mig

Unité Mathématique, Informatique et Génome

Ontology design

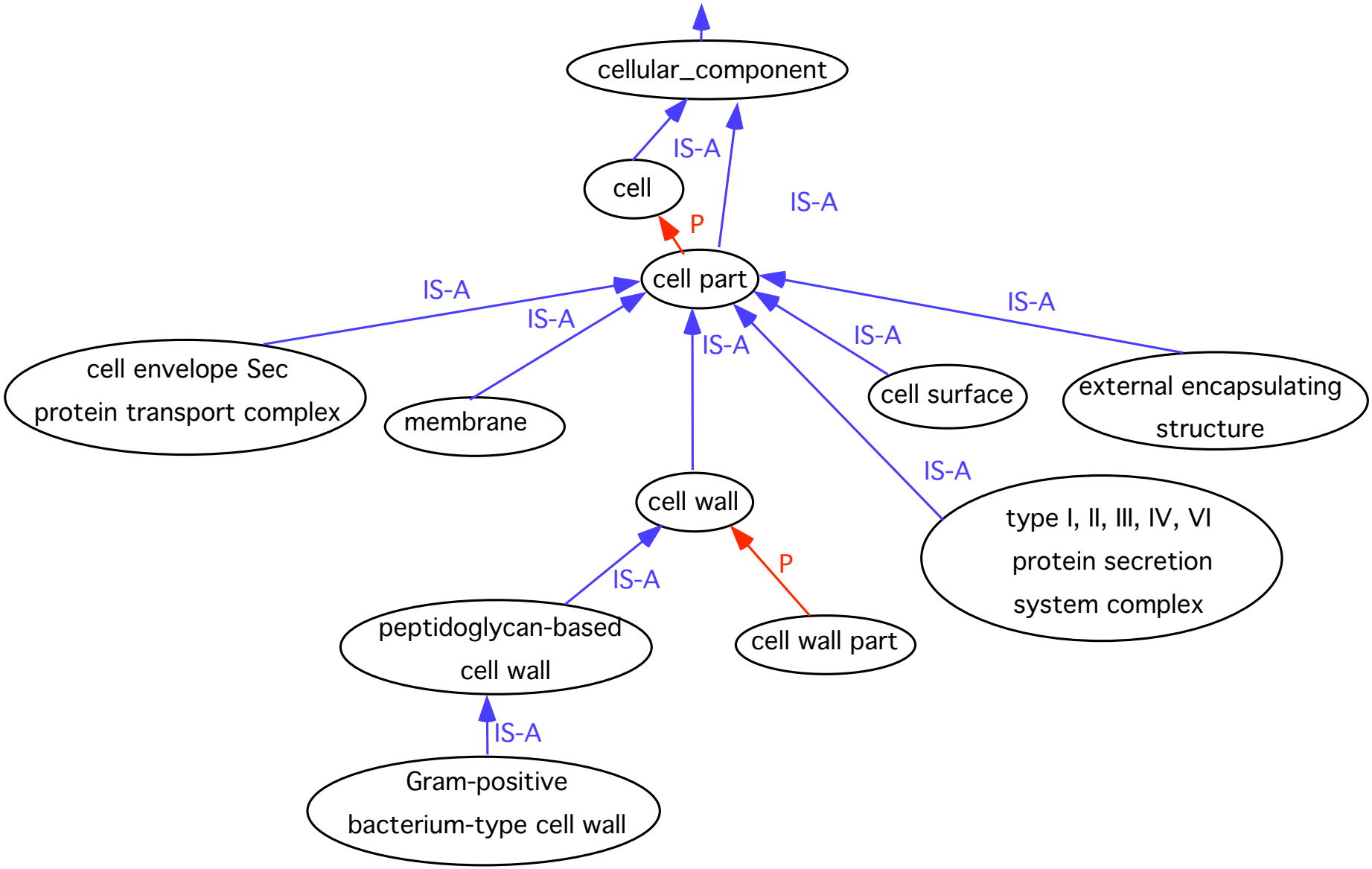
Transys training session

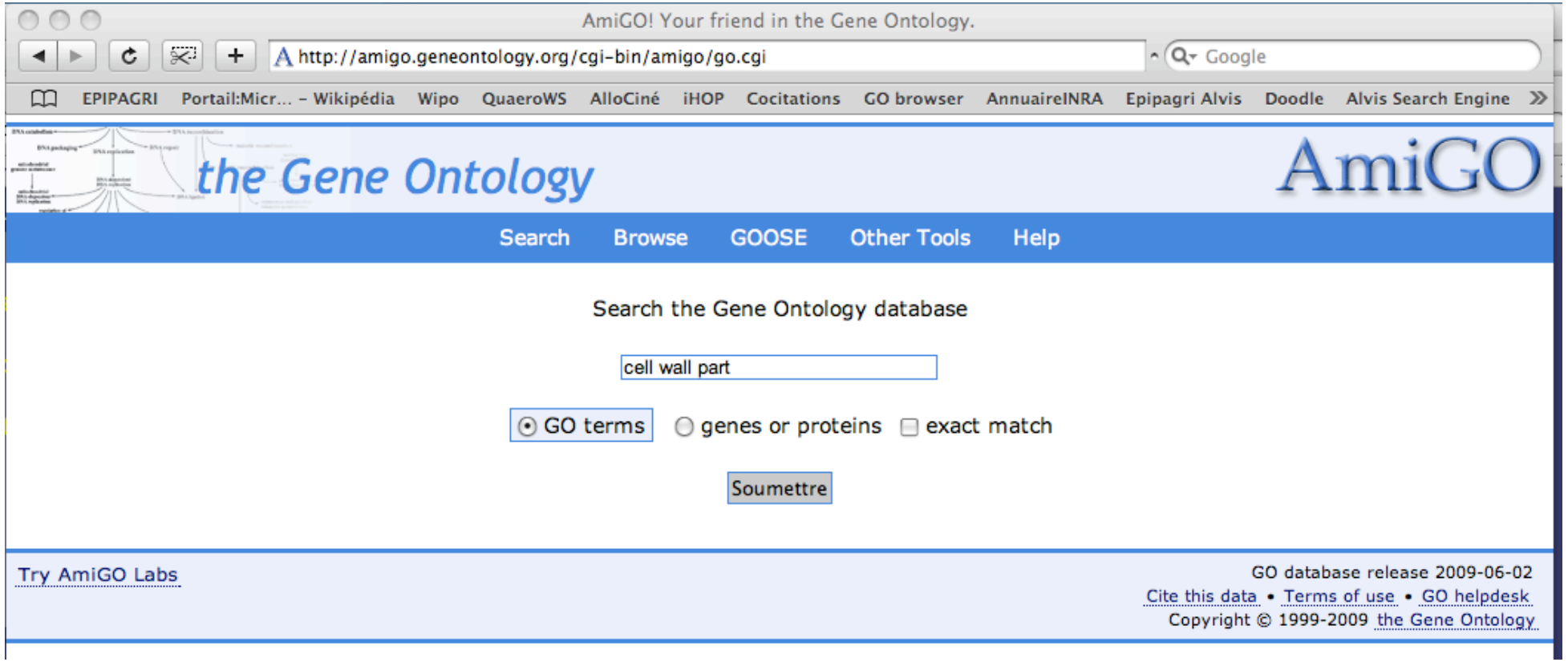
Jouy-en-Josas – 3rd June 2009

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

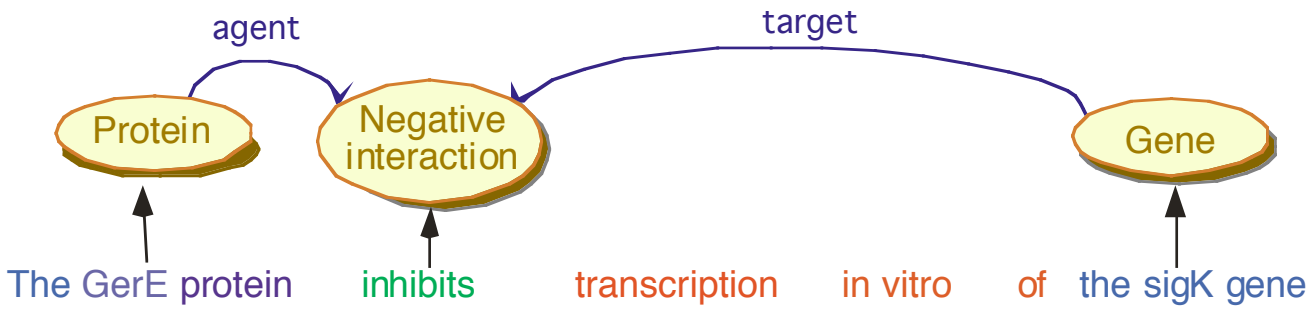
Example of Ontology from Gene Ontology




AmiGO! Your friend in the Gene Ontology.
http://amigo.geneontology.org/cgi-bin/amigo/go.cgi
Google
EPIPAGRI
Portail:Micr... – Wikipédia
Wipo
QuaeroWS
AlloCiné
iHOP
Cocitations
GO browser
AnnuaireINRA
Epipagri Alvis
Doodle
Alvis Search Engine
the Gene Ontology
AmiGO
Search
Browse
GOOSE
Other Tools
Help
Search the Gene Ontology database
cell wall part
GO terms
genes or proteins
exact match
Soumettre
Try AmiGO Labs
GO database release 2009-06-02
Cite this data • Terms of use • GO helpdesk
Copyright © 1999-2009 the Gene Ontology

