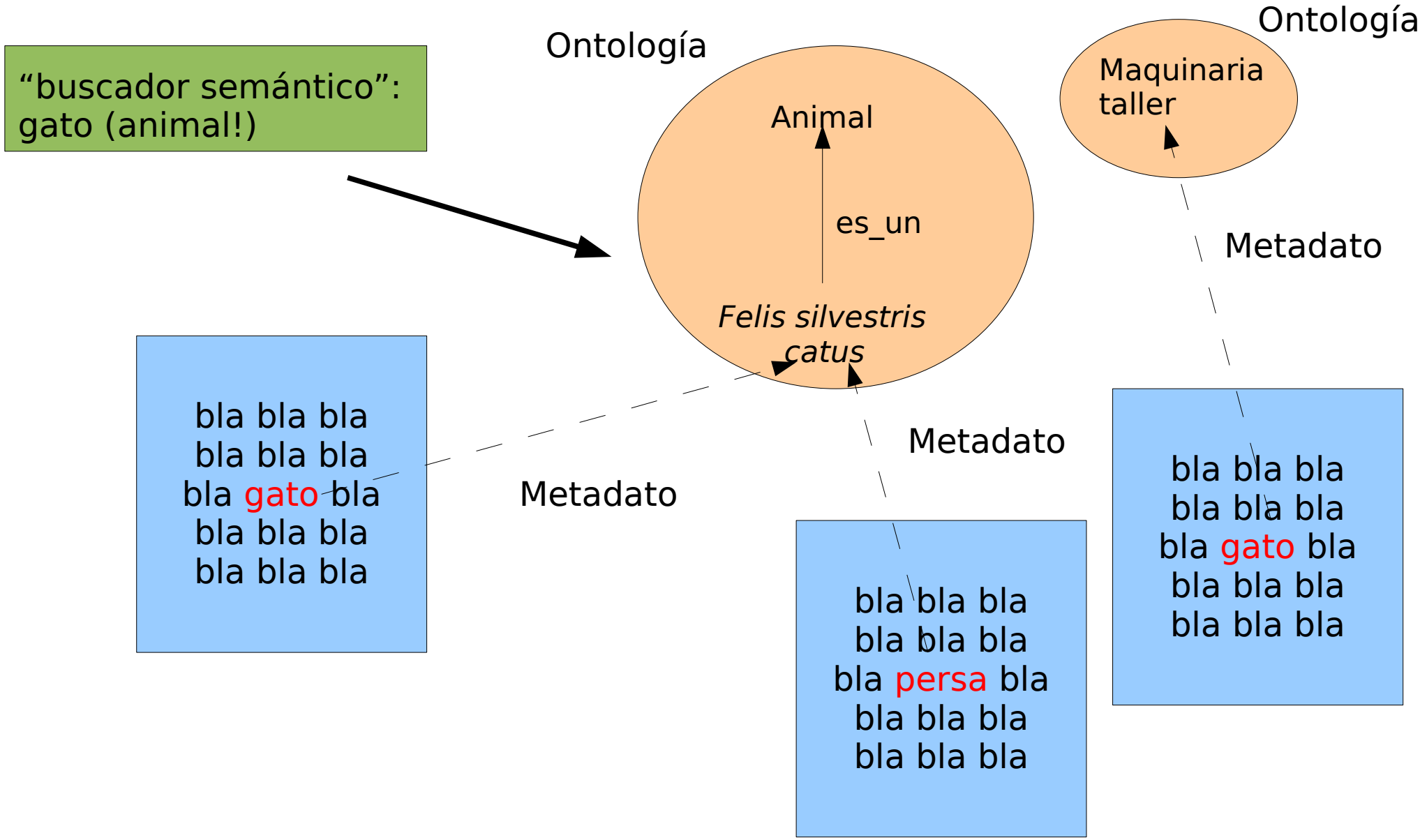
Buscar en: la Web páginas en español páginas de España

Gato

Persa



Web Semántica: metadatos y ontologías



La Red semántica. Berners-Lee, Tim; Hendler, James y Lassila, Ora. Investigación y Ciencia: 298 - JULIO 2001

Web Semántica

Web Semántica → web de “conceptos” (vs docs) → “preguntas”

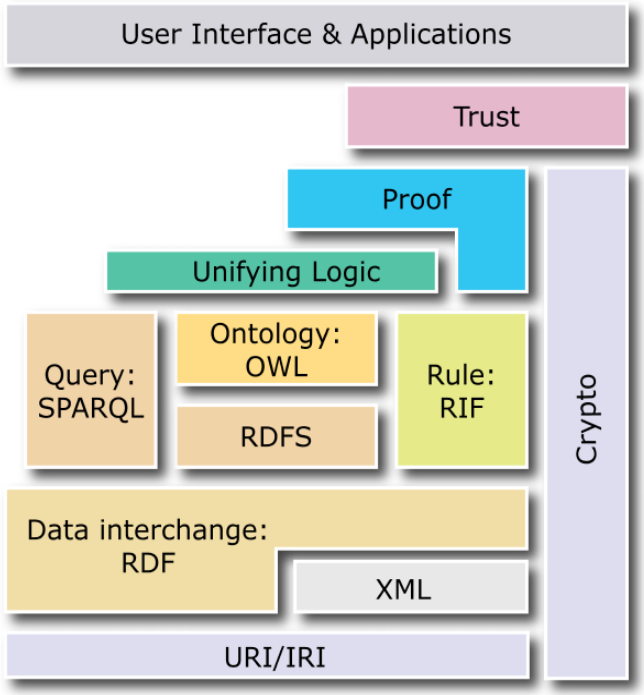
“Necesito buscar gatos negros que hayan residido sus dos primeros años en un centro de acogida dirigido por un equipo internacional, cerca de mi residencia”

World Wide Web Consortium (W3C): http://www.w3.org/

W3C Semantic Web Activity:
http://www.w3.org/2001/sw/

Resource Description Framework (RDF):
Representar datos
http://www.w3.org/RDF/

Web Ontology Language (OWL):
Representar conocimiento sobre esos datos en **ontologías**
http://www.w3.org/2004/OWL/



¿Qué es una ontología?

