AMIA 2006: Biomedical and Health Informatics
November 11-15, 2006 Hilton Washington & Towers Washington, DC

Towards a Top-Level Ontology for Molecular Biology

Stefan Schulz [1], Elena Beisswanger [2],
Udo Hahn [2], Joachim Wermter [2],

[1] Freiburg University Hospital, Department of Medical Informatics, Germany
[2] Jena University Language and Information Engineering (JULIE) Lab

Boot Strep

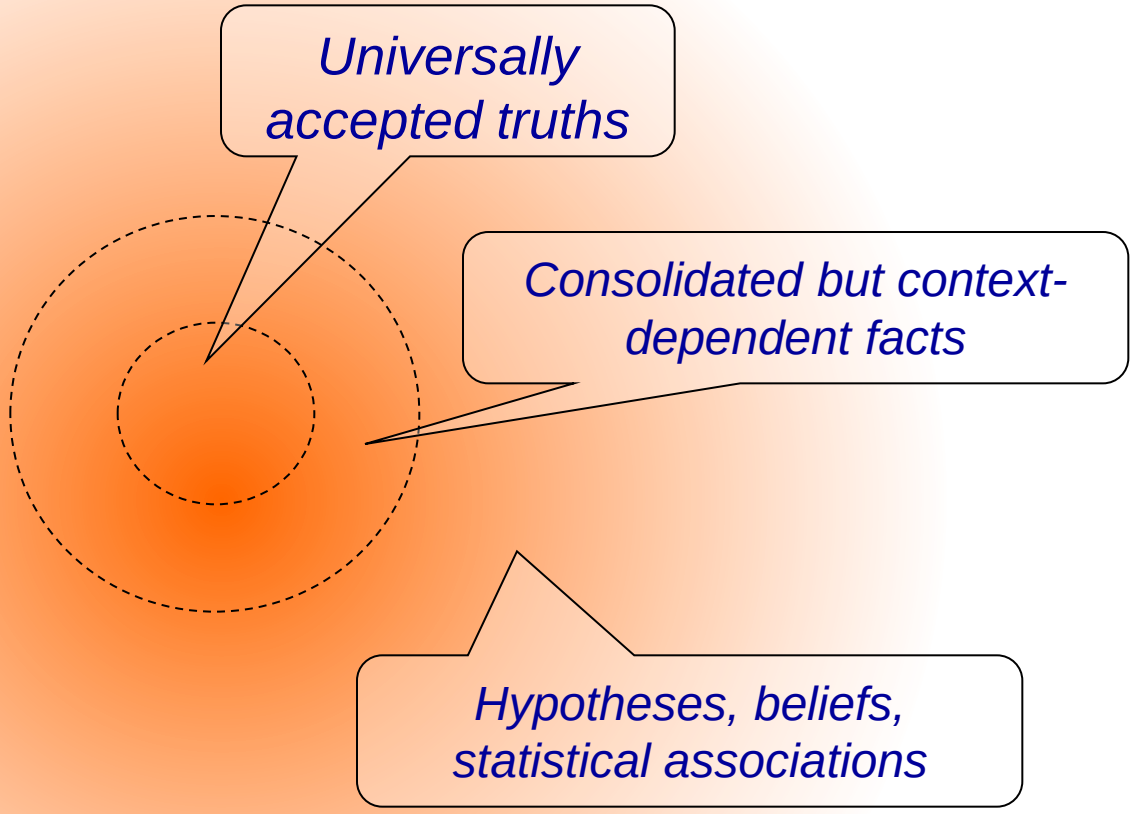
StemNet

Information Society Technologies

Semantic Mining
—— *Semantic Interoperability and Data Mining in Biomedicine*

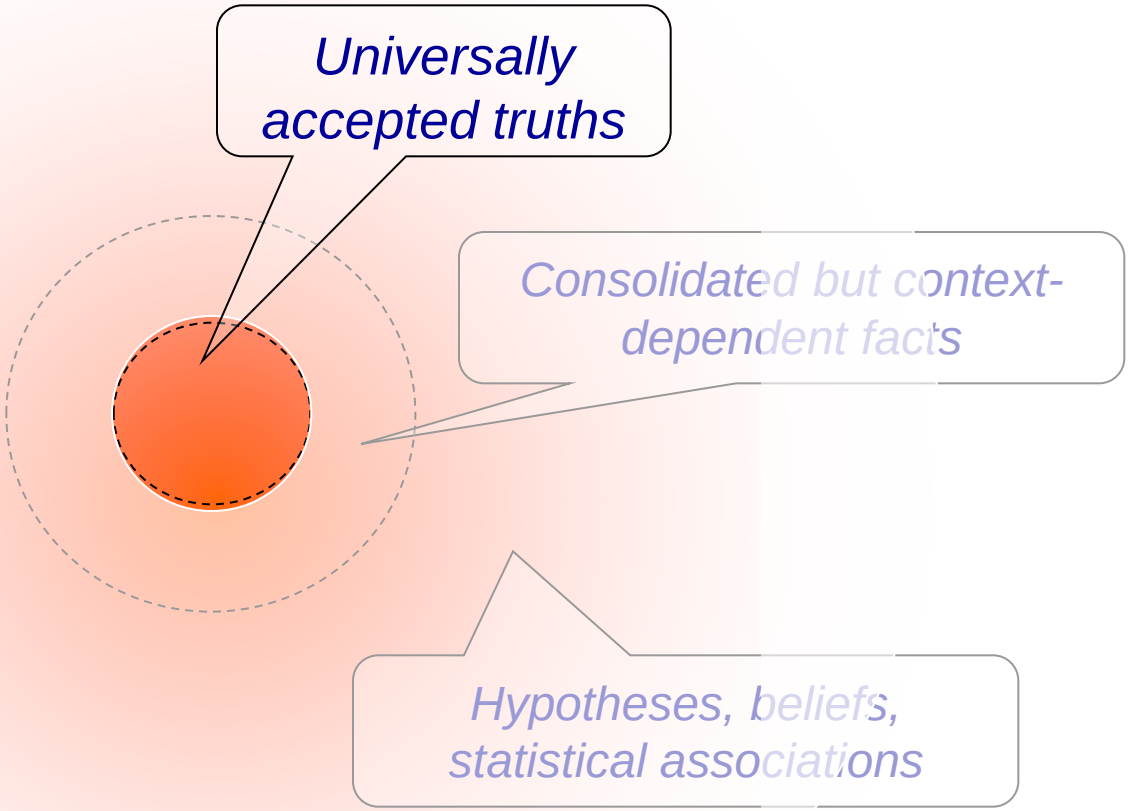
What's an ontology…?

Our notion of ontology



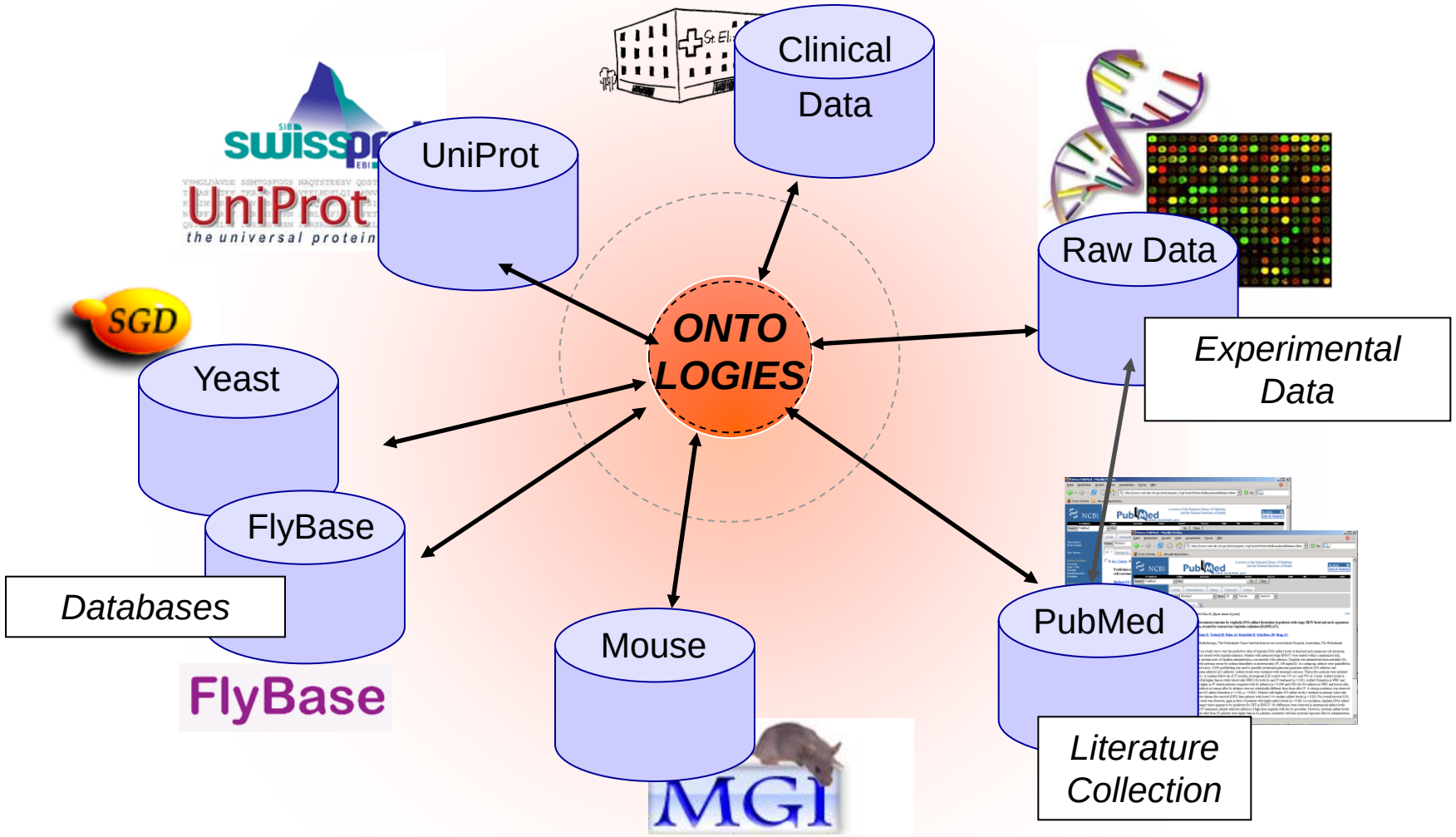
Domain Knowledge

Ontology !