Ontology for text interpretation

Simple example of text annotation by an ontology

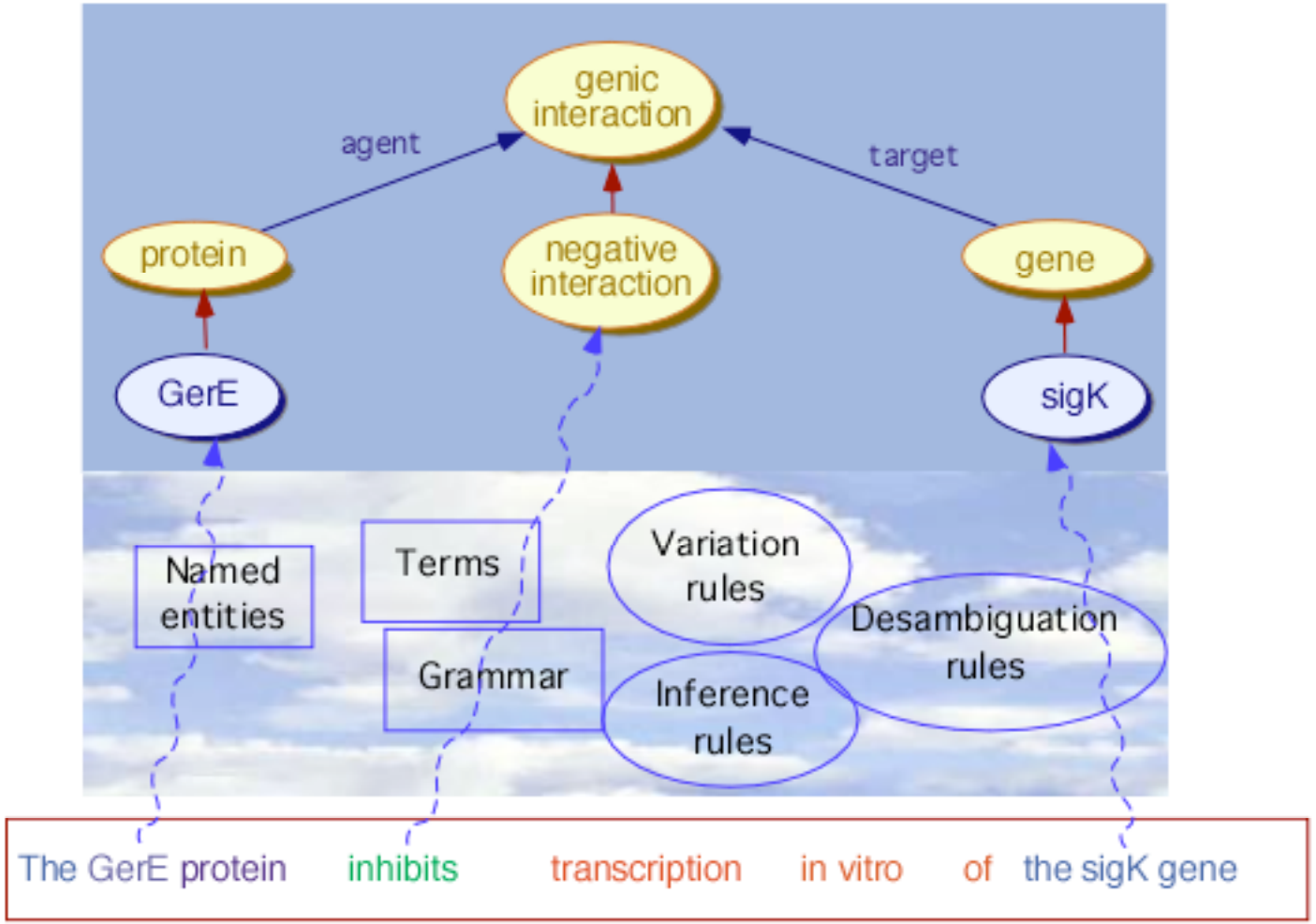


Associate text *fragments to concepts and relations of the ontology*

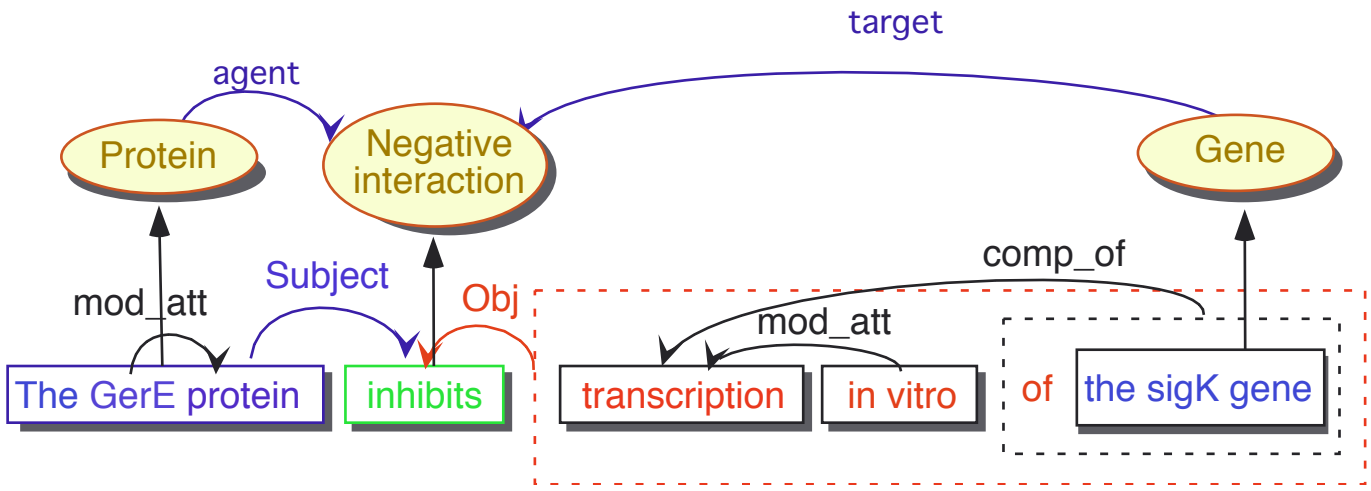
In the simplest case

- The text terms coorespond to labels of ontology concepts (*GerE*, *inhibit*, *sigK*)
- The relations are deried from the order of the words (protein and gene surround and interaction verb)

Text to ontology



From text to semantic annotation



Semantic relation	agent(Ger_protein, inhibit), target(sigK gene, inhibit), …
Semantic category	is_a(Ger_protein, protein), is_a(inhibit, negative_interaction), …
Syntactic dependency	subject(Ger_protein, inhibit), obj(transcription, inhibit), …
Term	term(GerE protein), term(in vitro) term(sigK gene) , …
Grammatical category	cat(the, det), cat(GerE, nom), cat(inhibit, verbe), …
Named entity	entity(GerE, protein), entity(sigK, gene), entity(sigma K, protein) , …
Segmented text	word(the), word(Ger_protein), word(inhibit), word(transcription), …

Ontology as a formal knowledge model

New facts can be automatically derived from the formal representation of knowledge in a model with guarantees of validity and completeness.

Simple example

A first order logic model on ability of birds to fly.
Specific classes iherit from properties of more general classes.

Ontology =

{Bird(X) → Fly (X),	All birds fly
Canary(X) → Bird(X)}	All canaries are birds
———————————	
Canary(X) →Fly(X)	Then, all canaries fly

Observation: Canary(titi)
Derived fact: Bird(titi), Fly(titi)

Ontology and model

The ontology is the *science of being, of categories, and structures of the objects, of properties, events, processes and relations*.