ONTOLOGIA

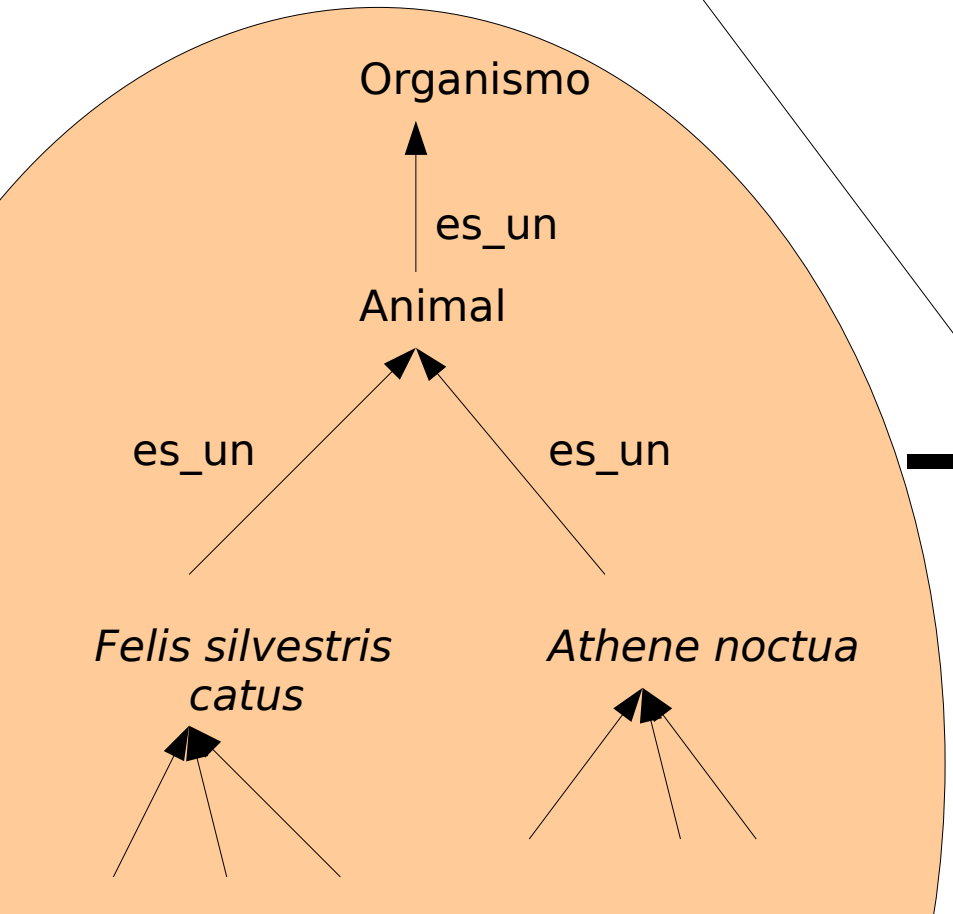
[filosofía] disciplina que describe lo existente.

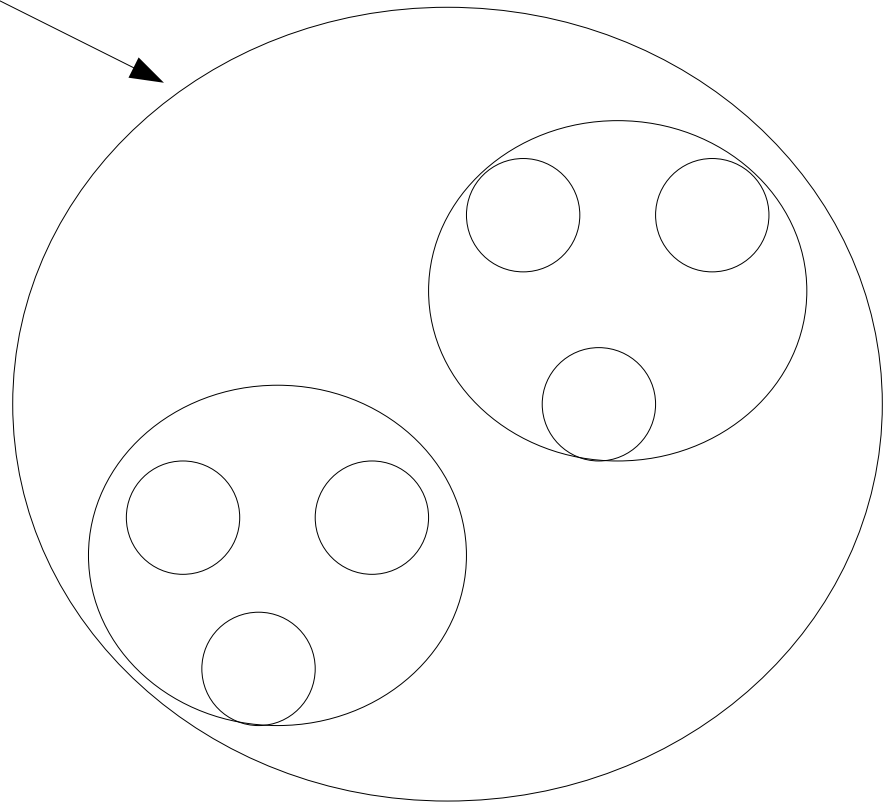
[informática] conceptos + relaciones.