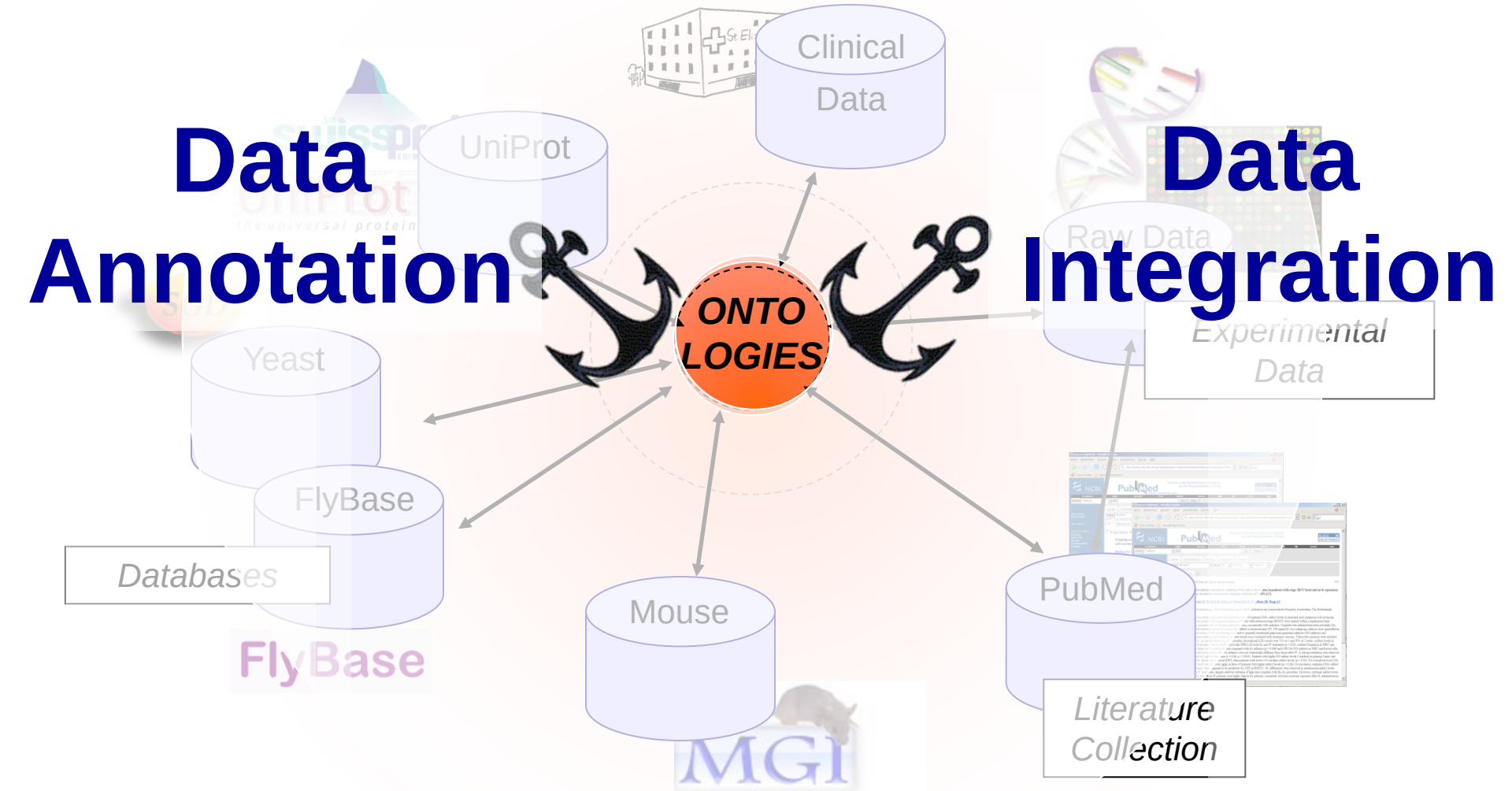


Domain Knowledge

Data Sources in Biomedicine

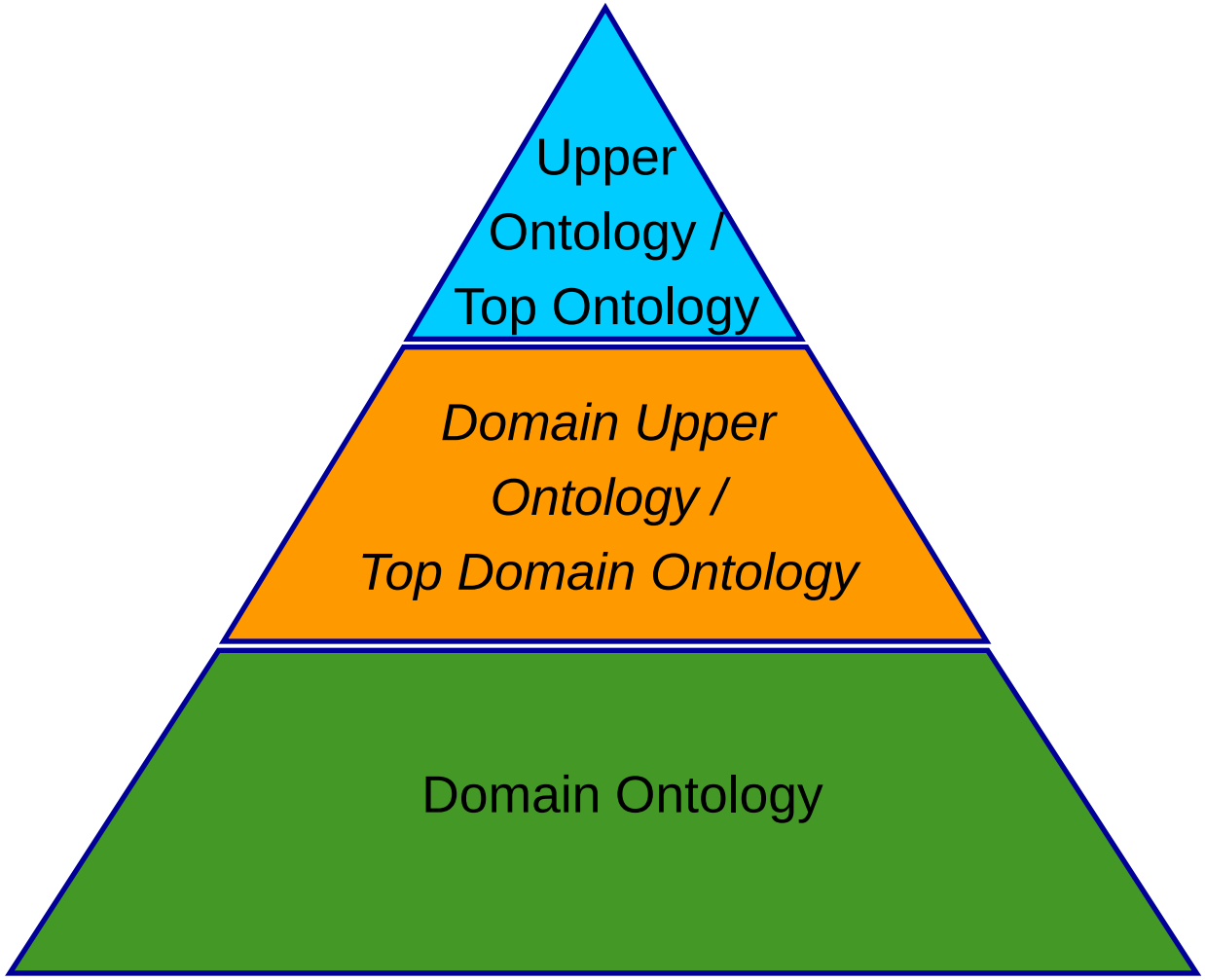


Bioontologies are devised to support

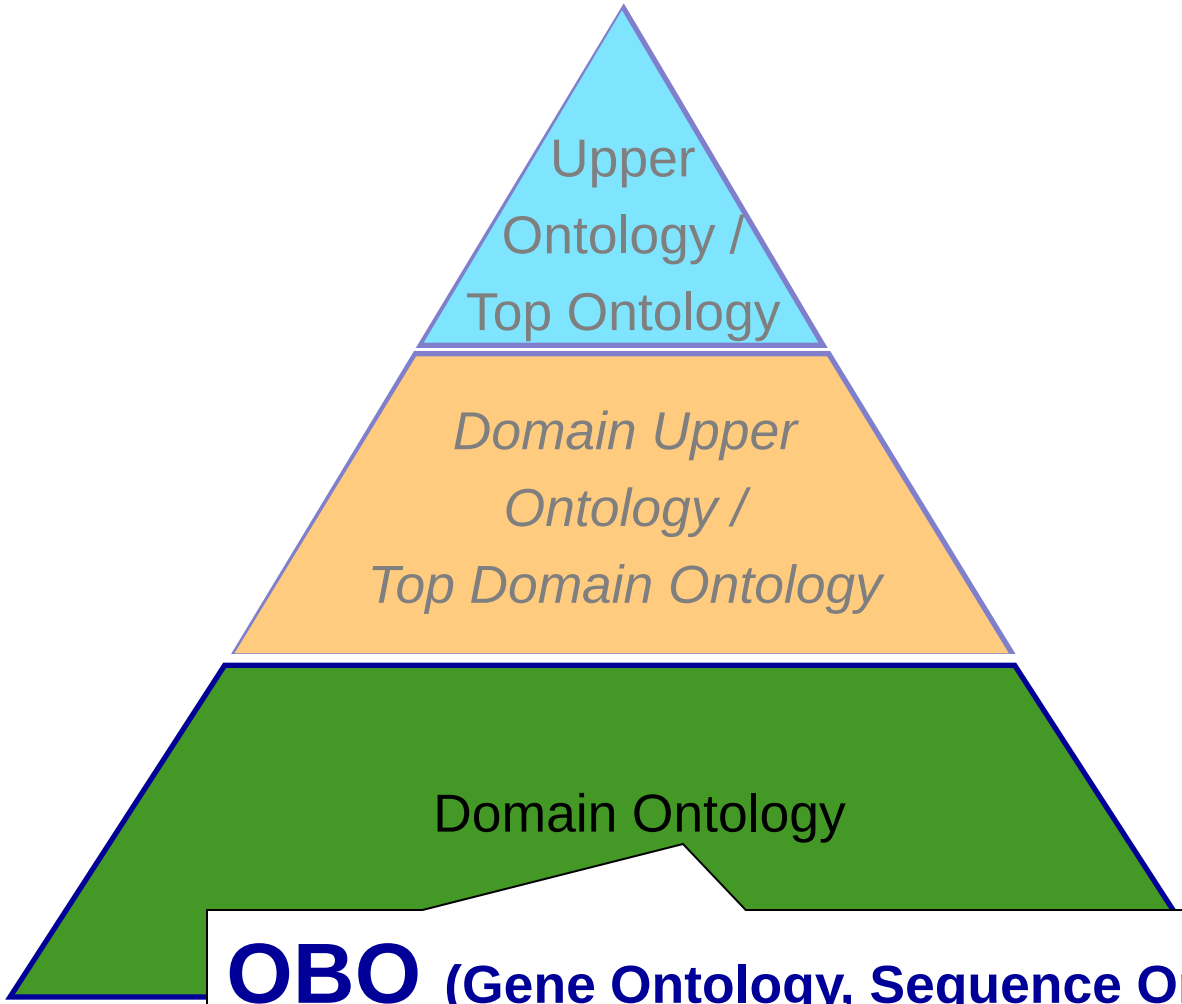


How can ontologies be structured?