Concept used in *Artificial Intelligence*,

Because of the need for declarative representations as reusable and general as possible,

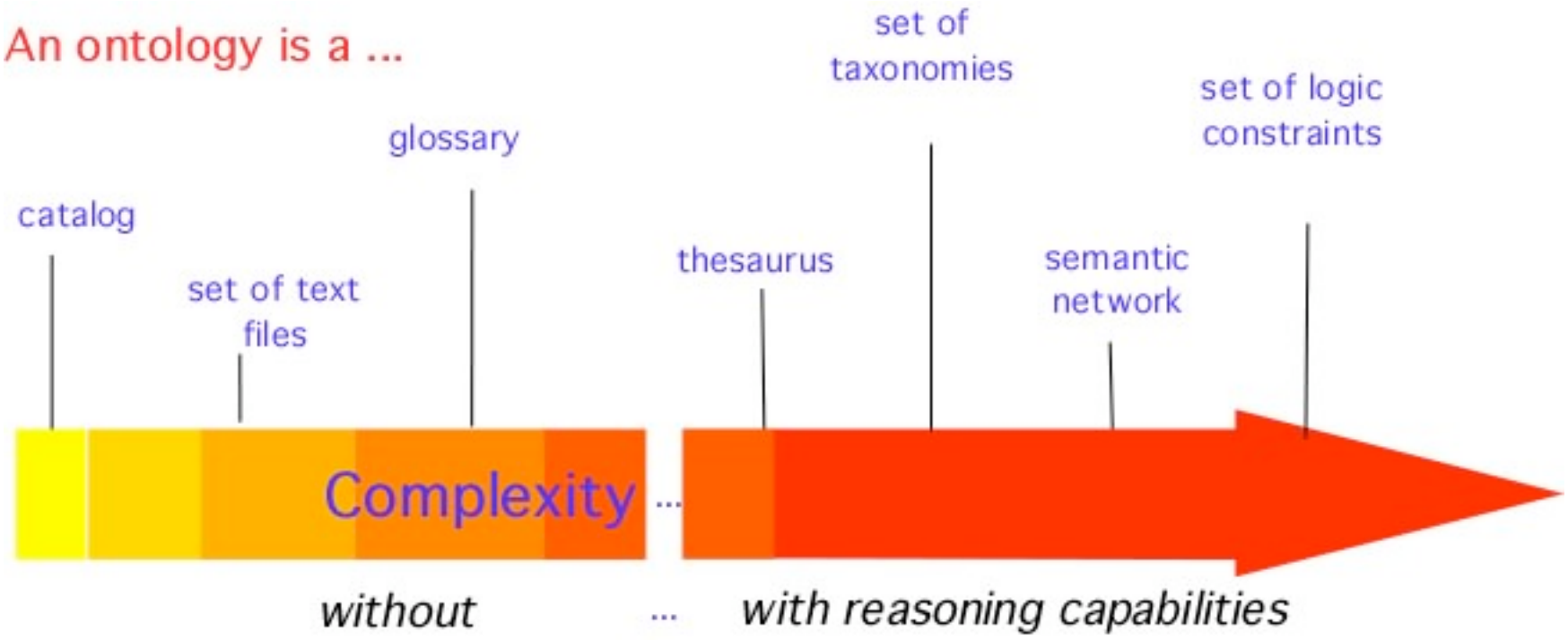
Corresponding the objects and processes that they represent.

Formal ontology is an active research field in particular in *logic models* that are able to simulate reasoning knowledge by automatic reasoning mecanisms. (ex. *Owl* language).

Formal approaches and empirical knowledge bases converge in particular in the biomedical domain.

Applied Ontologies

As opposed to formal ontologie, many kinds of representation are called ontologies (*C. Welty et al.)*



Unstructured ontologies

Catalogues, nomenclature : item lists with unique identifiers,
Example :

SwissProt

Hugo (Human Gene Nomenclature) and links to d'Ingenuity bibliographic database

Glossaries, terminologies or dictionnaries : « keywords» with, or without, definitions and synonyms. For document indexing and search by « string-matching ».

Examples

CadInteract
(cocitation of gene names)

Gene Interactions Database

Sentence	PMID
A low concentration of **gerE** activated cotB transcription by final **sigma(K)** RNA polymerase, whereas a higher concentration was needed to activate transcription of cotX or cotC.	10788508 Abstract Annotation
Transcription of the cotB, cotC, and cotX genes by final **sigma(K)** RNA polymerase is activated by a small, DNA-binding protein called **gerE**.	10788508 Abstract Annotation
In extension of recent work on the transcription of cot gene A and the mother-cell regulatory genes **gerE**, **sigK** and spoIIID, we show that genes involved in coat formation are turned on in a regulatory cascade of at least four co-ordinately controlled gene sets.	1691789 Abstract

Ontology and reasoning

Thesaurus : terms are organized in generality hierarchies. The properties of more general classes are supposed to be true for more specific classes.