CONTENIDO SEMANTICO

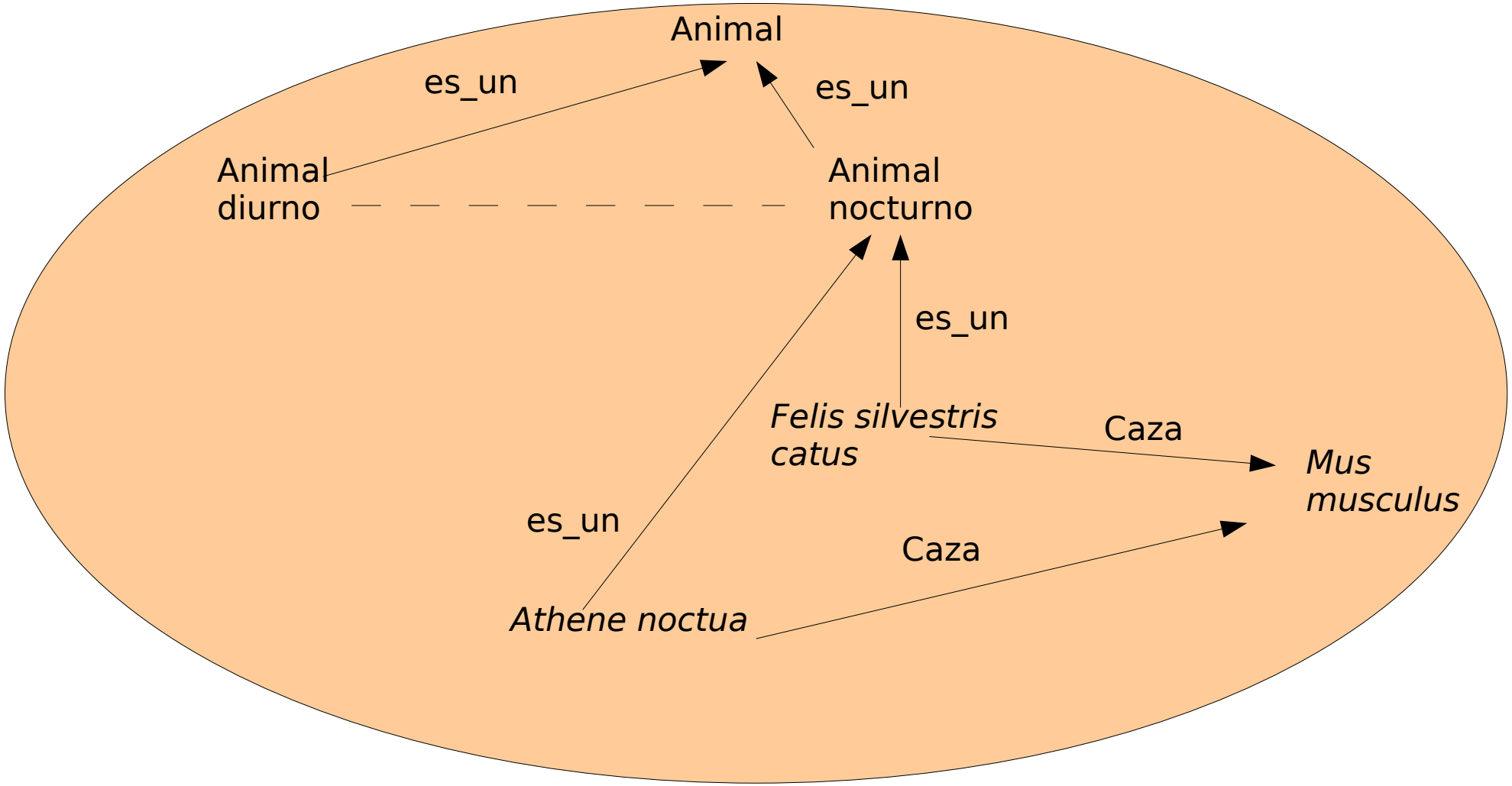
conceptos + relaciones → estructura → procesable por un ordenador (indep. nombres relaciones + conceptos): Reasoning

formalismo

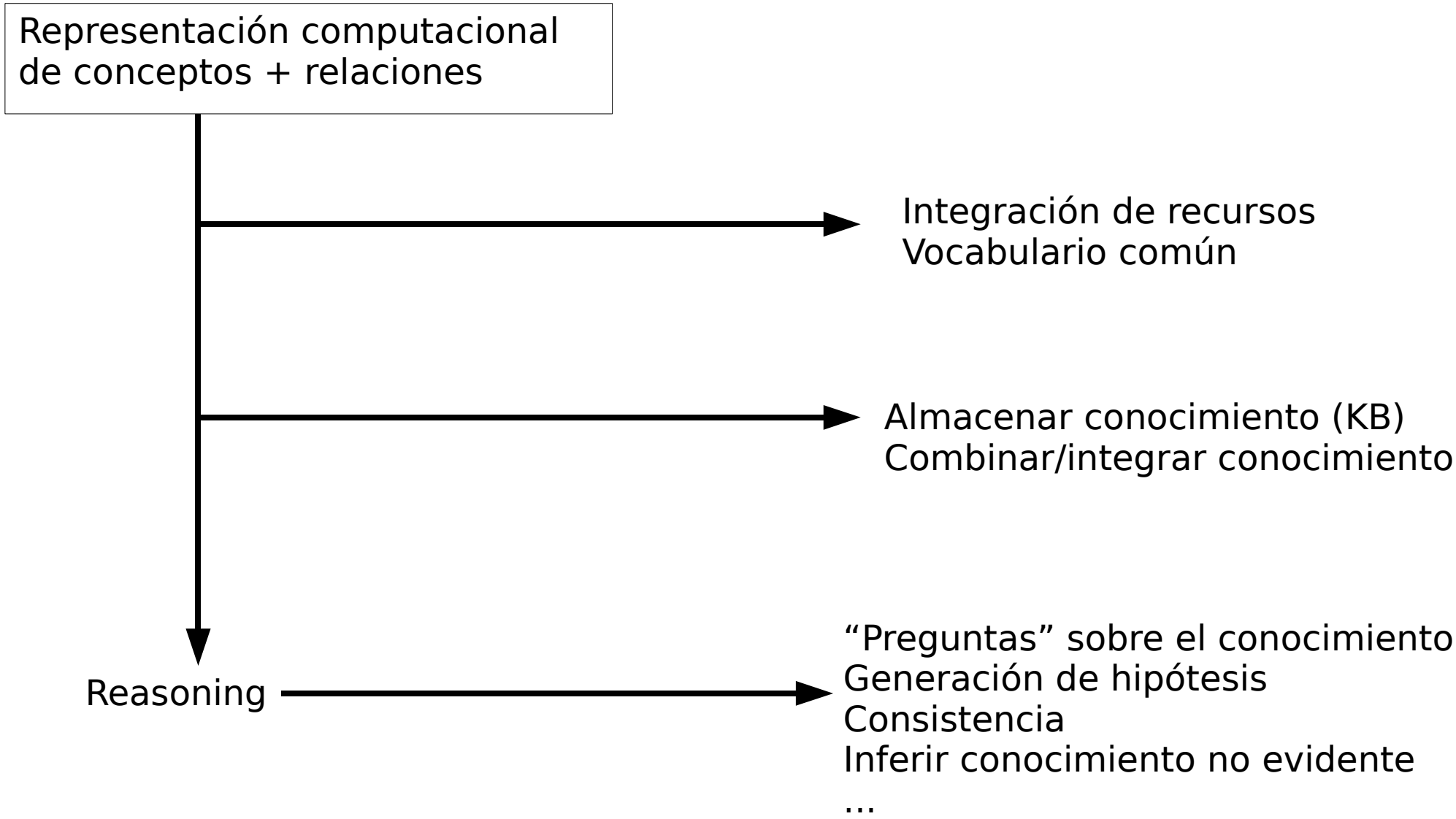




¿Qué es una ontología?



Funciones de una ontología



¿Qué es una ontología?