Ontological Layers

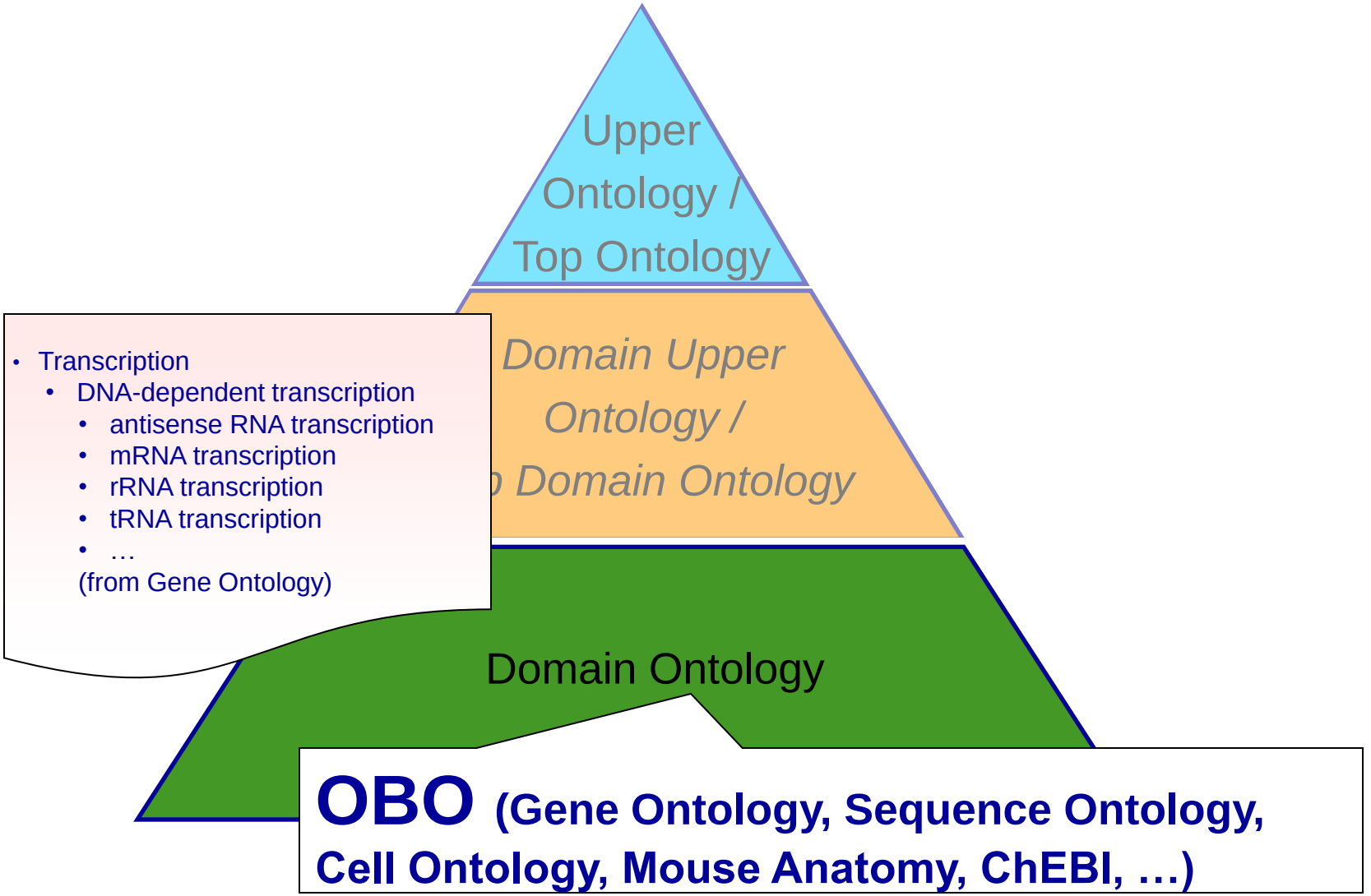


Ontological Layers

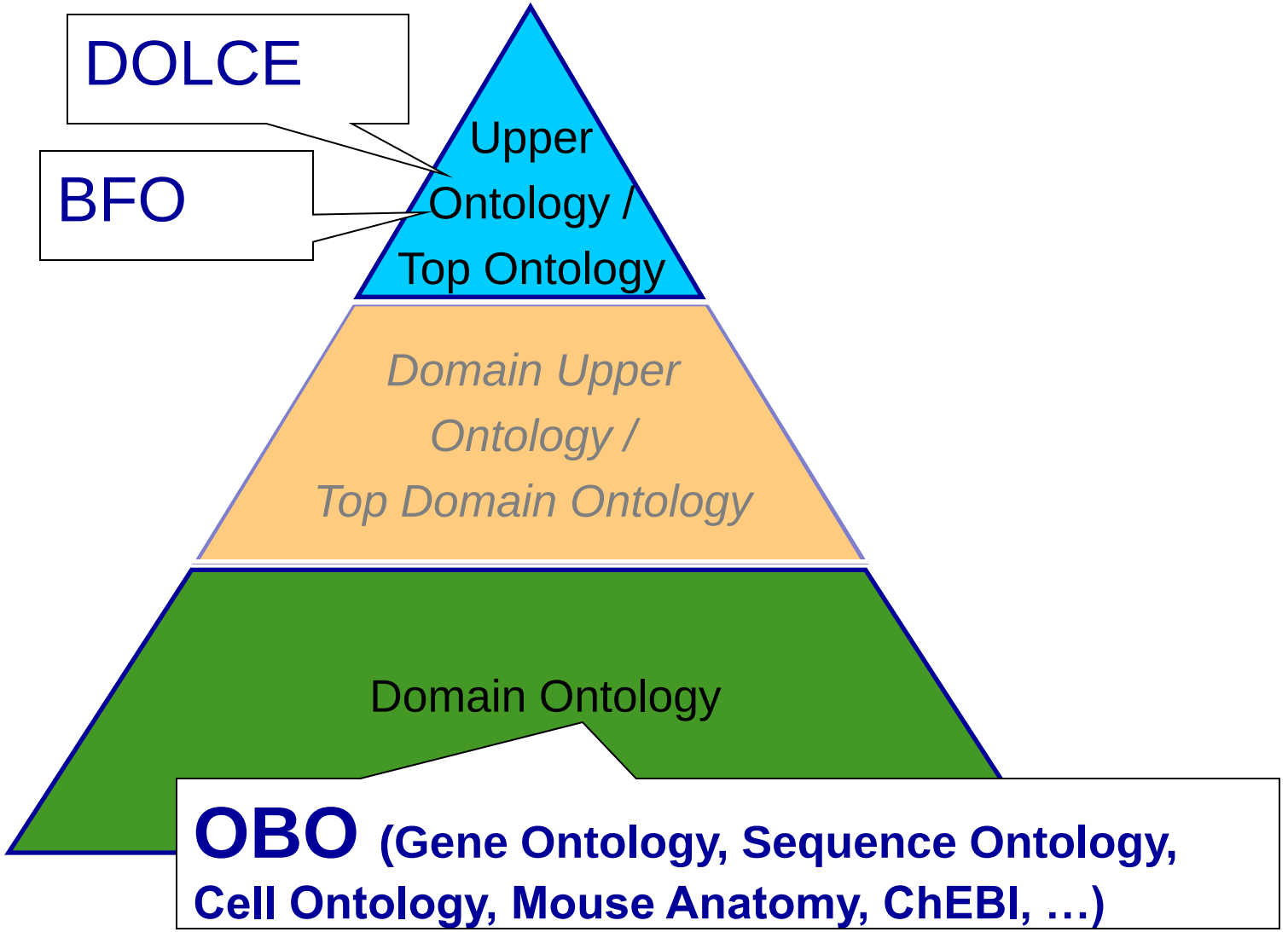


OBO (Gene Ontology, Sequence Ontology, Cell Ontology, Mouse Anatomy, ChEBI, …)