Exemples

MeSH

Cells [A11]
Cellular Structures [A11.284]
Cell Membrane [A11.284.149] +
Cell Surface Extensions [A11.284.180] +
▶ Cell Wall [A11.284.183]
Cell Wall Skeleton [A11.284.183.200]

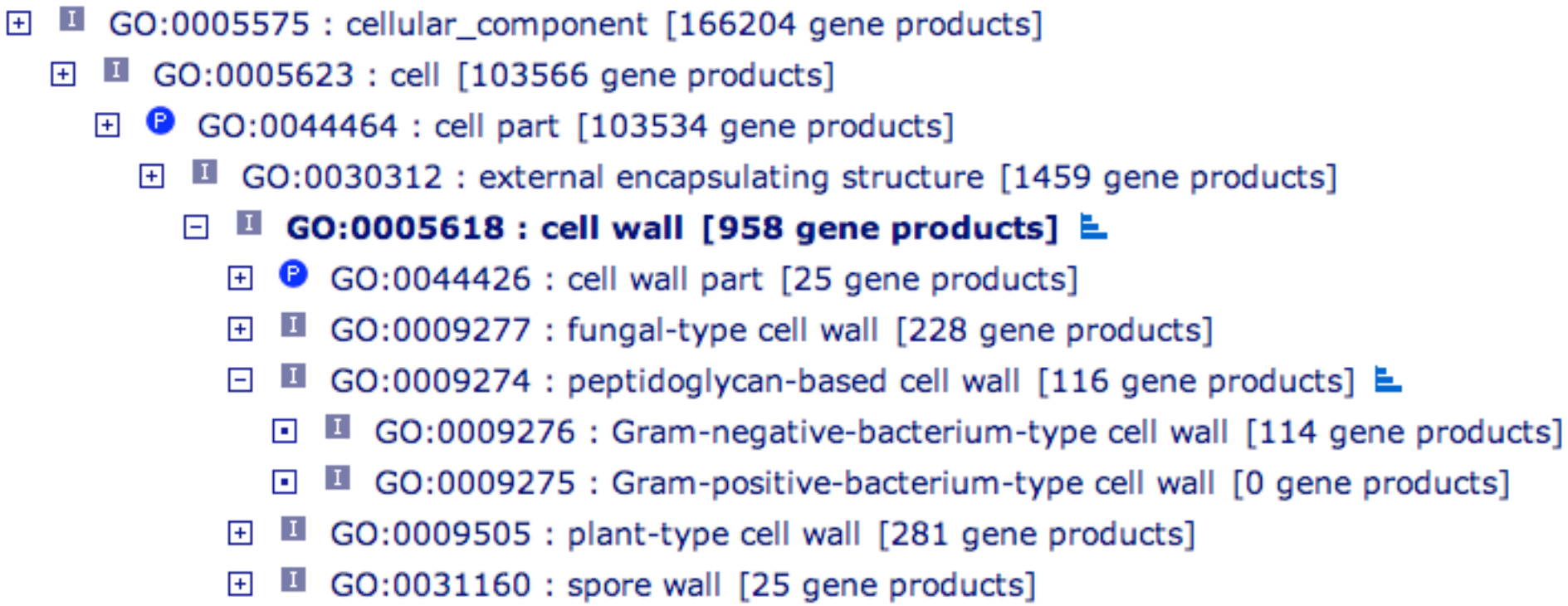
The taxinomic relations in thesaurii are often unformal, weak, and may produce invalid inferences.

Example from MeSH : mutation → mutagenesis
Instead of *mutation* is_a_result_of *mutagenesis*

☹ Unexpected result: PubMed query: « *mutagenesis* » retrieves documents on *Expanded DNA Repeats*