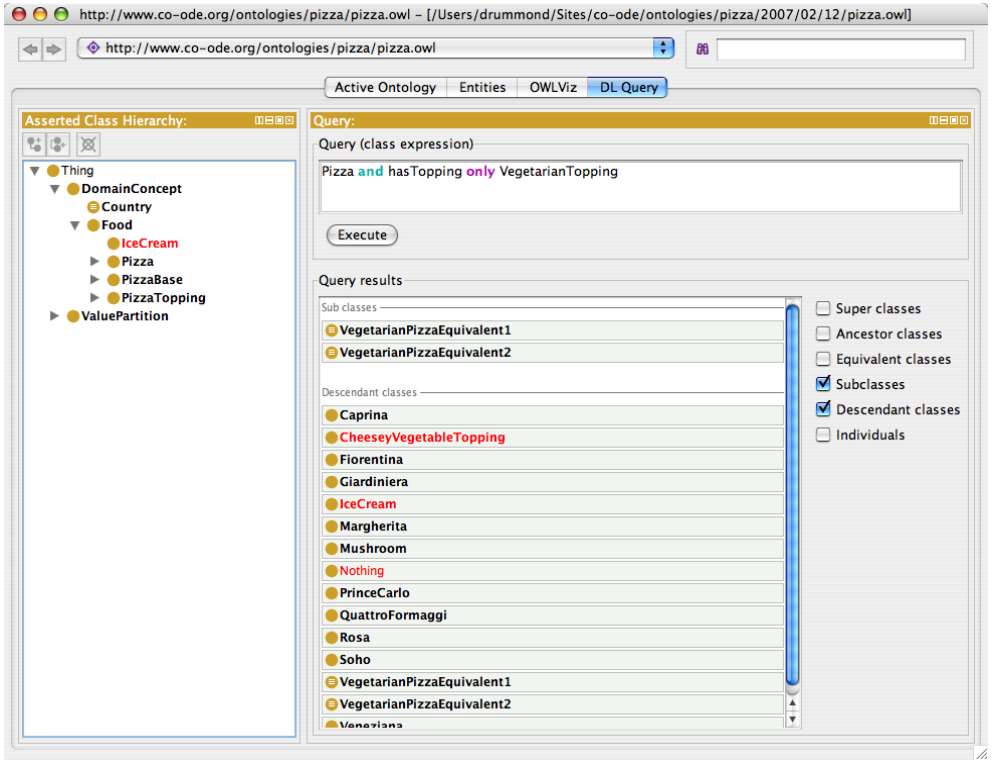
Se crean usando un lenguaje de representación de conocimiento: OWL, ...

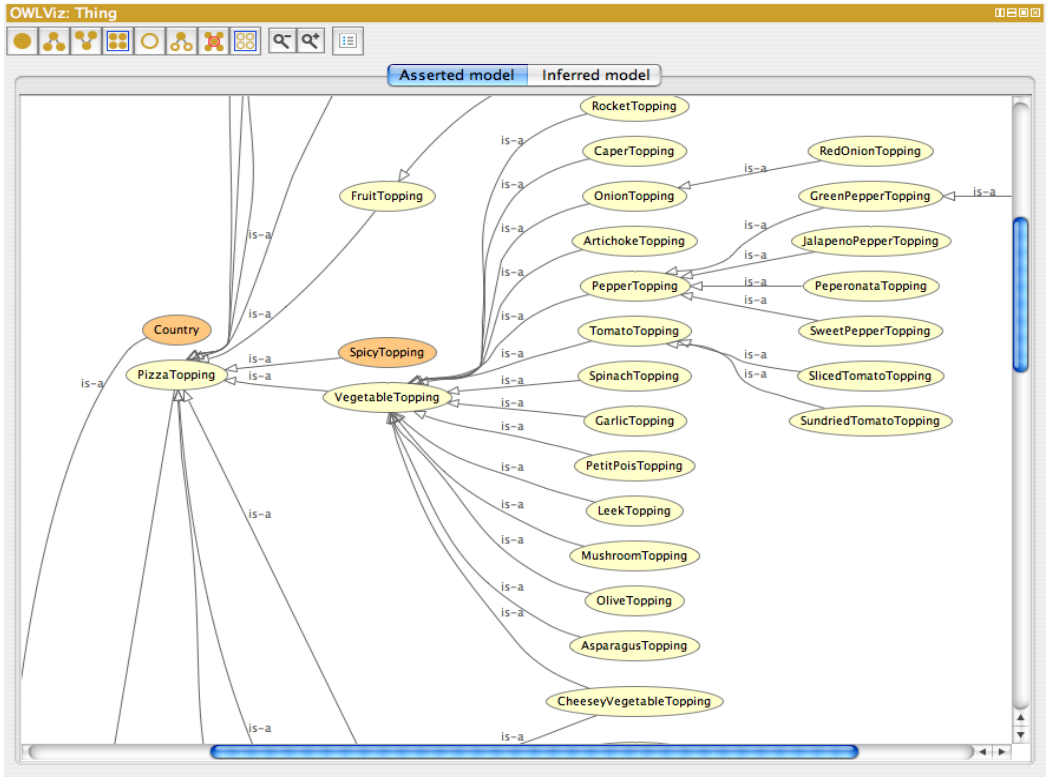
Se “encarnan” en archivos.

```
<owl:Class rdf:about="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0002060">
    <rdfs:label xml:lang="en">NEB2_HUMAN</rdfs:label>
    <oboInOwl:hasDefinition>
        <oboInOwl:Definition>
            <rdfs:label xml:lang="en">Neurabin-2</rdfs:label>
            <oboInOwl:hasDbXref>
                <oboInOwl:DbXref>
                    <rdfs:label>UniProt:Q96SB3</rdfs:label>
                    <oboInOwl:hasURI rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                        http://www.cellcycleontology.org/ontology/owl/UniProt#UniProt_Q96SB3
                    </oboInOwl:hasURI>
                </oboInOwl:DbXref>
            </oboInOwl:hasDbXref>
        </oboInOwl:Definition>
    </oboInOwl:hasDefinition>
    <oboInOwl:hasDbXref>
        <oboInOwl:DbXref>
            <rdfs:label>UniProt:Q8TCR9</rdfs:label>
            <oboInOwl:hasURI rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                http://www.cellcycleontology.org/ontology/owl/UniProt#UniProt_Q8TCR9
            </oboInOwl:hasURI>
        </oboInOwl:DbXref>
    </oboInOwl:hasDbXref>
    <rdfs:subClassOf
rdf:resource="http://www.cellcycleontology.org/ontology/owl/CCO#CCO_B0000000"/>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty>
                <owl:ObjectProperty rdf:about=
                "http://www.cellcycleontology.org/ontology/owl/CCO#belongs_to"/>
            </owl:onProperty>
```

¿Qué es una ontología?

Se manipulan con programas: Protégé, OBOEdit, ...