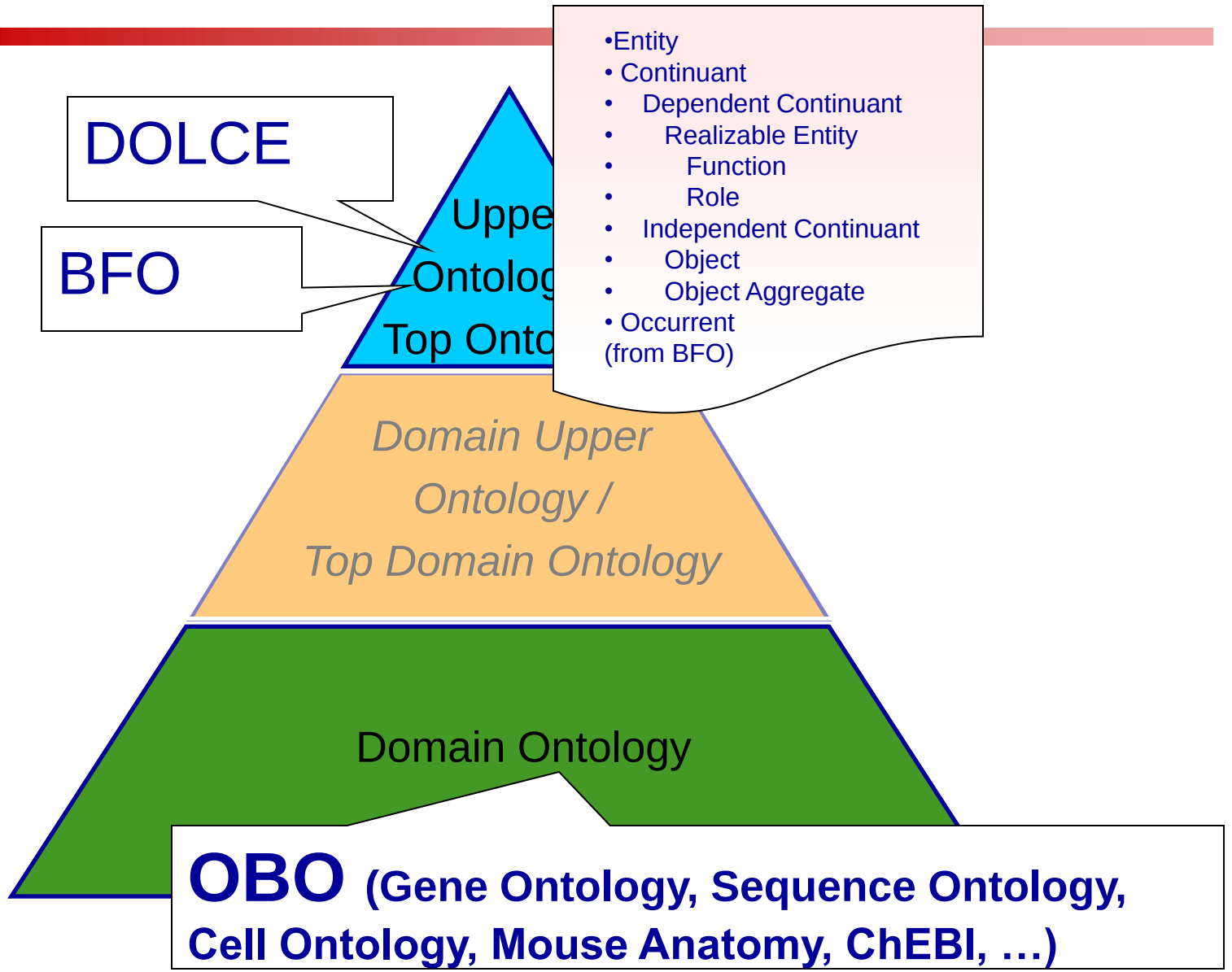
Ontological Layers



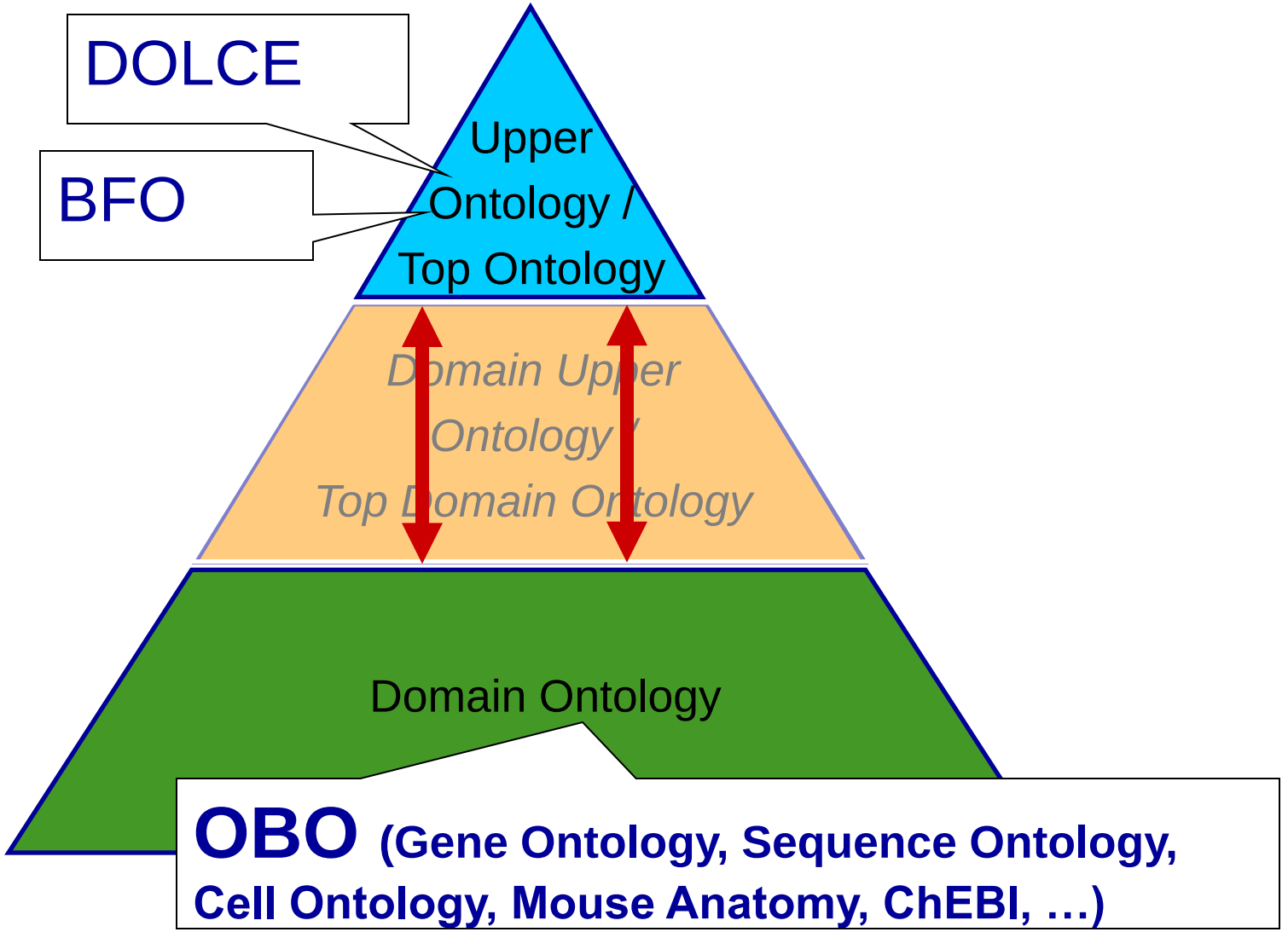
Ontological Layers



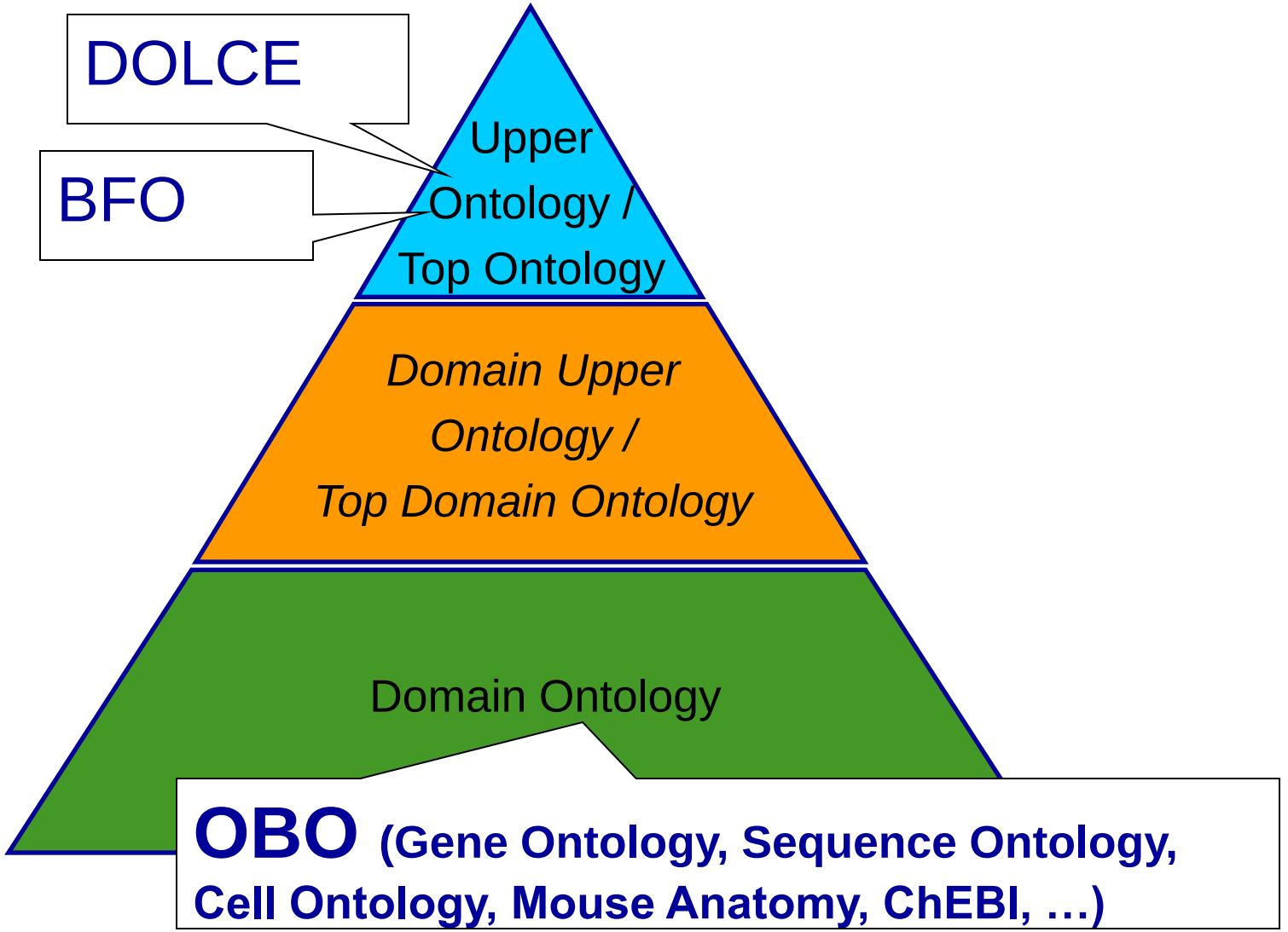
Ontological Layers



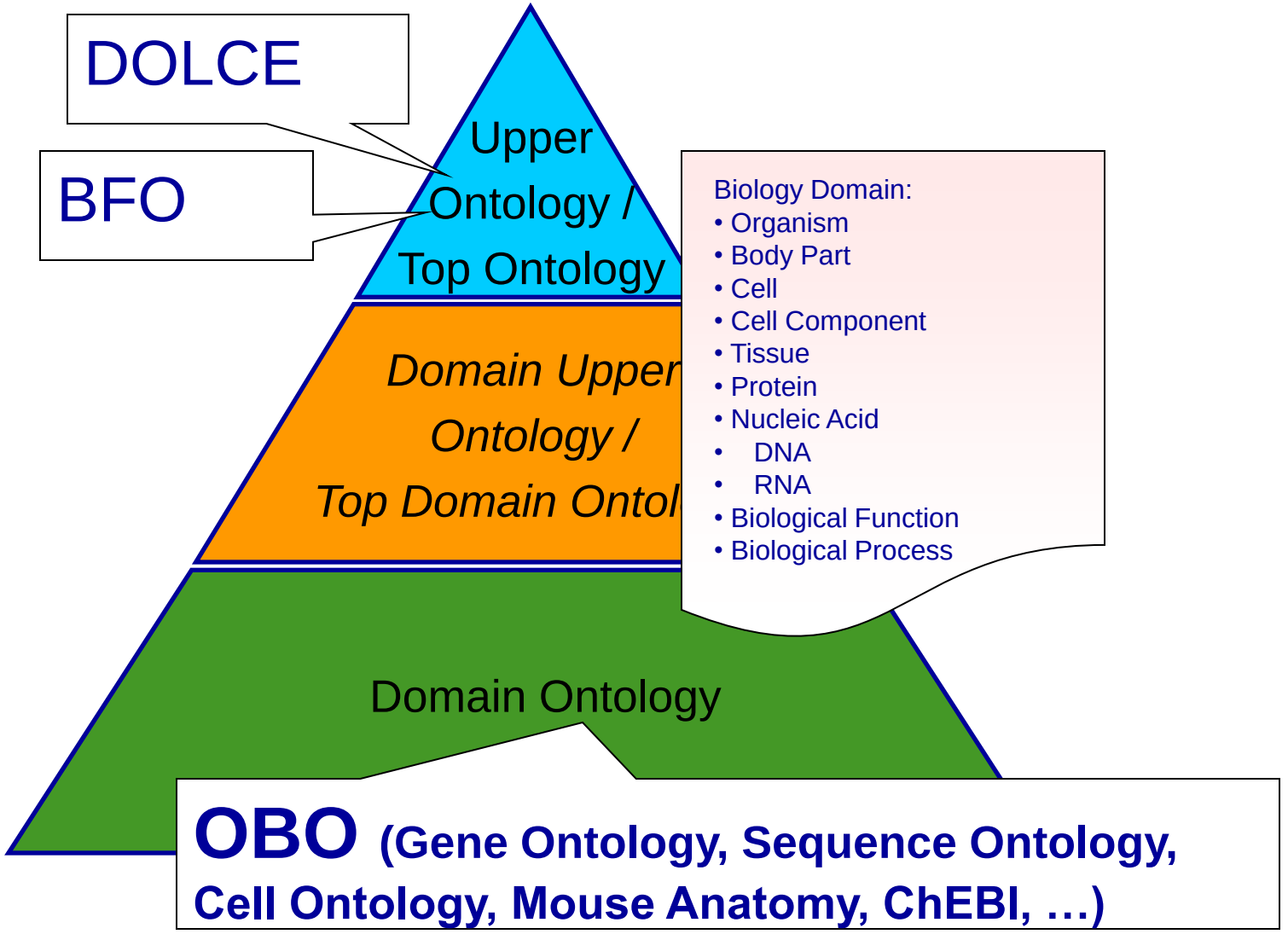
Ontological Layers



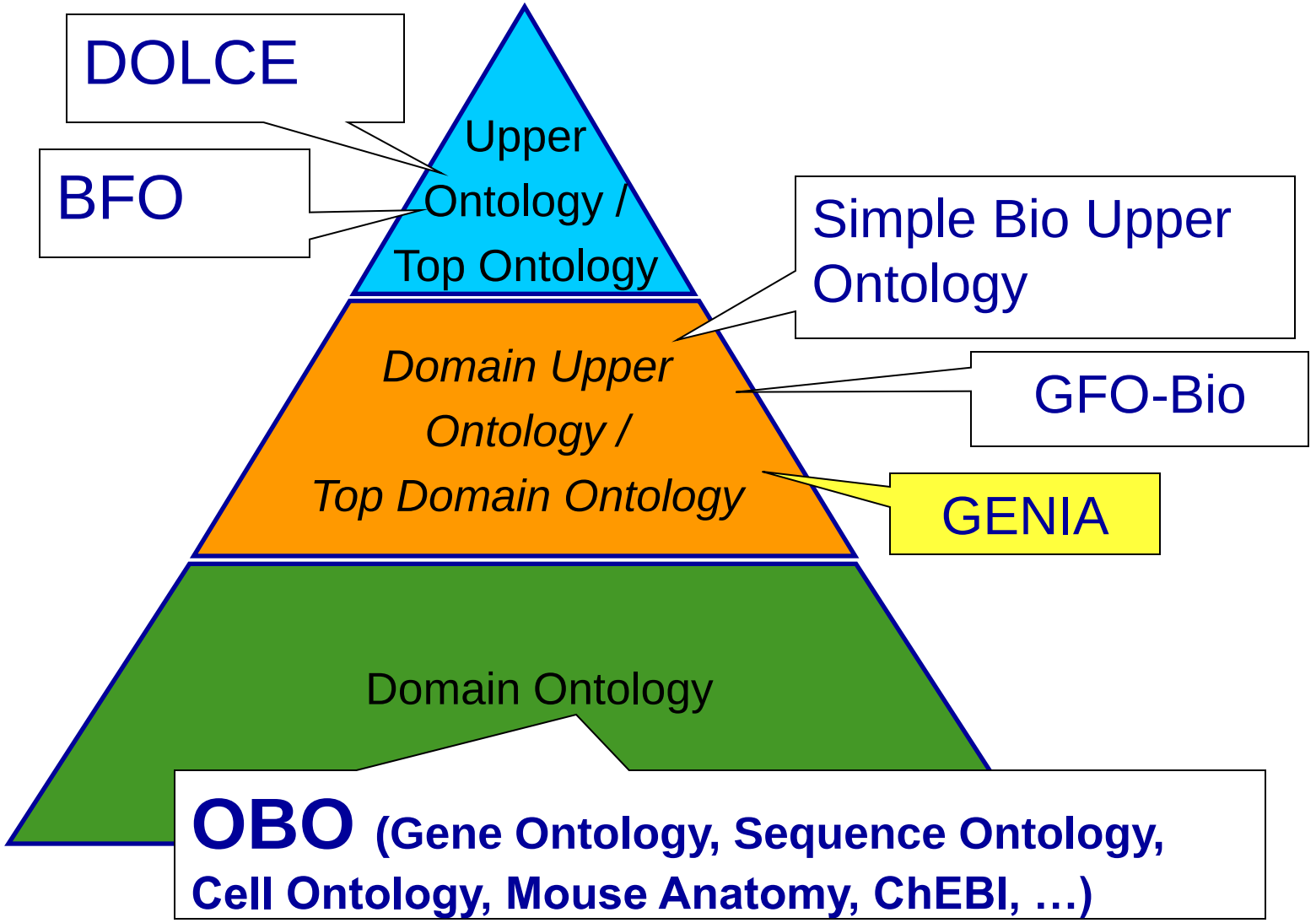
Ontological Layers



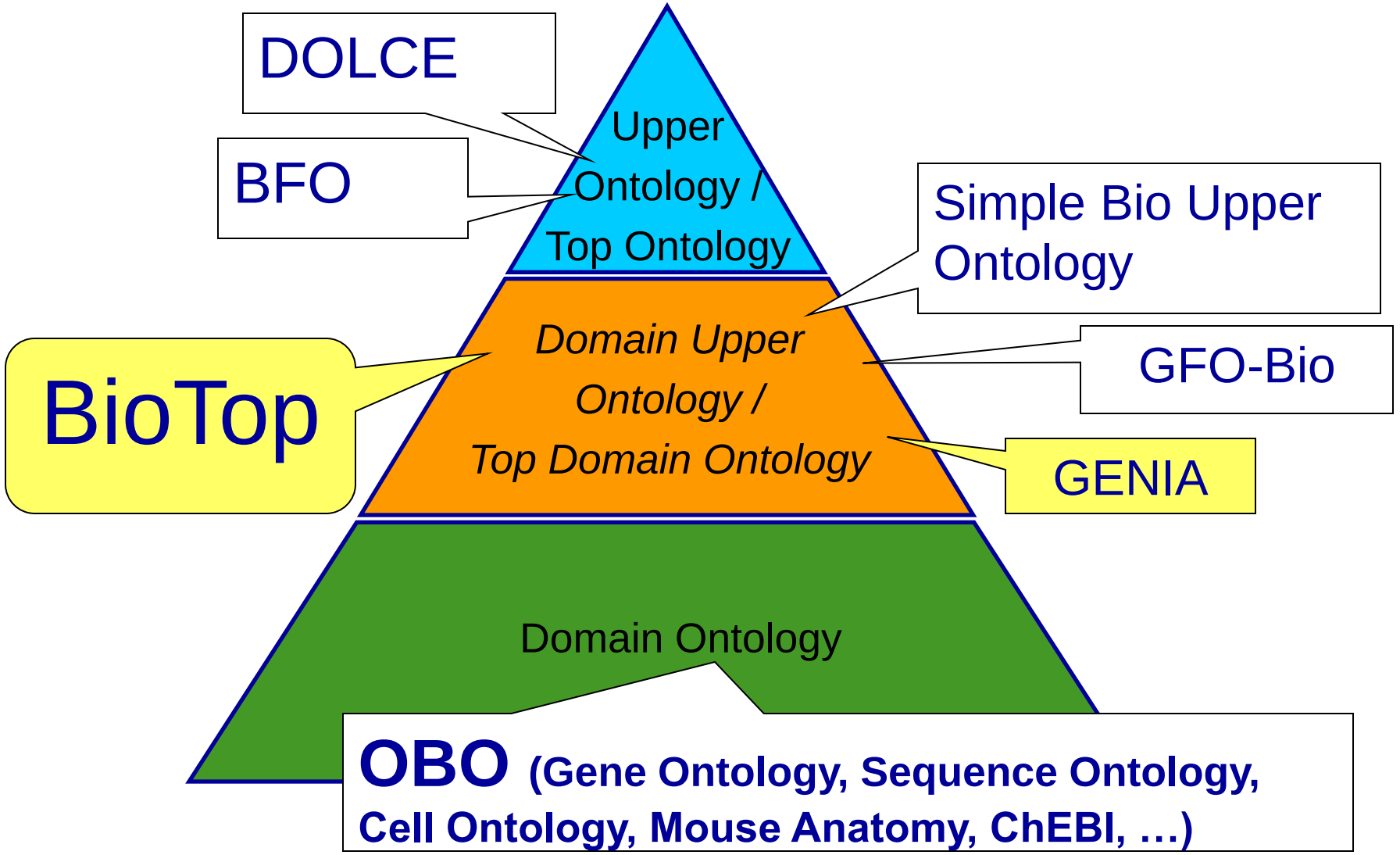
Ontological Layers



Ontological Layers



Ontological Layers



GENIA as example for top level domain ontology.

The GENIA Ontology

"The GENIA ontology is intended to be a formal model of cell signaling reactions in human. It is to be used as a basis of thesauri and semantic dictionaries for natural language processing applications [...]