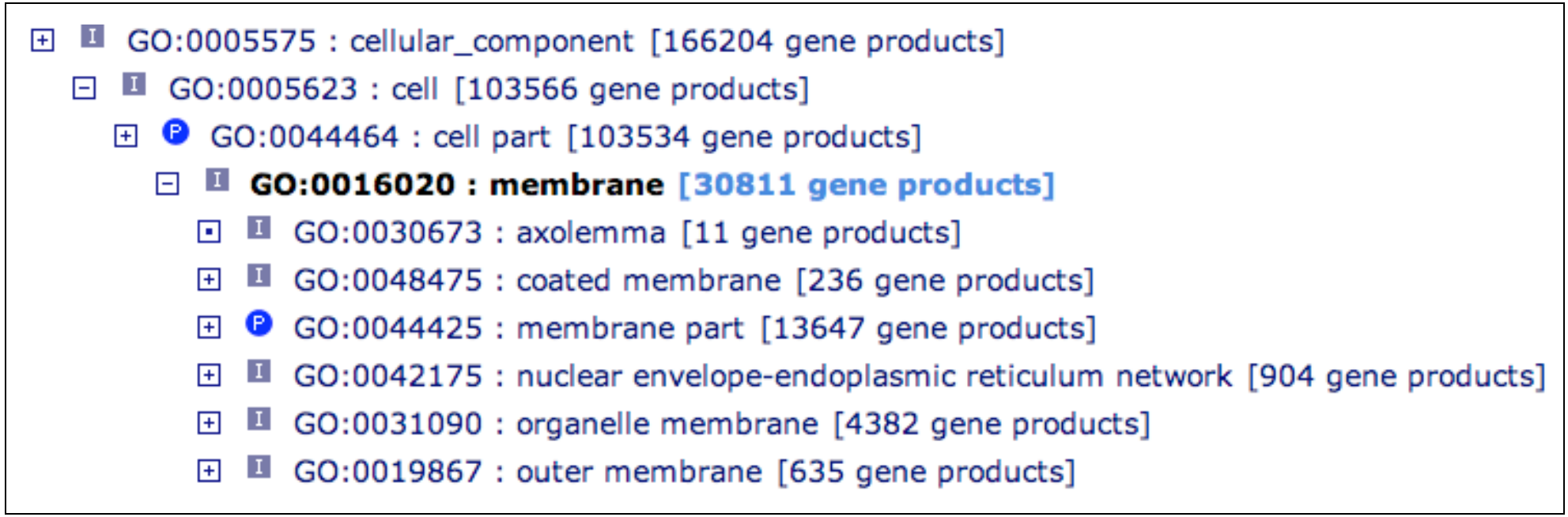
Formal ontology and inheritance

Gene Ontology extract



Part-of relation is distinguished from IS-A relation

Gene Ontology extract



Subtilist classification extract

1 Cell envelope and cellular processes
- 1.1 Cell wall
- 1.2 Transport/binding proteins and lipoproteins
- 1.3 Sensors (signal transduction)
- 1.4 Membrane bioenergetics (electron transport chain and ATP synthase)
- 1.5 Mobility and chemotaxis
- 1.6 Protein secretion
- 1.7 Cell division
- 1.8 Sporulation
- 1.9 Germination
- 1.10 Transformation/competence

The taxinomic link does not denote here an ***is-a***, nor a ***kind-of*** relation.
It means, « is a function related to ».

☺ Useful for gene annotation.
☹ Unexpected side-effect if used for automatic derivation.

MultiFun (E. Coli) classification extract

6 Cell structure
- 6.1 Membrane
- 6.2 Peptidoglycan (murein)
- 6.3 Surface antigens (ECA, O antigen of LPS)
- 6.4 Flagellum
- 6.5 Pilus
- 6.6 Ribosome
- 6.7 Capsule (M and K antigens)

☹ Membrane is defined as a cell structure instead of ***part-of*** the cell structure

Basic Ontology Design Rules:

No conjunction of coordination (or, and)

No *other* (Other functions) or *Miscellaneous*

Use labels that are at most as possible self-sufficient, or at least comprehensible in the hierarchy context. (*Similar to unknown proteins / From B. subtilis ; No similarity*)