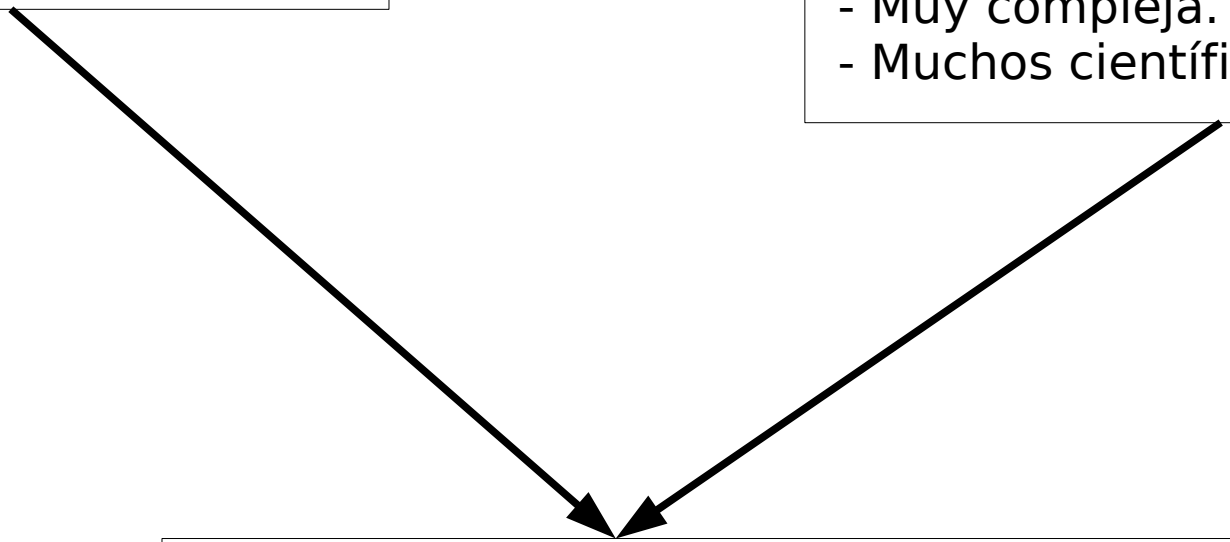
“Reasoners”: Pellet, FaCT++, Racer, ...

Implantación de la Web Semántica en la bioinformática actual

Web Semántica en bioinformática

Tecnología semántica
(OWL, RDF, ...)

Bioinformática:
- Mucha información.
- Muy compleja.
- Muchos científicos creando metadatos.



Matrimonio perfecto:
- Bioinformáticos: explotar conocimiento.
- Investigadores productores de tecnología semántica: casos de uso reales, requerimientos ...

Bio-ontologías

Web Semántica en bioinformática

W3C Semantic Web Health Care and Life Sciences Interest Group:
http://www.w3.org/2001/sw/hcls/

Good BM, Wilkinson MD.The Life Sciences Semantic Web is full of creeps! Brief Bioinform. 2006 Sep;7(3):275-86.

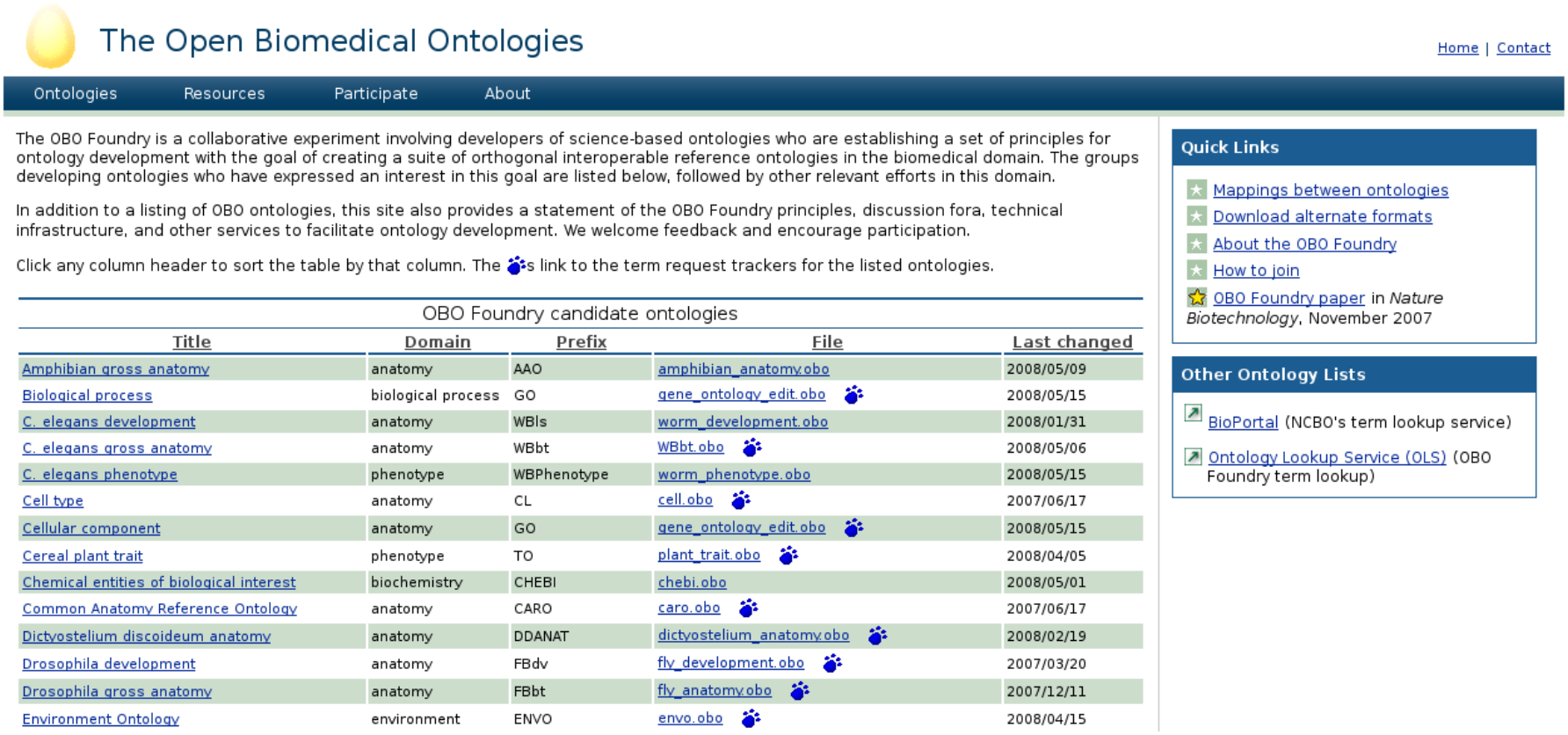
Olivier Bodenreider, Robert Stevens.Bio-ontologies: current trends and future directions. Brief Bioinform, 7(3):256–274, 2006.