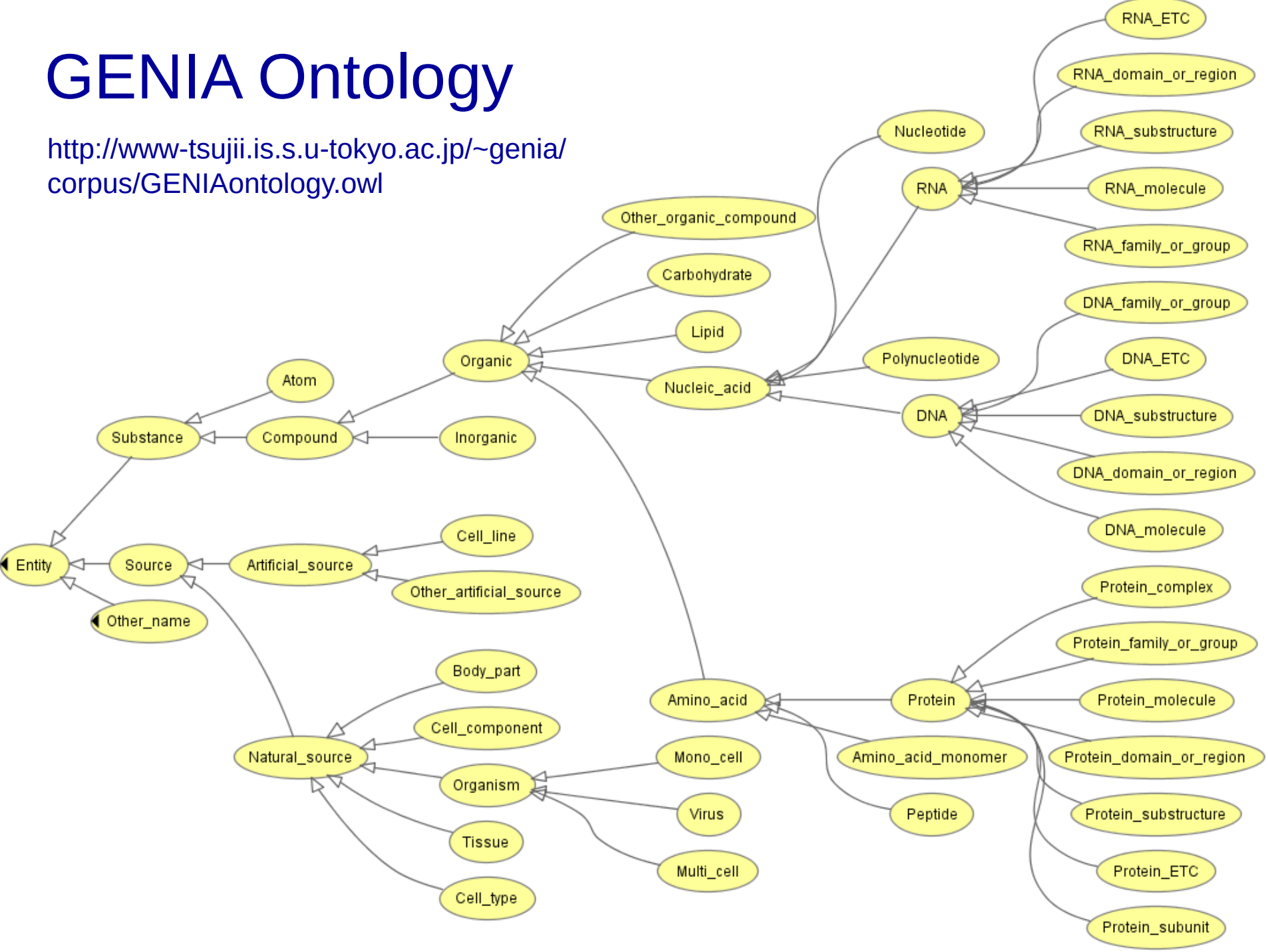
Another use of the GENIA ontology is to provide a basis for integrated view of multiple databases"

http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html

- Developed at Tsujii Laboratory, University Tokyo
- Taxonomy, 48 classes
- Verbal definitions ("scope notes")
- No relations other than subclass relations
- Purpose: Semantic annotation of biological papers

GENIA Ontology

http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/corpus/GENIAontology.owl



GENIA Critique

Insufficient Definitions

- Amino Acid Monomer:

  "An amino acid monomer, e.g. tyrosine, serine, tyr, ser"

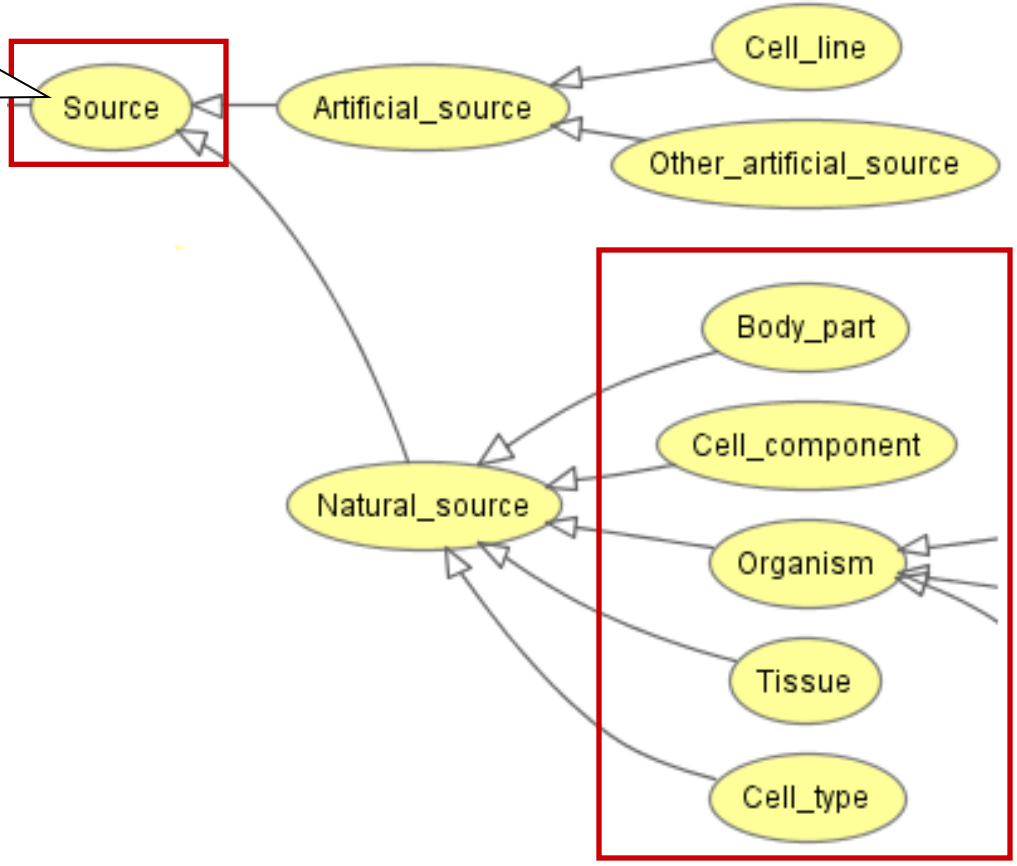
- DNA:

  "DNAs include DNA groups, families, molecules, domains, and regions"

False Taxonomic Parent

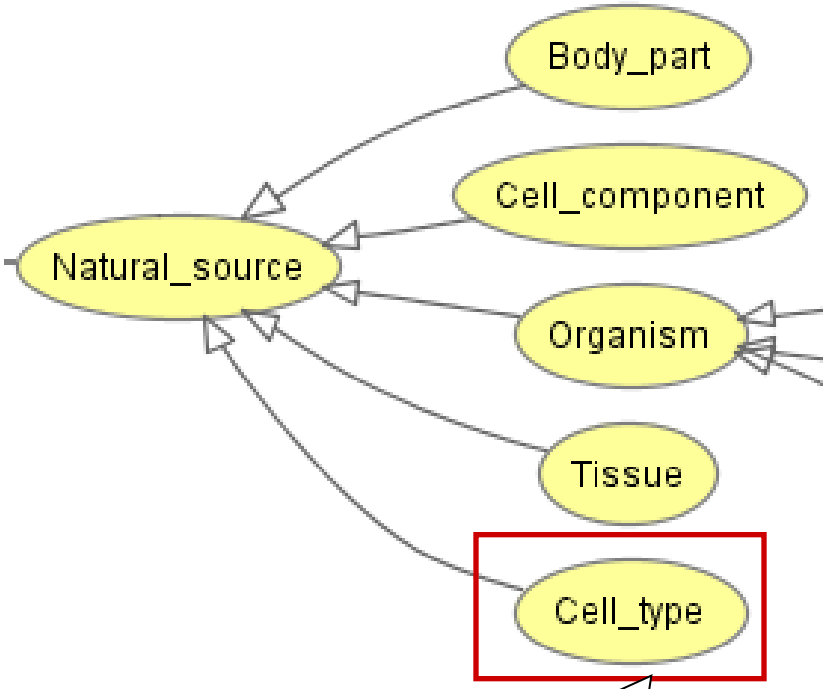
GENIA: Source

"Sources are biological locations where substances are found …"



- 'Organism', 'Tissue', … are objects and not primarily sources!
- 'Source' is a role …

Misconception of “Type”
GENIA: Cell Type

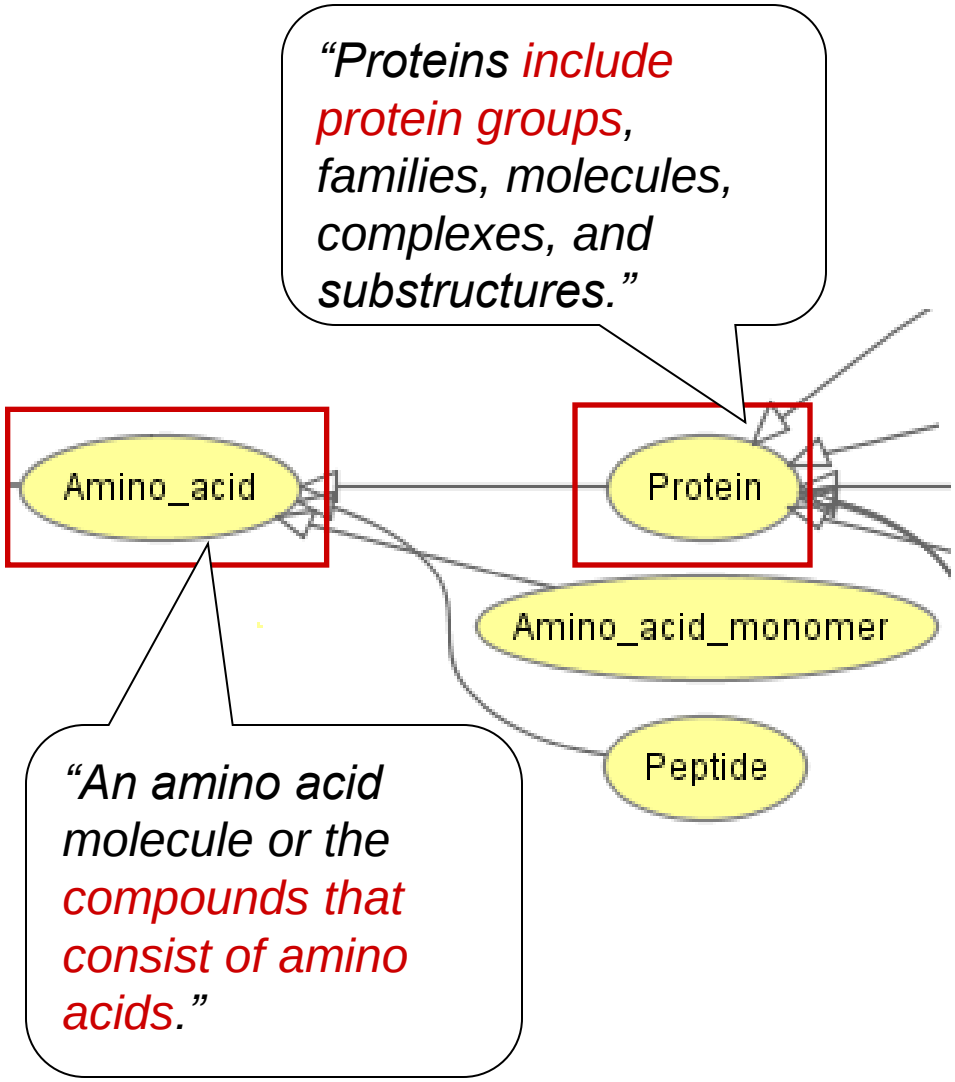


“A cell type, e.g. T-Lymphocyte, T-cell, astrocyte, fibroblast”

- Are the instances of ‘Cell *Type*’ classes ?
- Otherwise rename the class as ‘Cell’!

Non-Conformant Naming Policy

GENIA: Amino Acid, GENIA: Protein



- Names suggest single molecules
- Don't work against biologists' intuition!