Ontology design in Transys

Two short-term main goals

Information retrieval = Bibliographic search with BioAlvis

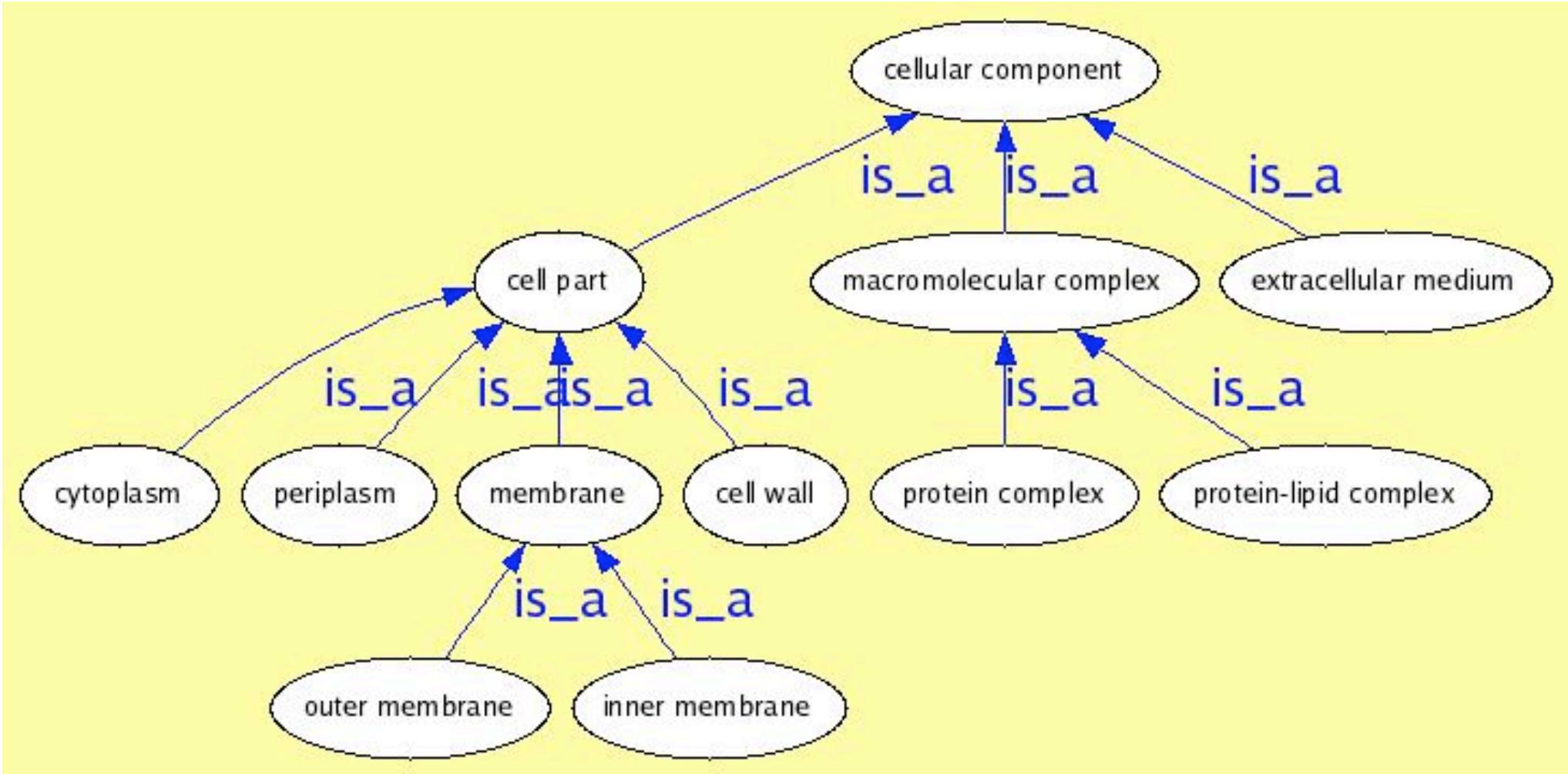
Domain modeling for sharing a common representation among Transys partners

Méthodology

1. Specify the domain and the role of the ontology
2. Acquire the knowledge from biologist expertise, scientific papers, databanks, thesaurus
3. Conceptualize
   - identify the key concept of the ontology, their properties and relations
   - Identify the terms to be used as concept labels
   - Build the structure
4. Integrate / align existing thesaurus / ontology (GO, MeSH, BactriOnto)

Non linear process.

Transys core ontology on cell component



Examples of term extracted from scientific papers

aminoacyl-tRNA synthetase	antibiotic production	metabolism of sulfur
cell division cycle	cell division	motility and chemotaxis
elongation of protein synthesis	cell-wall	phosphate metabolism
elongation of RNA synthesis	detoxification	protein folding
enzyme synthesis	DNA recombination	protein modification
initiation of protein synthesis	DNA restriction-modification and	protein secretion
initiation of RNA synthesis	repair	protein synthesis
late sporulation	DNA-packaging and segregation	RNA modification
main glycolytic pathway	germination	RNA synthesis
phosphorylation-dephosphorylation	membrane bioenergetics	sporulation
proteolytic processing	metabolism of amino-acid and related molecule	transformation-competence
regulation of RNA synthesis	metabolism of carbohydrate and	transport-binding protein and lipoprotein
ribosomal protein	metabolism of coenzyme and	transposon and IS
TCA cycle	prosthetic group	cell envelope and cellular process
termination of protein synthesis	metabolism of lipid	information pathway
termination of RNA synthesis	metabolism of nucleotide and	intermediary metabolism
adaptation to atypical condition	nucleic acid	

Ontology population ?