Xiaoshu Wang, Robert Gorlitsky, Jonas S Almeida. From XML to RDF: how semantic web technologies will change the design of omic standards. Nat Biotechnol. 2005 Sep;23(9):1099-103.

Web Semántica en bioinformática

OBO: Open Biomedical Ontologies (http://obofoundry.org/)

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Gold
The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.Nat Biotechnol. 2007 Nov; 25(11):1251-5.



The Open Biomedical Ontologies

Home | Contact

Ontologies | Resources | Participate | About

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other relevant efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the OBO Foundry principles, discussion fora, technical infrastructure, and other services to facilitate ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The 🐾s link to the term request trackers for the listed ontologies.

OBO Foundry candidate ontologies

Title	Domain	Prefix	File	Last changed
Amphibian gross anatomy	anatomy	AAO	amphibian_anatomy.obo	2008/05/09
Biological process	biological process	GO	gene_ontology_edit.obo 🐾	2008/05/15
C. elegans development	anatomy	WBls	worm_development.obo	2008/01/31
C. elegans gross anatomy	anatomy	WBbt	WBbt.obo 🐾	2008/05/06
C. elegans phenotype	phenotype	WBPhenotype	worm_phenotype.obo	2008/05/15
Cell type	anatomy	CL	cell.obo 🐾	2007/06/17
Cellular component	anatomy	GO	gene_ontology_edit.obo 🐾	2008/05/15
Cereal plant trait	phenotype	TO	plant_trait.obo 🐾	2008/04/05
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo	2008/05/01
Common Anatomy Reference Ontology	anatomy	CARO	caro.obo 🐾	2007/06/17
Dictyostelium discoideum anatomy	anatomy	DDANAT	dictyostelium_anatomy.obo 🐾	2008/02/19
Drosophila development	anatomy	FBdv	fly_development.obo 🐾	2007/03/20
Drosophila gross anatomy	anatomy	FBbt	fly_anatomy.obo 🐾	2007/12/11
Environment Ontology	environment	ENVO	envo.obo 🐾	2008/04/15

Quick Links

- Mappings between ontologies
- Download alternate formats
- About the OBO Foundry
- How to join
- OBO Foundry paper in *Nature Biotechnology*, November 2007

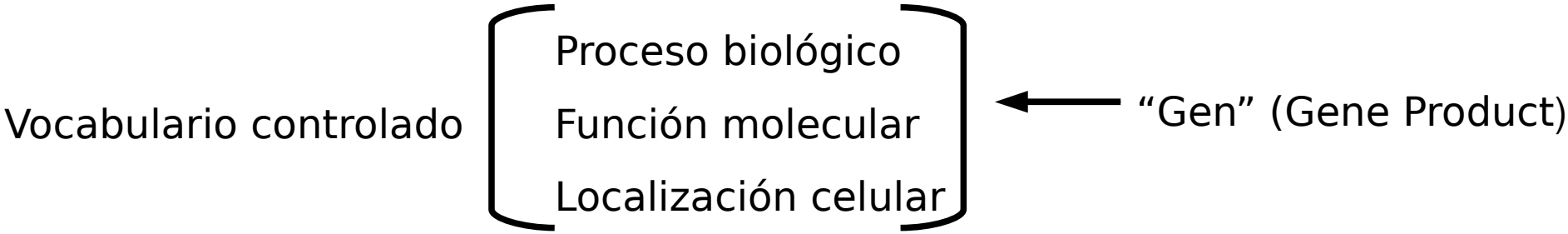
Other Ontology Lists

- BioPortal (NCBO's term lookup service)
- Ontology Lookup Service (OLS) (OBO Foundry term lookup)

Web Semántica en bioinformática

Gene Ontology (GO):
http://www.geneontology.org/

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

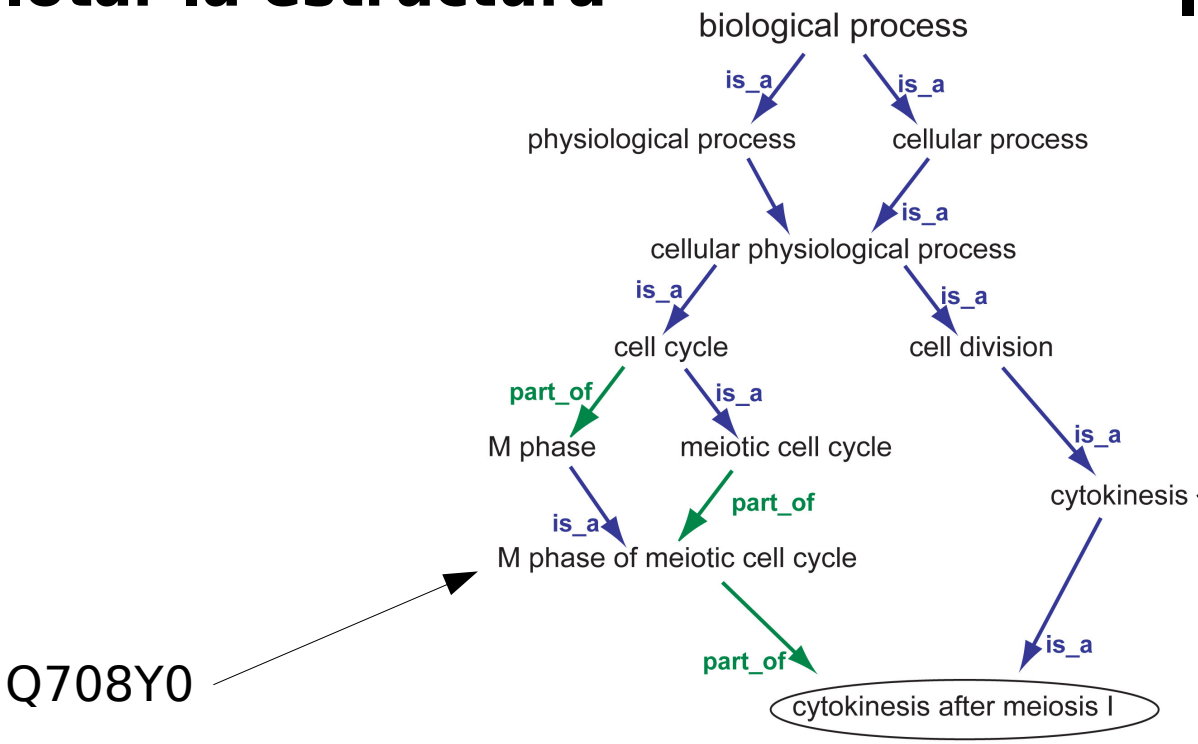


Cierta structura: is_a, part_of, regulates.