GENIA Redesign:

BioTop

BioTop

- Ontologically founded Top level ontology for biology
- Extensive use of formal relations (OBO)
- Formal semantics (OWL-DL)
- Use as a semantic glue between existing ontologies
- Scope: as GENIA (in a 1st phase)

BioTop Relations

- *partOf*
- *properPartOf*
- *locatedIn*
- *derivesFrom*
- *hasParticipant*

BioTop Relations

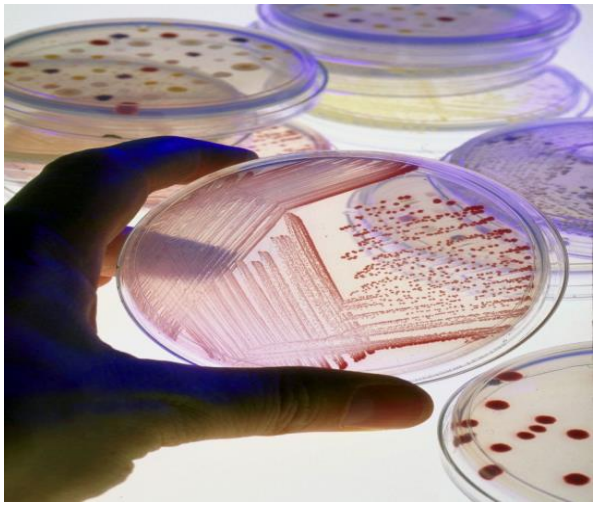
- *partOf*
- *properPartOf*
- *locatedIn*
- *derivesFrom*
- *hasParticipant*
- *hasFunction (functionOf)*

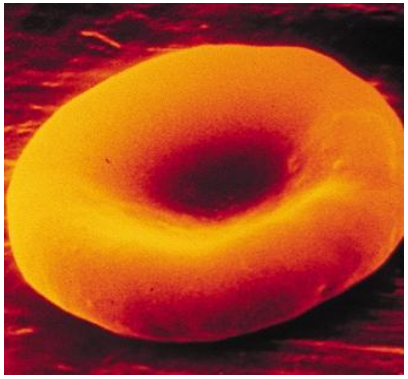
BioTop Relations

- *partOf*
  - *grainOf (hasGrain)*
  - *componentOf (hasComponent)*
- *properPartOf*
- *locatedIn*
- *derivesFrom*
- *hasParticipant*
- *hasFunction (functionOf)*

Collectives and *hasGrain*

- *hasGrain* non-transitive subrelation of *hasPart*
- Example: “Collective of Cell *hasGrain* only Cell”

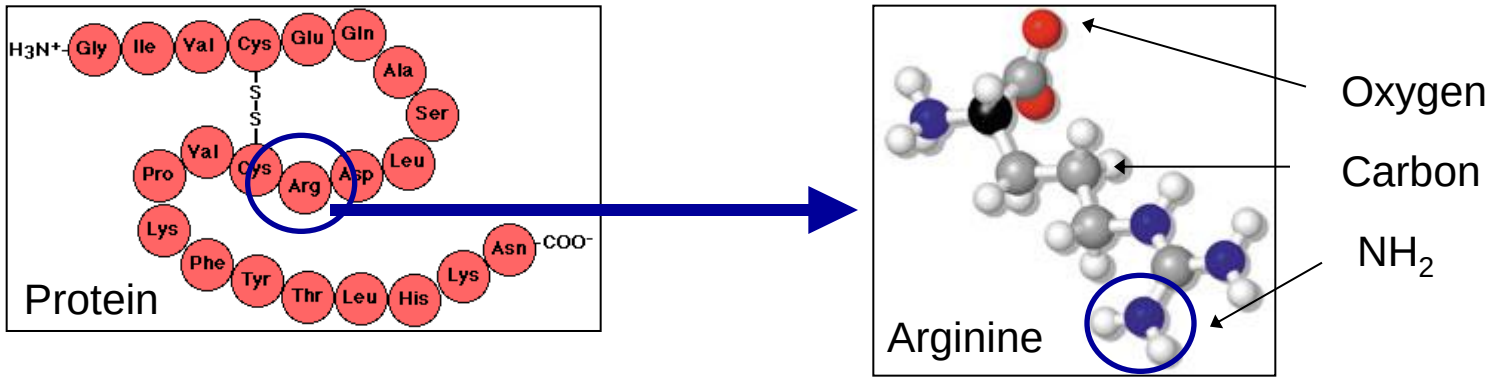




- Collectives can gain or lose grains without changing their identity

Compounds and *hasComponent*

- *hasComponent* non-transitive subrelation of *hasPart*
- Example: Definition of Protein



- Compounds are determined by their components

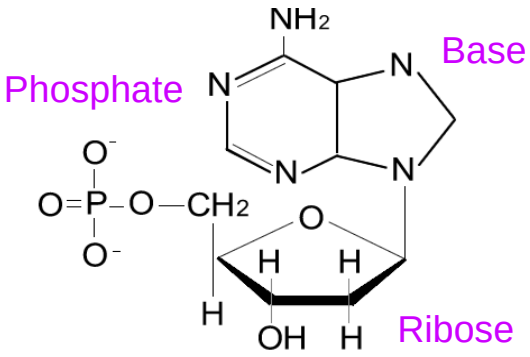
Full Definitions

- Classes defined by necessary and sufficient conditions whenever possible
- Rationales
  - Precise understanding of meaning
  - Empowering the classifier for automated validation processes
- Required introduction of new classes, e.g. Phosphate, Ribose

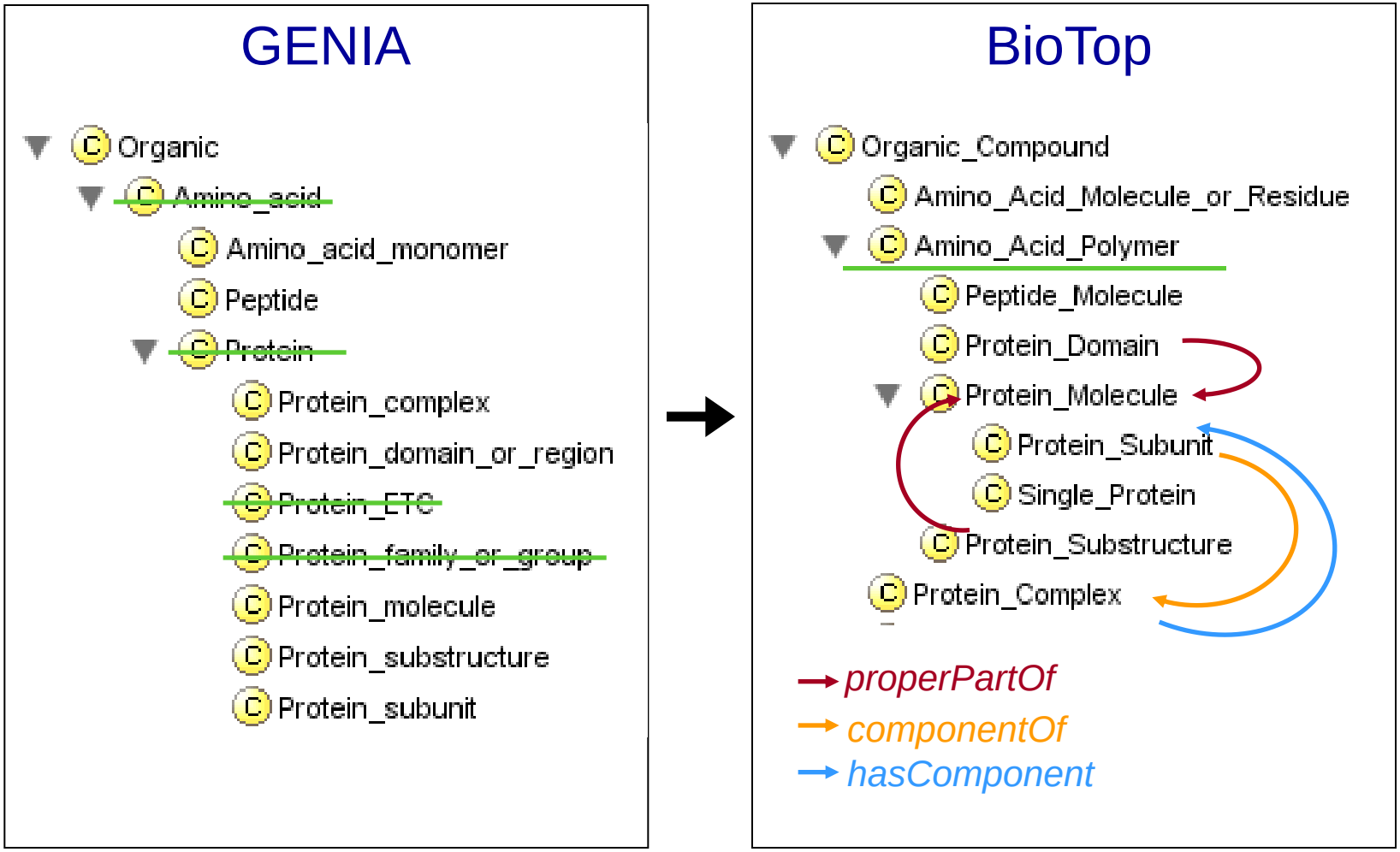
Sufficient conditions for class ‘Nucleotide’

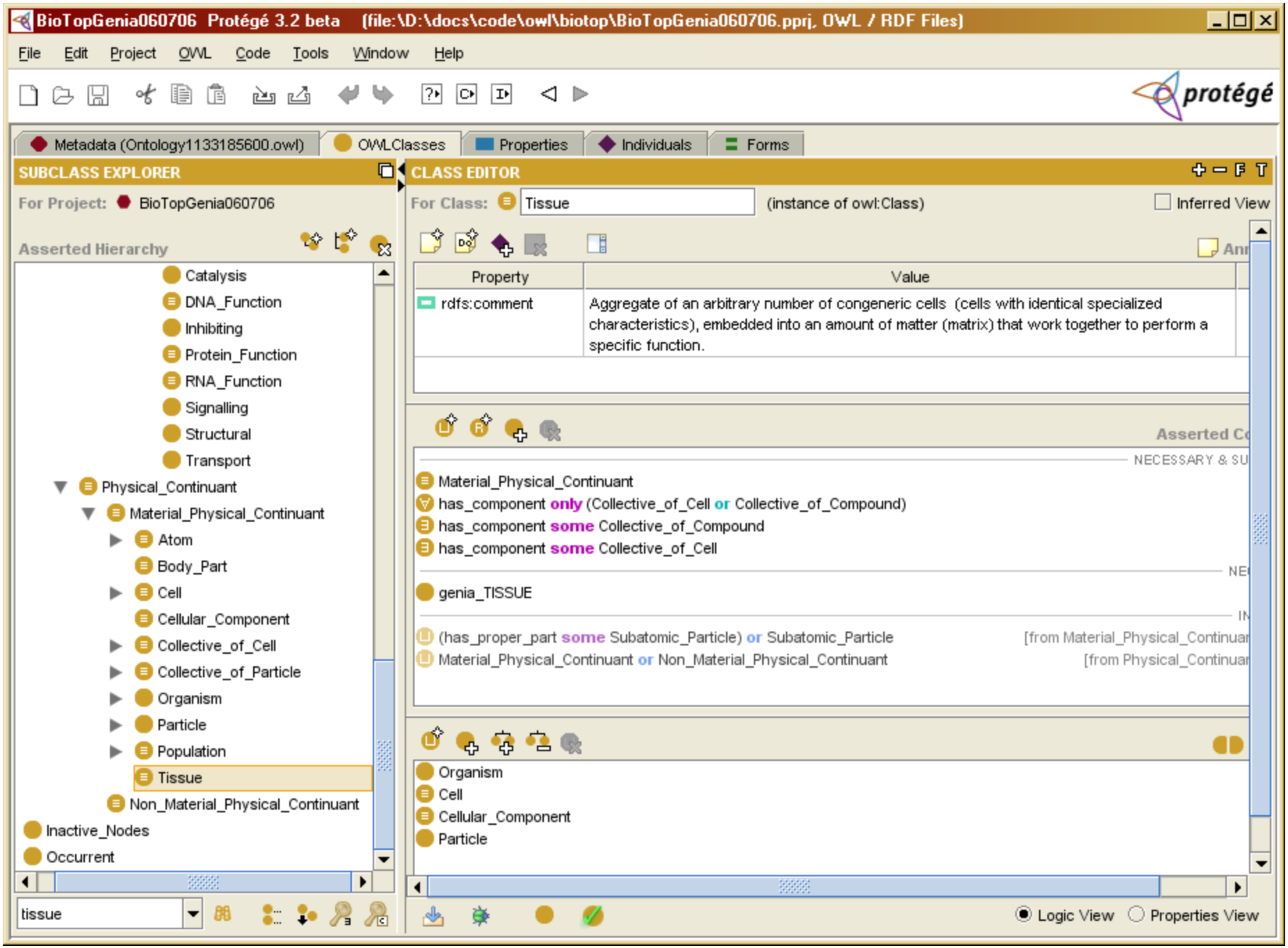
- ∀ hasComponent **only** (Phosphate **or** Ribose **or** HeterocyclicBase)
- = hasComponent **exactly** 1 Phosphate
- = hasComponent **exactly** 1 Ribose
- = hasComponent **exactly** 1 HeterocyclicBase

Nucleotide



Rearranged Classes




BioTopGenia060706 Protégé 3.2 beta (file:\D:\docs\code\owl\biotop\BioTopGenia060706.pprj, OWL / RDF Files)
File Edit Project OWL Code Tools Window Help
protégé
Metadata (Ontology1133185600.owl)
OWLClasses
Properties
Individuals
Forms
SUBCLASS EXPLORER
For Project: BioTopGenia060706
Asserted Hierarchy
Catalysis
DNA_Function
Inhibiting
Protein_Function
RNA_Function
Signalling
Structural
Transport
Physical_Continuant
Material_Physical_Continuant
Atom
Body_Part
Cell
Cellular_Component
Collective_of_Cell
Collective_of_Particle
Organism
Particle
Population
Tissue
Non_Material_Physical_Continuant
Inactive_Nodes
Occurrent
tissue
CLASS EDITOR
For Class: Tissue
(instance of owl:Class)
Inferred View
Property
Value
rdfs:comment
Aggregate of an arbitrary number of congeneric cells (cells with identical specialized characteristics), embedded into an amount of matter (matrix) that work together to perform a specific function.
Asserted Co
NECESSARY & SU
Material_Physical_Continuant
has_component only (Collective_of_Cell or Collective_of_Compound)
has_component some Collective_of_Compound
has_component some Collective_of_Cell
genia_TISSUE
(has_proper_part some Subatomic_Particle) or Subatomic_Particle
[from Material_Physical_Continuan
Material_Physical_Continuant or Non_Material_Physical_Continuant
[from Physical_Continuan
Organism
Cell
Cellular_Component
Particle
Logic View
Properties View