Web Semántica en bioinformática

Gene Ontology

Explotar la estructura



Integración de recursos

Base de datos
Q708YF

Base de datos
Q708YF
Q8M34J

Base de datos
CYC4;D

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

Web Semántica en bioinformática

EXITO DE GENE ONTOLOGY

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.

PROBLEMAS DE GENE ONTOLOGY

Solo is_a, part_of, regulates, no muy expresivo.

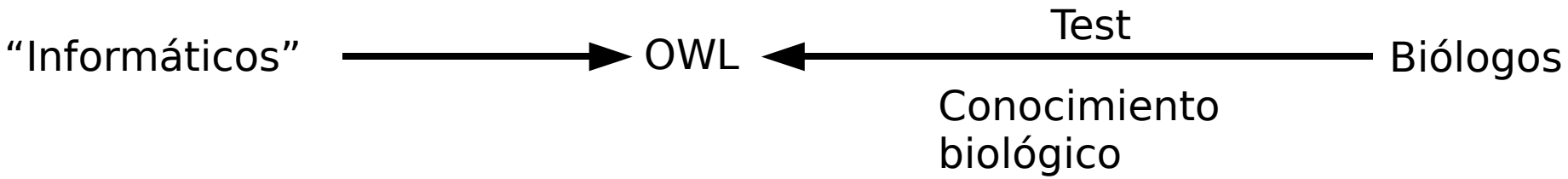
OBO format (sin modelo semántico) vs RDF/OWL:

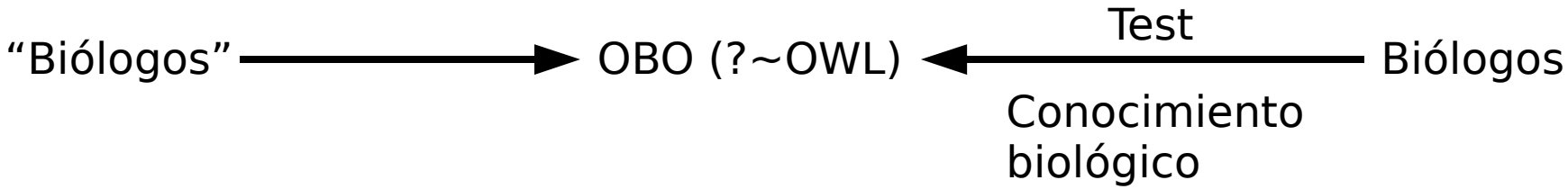
Mikel Egaña Aranguren, Sean Bechhofer, Phillip Lord, Ulrike Sattler, Robert Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL.
BMC Bioinformatics 2007, 8:57.

No usa reasoning:

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

EL DIVORCIO

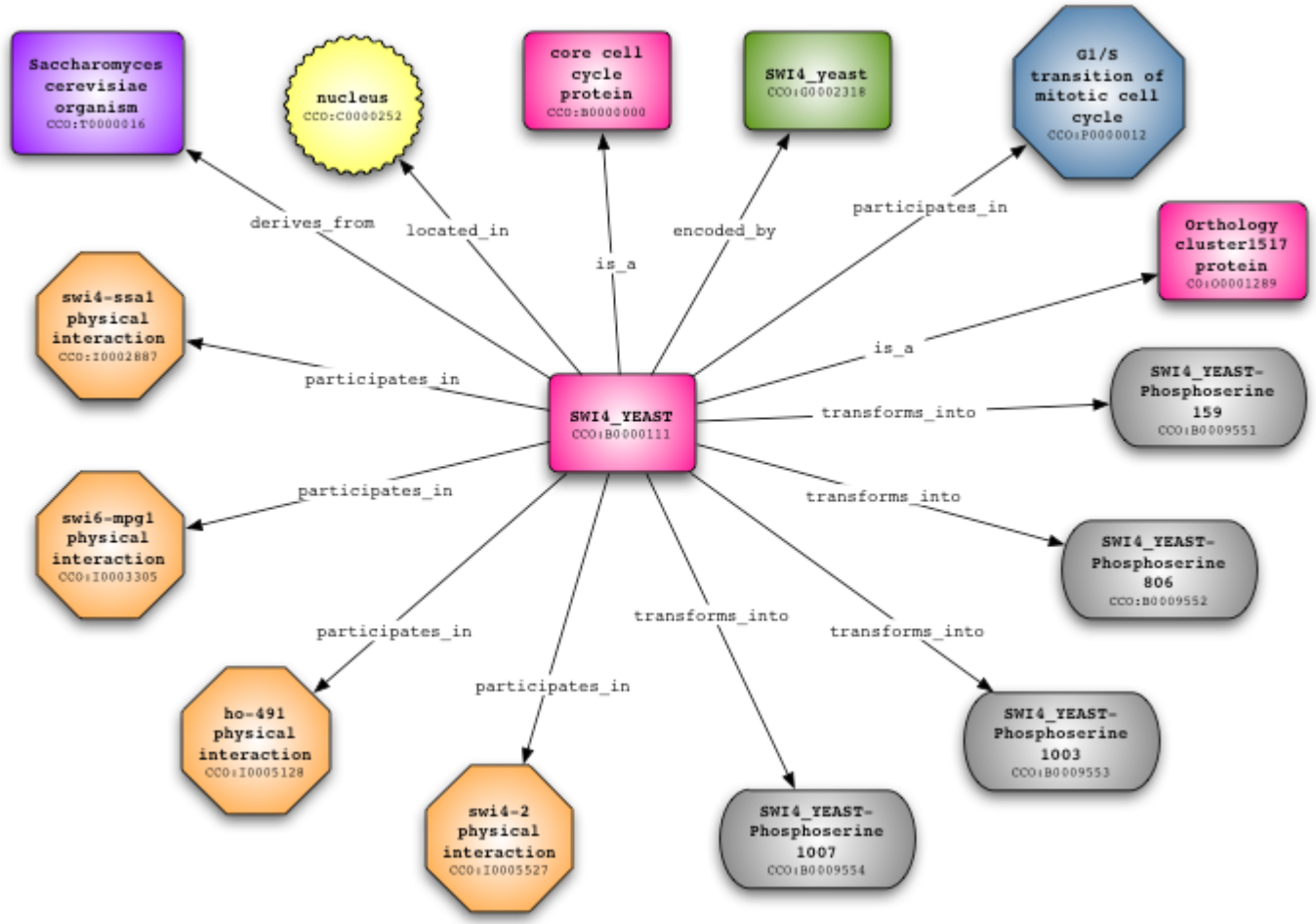




Carole Goble and Chris Wroe. The Montagues and the Capulets. Comparative and Functional Genomics. Volume 5 (2004), Issue 8, Pages 623-632

Cell Cycle Ontology (CCO)

http://www.cellcycleontology.org/



Web Semántica en bioinformática

Cell Cycle Ontology (CCO)

Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. Mikel Egaña Aranguren, Erick Antezana, Martin Kuiper, Robert Stevens. BMC bioinformatics 2008, 9(Suppl 5):S1.