BioTop Figures

- **`BioTop.owl`**:
  - 143 non-taxonomic relation instances (seven relation types)
  - 128 classes

- **`BioTopGenia.owl`**:
  - Imports GENIA
  - Maps BioTop classes to GENIA classes

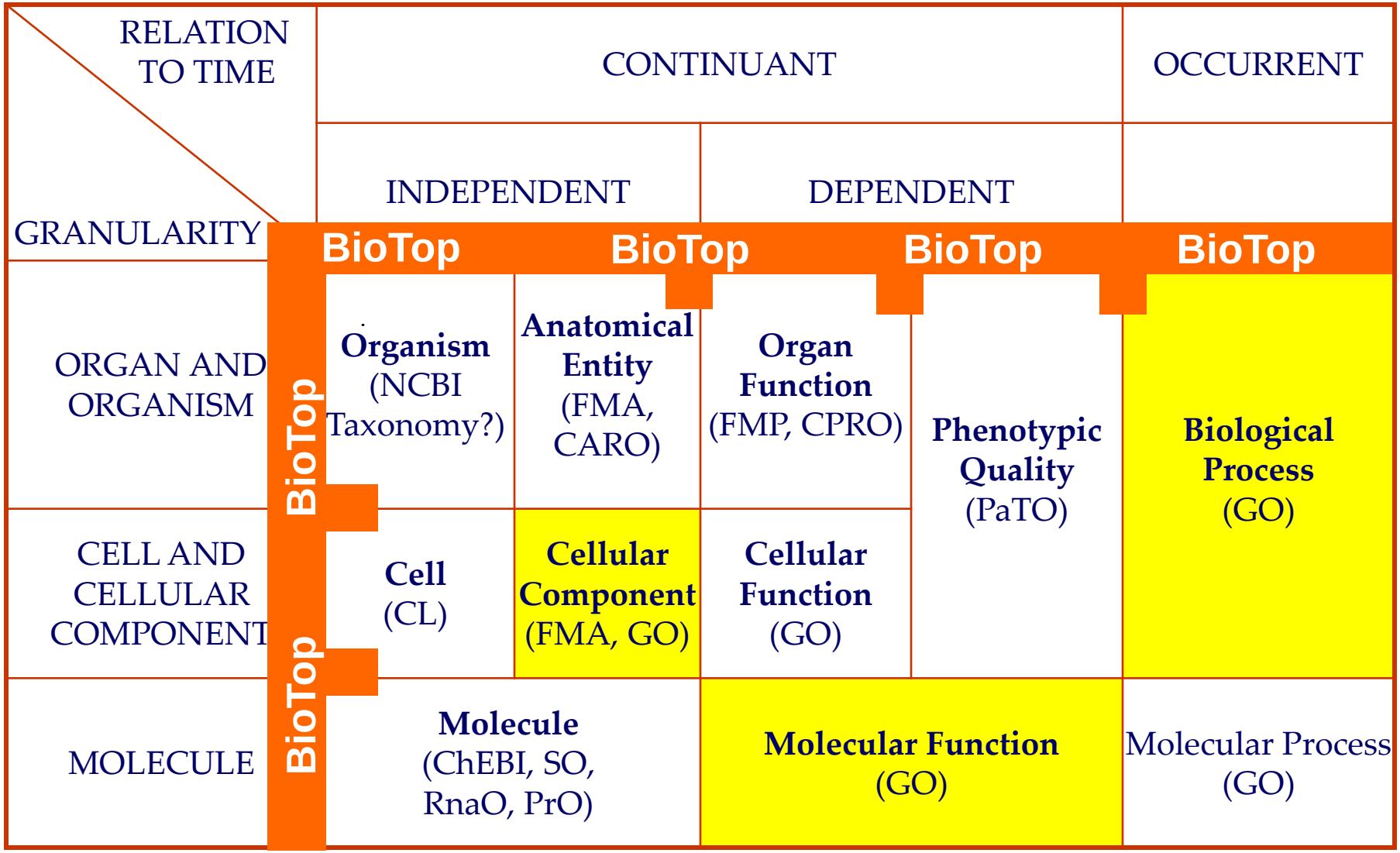
BioTop: Interfacing with OBO Ontologies

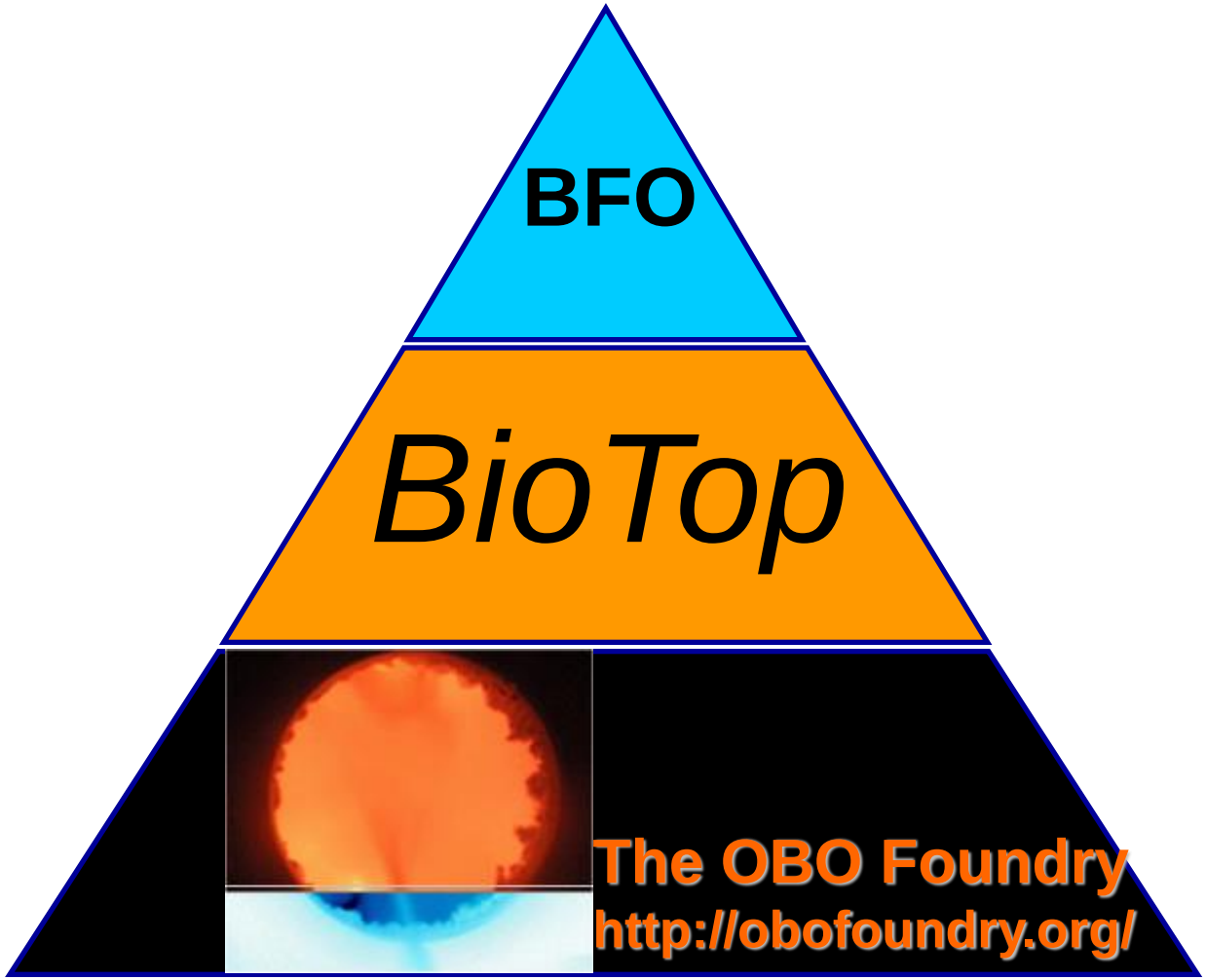
BioTop		OBO Ontologies
Biological Process	↔	Biological Process$_{Gene\ Ontology}$
Protein Function	↔	Molecular Function$_{Gene\ Ontology}$
Cell Component	↔	Cellular Component$_{Gene\ Ontology}$
Cell	↔	Cell$_{Cell\ Ontology}$ and Cell$_{FMA}$
Atom	↔	Atoms$_{ChEBI}$
Organic Compound	↔	Organic Molecular Entities$_{ChEBI}$
Tissue	↔	Tissue$_{FMA}$
DNA, RNA	↔	DNA$_{Sequence\ Ontology}$, RNA$_{Sequence\ Ontology}$
Protein	↔	Protein$_{Sequence\ Ontology}$

Proposal: BioTop to enrich OBO Foundry

RELATION TO TIME / GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	**Organism** (NCBI Taxonomy?)	**Anatomical Entity** (FMA, CARO)	**Organ Function** (FMP, CPRO)	**Phenotypic Quality** (PaTO)	**Biological Process** (GO)
CELL AND CELLULAR COMPONENT	**Cell** (CL)	**Cellular Component** (FMA, GO)	**Cellular Function** (GO)		
MOLECULE	**Molecule** (ChEBI, SO, RnaO, PrO)		**Molecular Function** (GO)		Molecular Process (GO)