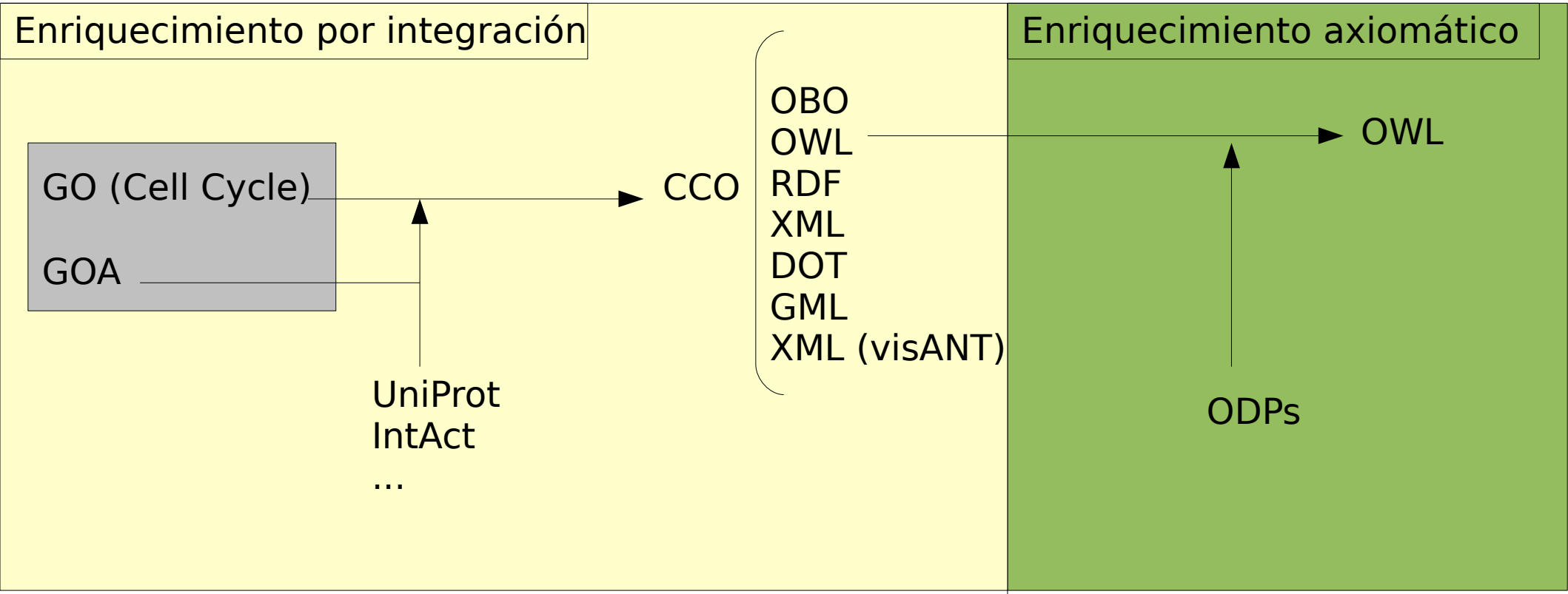
Transforming the Axiomisation of Ontologies: The Ontology Pre-Processor Language. M. Egaña; R. Stevens; E. Antezana. OWLED 2008, Fourth International Workshop - Washington, DC, USA.

ONTO-PERL: An API supporting the development and analysis of bio-ontologies. E. Antezana; M. Egana; B. De Baets; M. Kuiper; V. Mironov. Bioinformatics 2008 24: 885-887; doi:10.1093/bioinformatics/btn042

A cell-cycle knowledge integration framework.E. Antezana; E. Tsiporkova; V. Mironov; and Kuiper, M.DILS 2006, LNBI 4075, pp. 19-34, 2006.

Web Semántica en bioinformática

Cell Cycle Ontology (CCO)



OBO: BioPortal, OLS, OBO-Edit
OWL: Protégé, Ontology Server, reasoner
RDF: SPARQL (Virtuoso)
DOT: GraphViz
GML: Cytoscape

OWL: Protégé, Ontology Server, reasoner

Problemas de la Web Semántica en bioinformática

FRUSTRACIÓN POR ...

... PERCEPCIÓN DEMASIADO POSITIVA DE LA TECNOLOGÍA

Con meter todo en una ontología no es suficiente: hay que añadir axiomas, que es lo difícil.

El reasoner no es el genio de la lámpara.

Las ontologías no sirven para todo.

“Semantics” buzz word.

... PERCEPCIÓN DEMASIADO NEGATIVA DE LA TECNOLOGÍA

OWL es difícil de entender y no muy intuitivo (Unique Name Assumption, Open World Assumption).

Es difícil crear una buena ontología: es difícil ver los beneficios a corto plazo.

Tecnología todavía experimental:

- los reasoners fallan en cosas incomprensibles.
- trabajar con ontologías muy grandes es casi imposible.

Problemas de la Web Semántica en bioinformática

FALTA:

Mano de obra: biólogos que conozcan el dominio (¡CCO!).

Identificadores únicos para entidades:
LSID: http://lsids.sourceforge.net/
URI: http://www.w3.org/2001/sw/hcls/notes/uris/

Herramientas más “user friendly” para crear ontologías o introducir contenido semántico en páginas web.

Explicación “user friendly” de reasoning.

Mucho conocimiento sin codificar en ontologías, o peor todavía, mal codificado:
Contenido semántico “enterrado” en “anotaciones”.
Larisa N Soldatova and Ross D King. Are the current ontologies in biology good ontologies?. Nature Biotechnology 23, 1095 - 1098 (2005)

Más comunicación: diseñadores OWL – creadores de bio-ontologías.

OBO format?

“Killer app”.

El futuro de la Web Semántica en bioinformática

Sistemas que representan el mismo dominio de conocimiento con diferentes formalismos **integrados**: BD, ontologías, simulaciones metabólicas, ...

Publicar modelos (ontologías) en vez de/junto con artículos.

Intervención humana:

- Validación experimental.
- La tecnología semántica no sustituye al investigador, simplemente le ahorra tiempo.

Agradecimientos, licencia, ¿Preguntas?

EPSRC, University of Manchester

UM, Jesualdo Tomás Fernández Breis

Erick Antezana