OBO Foundry and BioTop

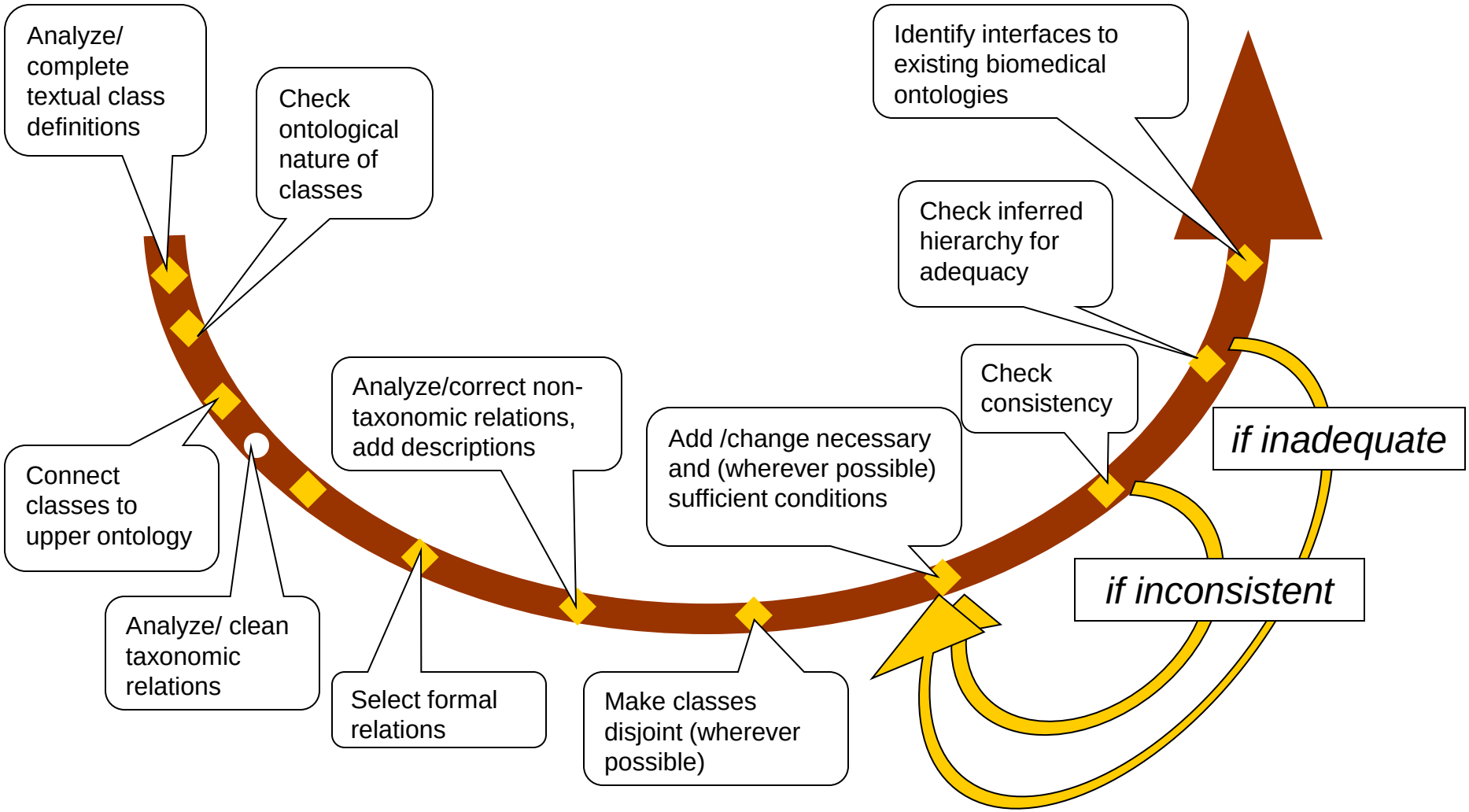



BFO
BioTop
The OBO Foundry
http://obofoundry.org/

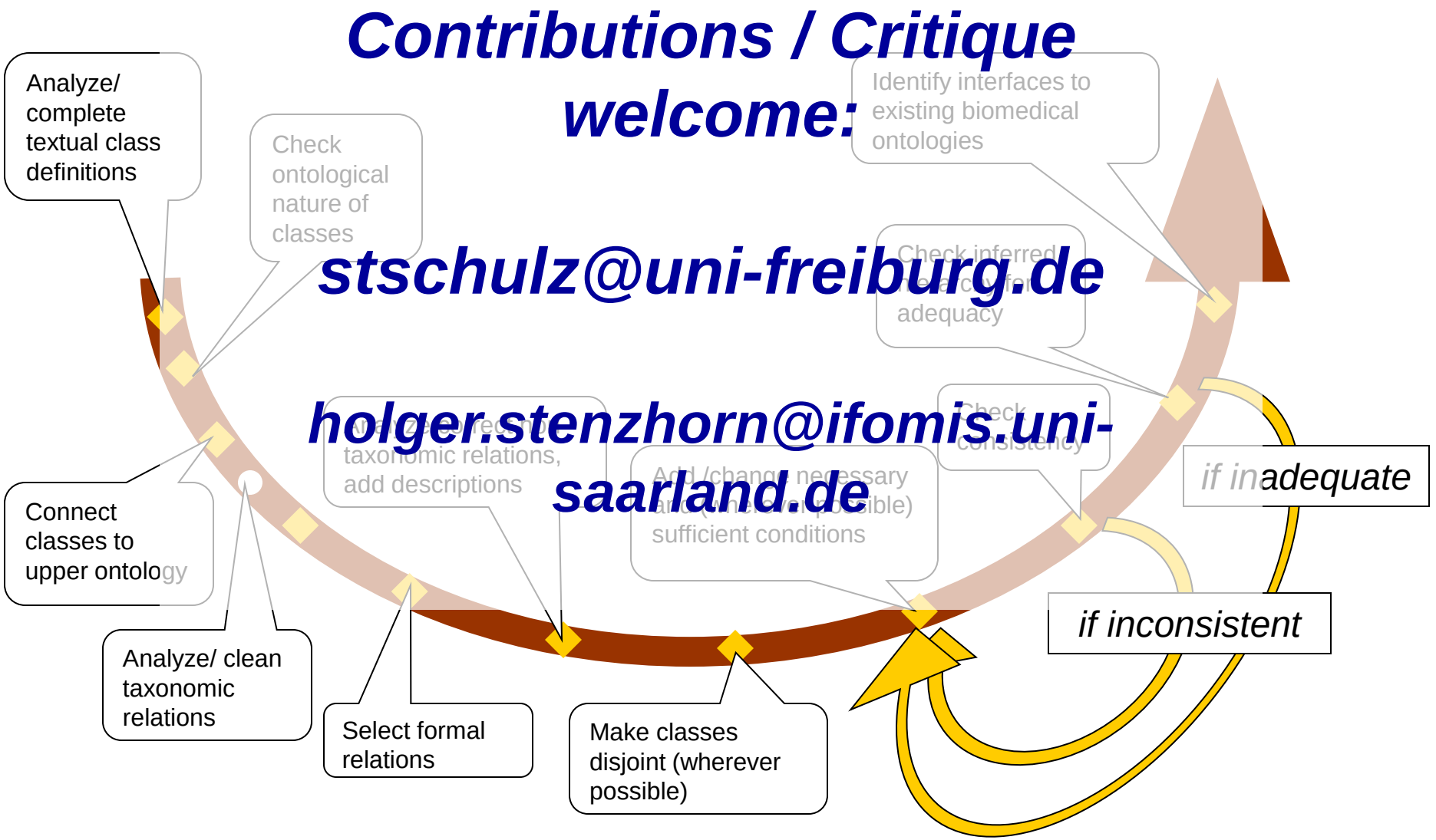
Outlook

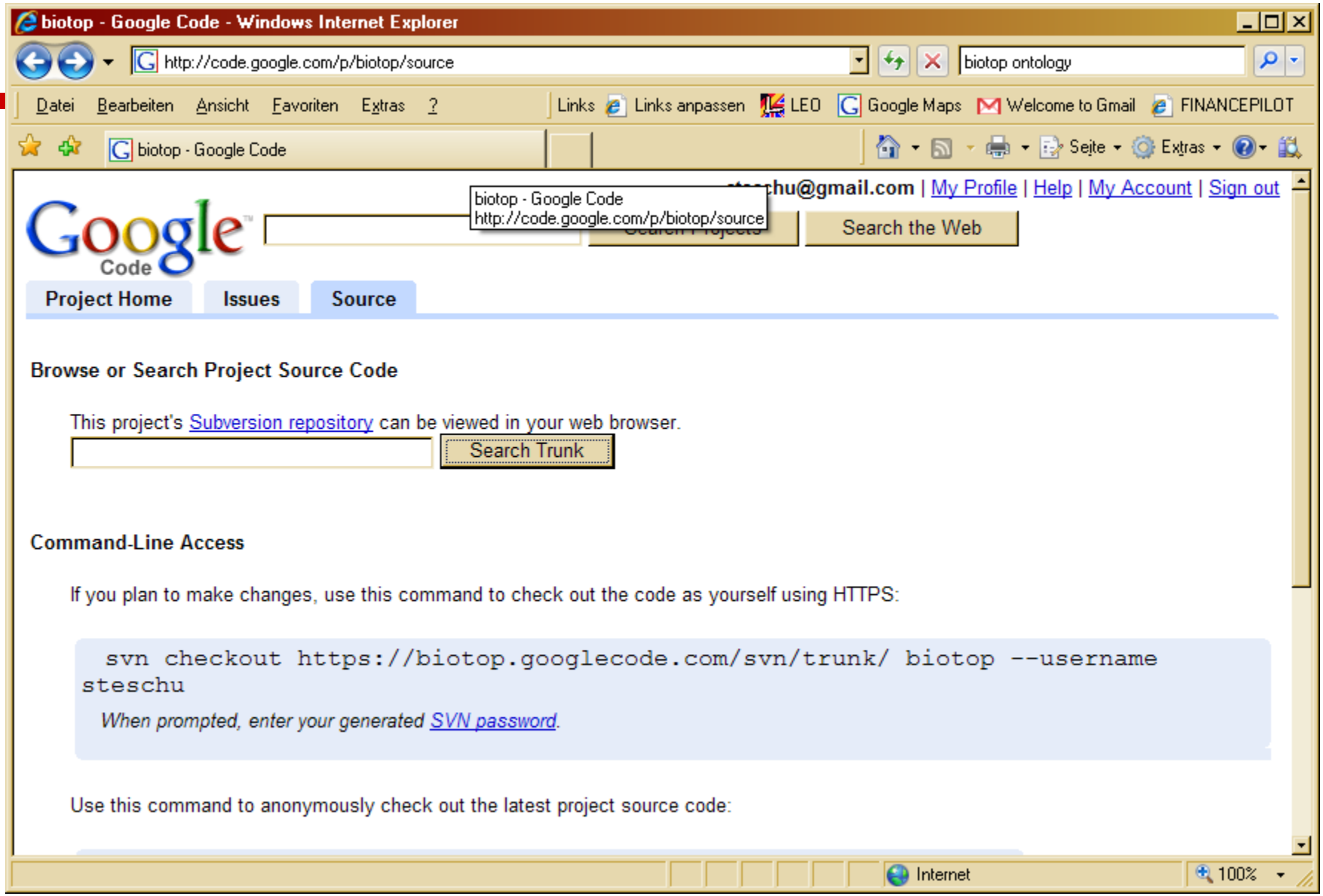
- Integration with BFO (work in process)
- Completion of textual and formal definitions
- Extension of process and function branch
- Integration of BioTop in OBO foundry
- Mapping BioTop to the UMLS Semantic Network
- Use of BioTop in text mining: Experimental validation of added value compared to GENIA / thesaurus approach
- Empirical Studies to create evidence of whether
  - Formal Ontologies better serve the needs of knowledge annotation / processing in biomedicine
  - Informal Thesauri are sufficient

BioTop Redesign is going on



BioTop Redesign is going on




biotop - Google Code - Windows Internet Explorer
http://code.google.com/p/biotop/source
biotop ontology
Datei Bearbeiten Ansicht Favoriten Extras ?
Links Links anpassen LEO Google Maps Welcome to Gmail FINANCEPILOT
biotop - Google Code
Seite Extras
biotop - Google Code
http://code.google.com/p/biotop/source
chu@gmail.com | My Profile | Help | My Account | Sign out
Google Code
Search the Web
Project Home
Issues
Source
Browse or Search Project Source Code
This project's Subversion repository can be viewed in your web browser.
Search Trunk
Command-Line Access
If you plan to make changes, use this command to check out the code as yourself using HTTPS:
svn checkout https://biotop.googlecode.com/svn/trunk/ biotop --username steschu
When prompted, enter your generated SVN password.
Use this command to anonymously check out the latest project source code:
Internet
100%

AMIA 2006: Biomedical and Health Informatics
November 11-15, 2006 Hilton Washington & Towers Washington, DC

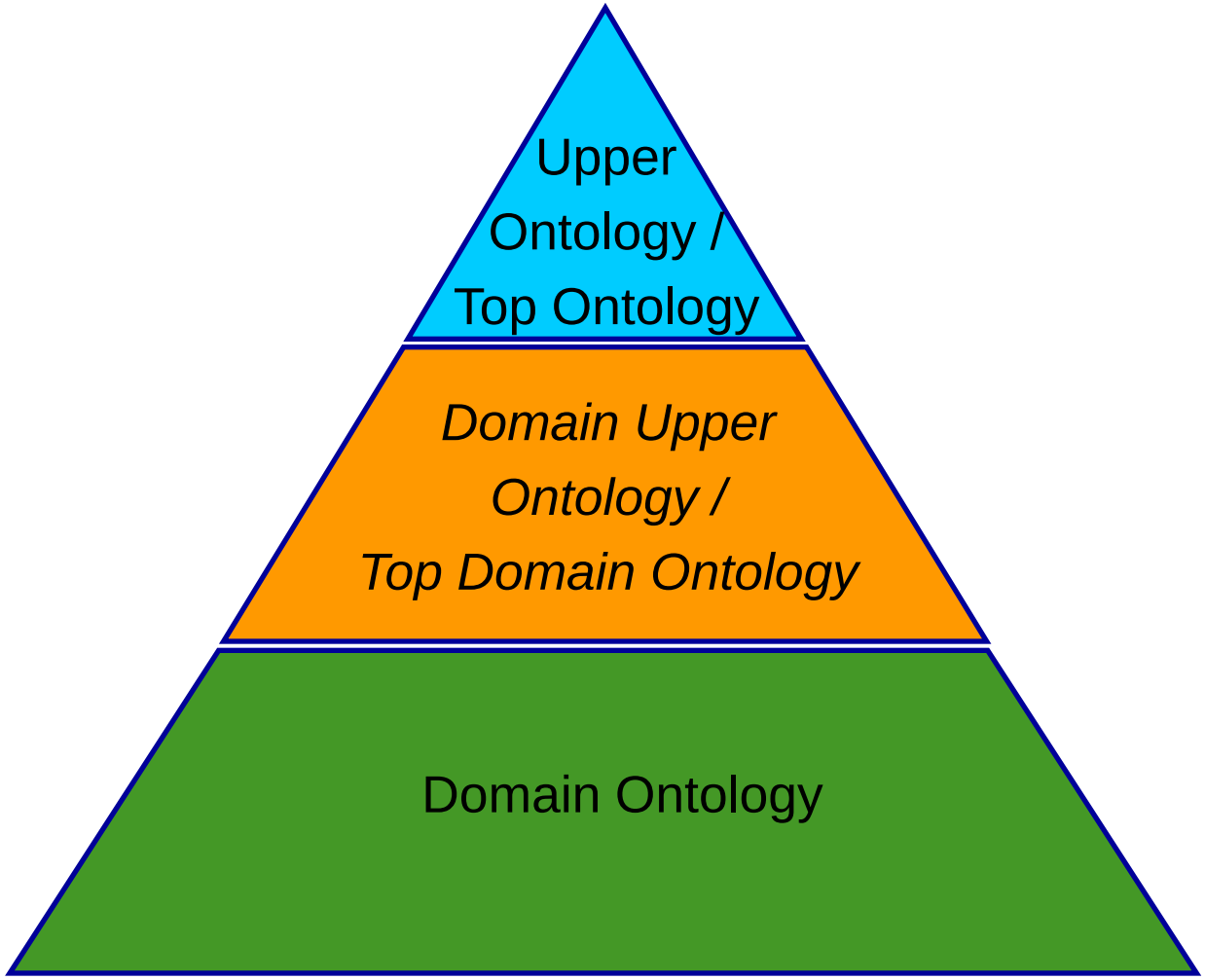
Towards a Top-Level Ontology for Molecular Biology

Stefan Schulz [1], Elena Beisswanger [2],
Udo Hahn [2], Joachim Wermter [2],

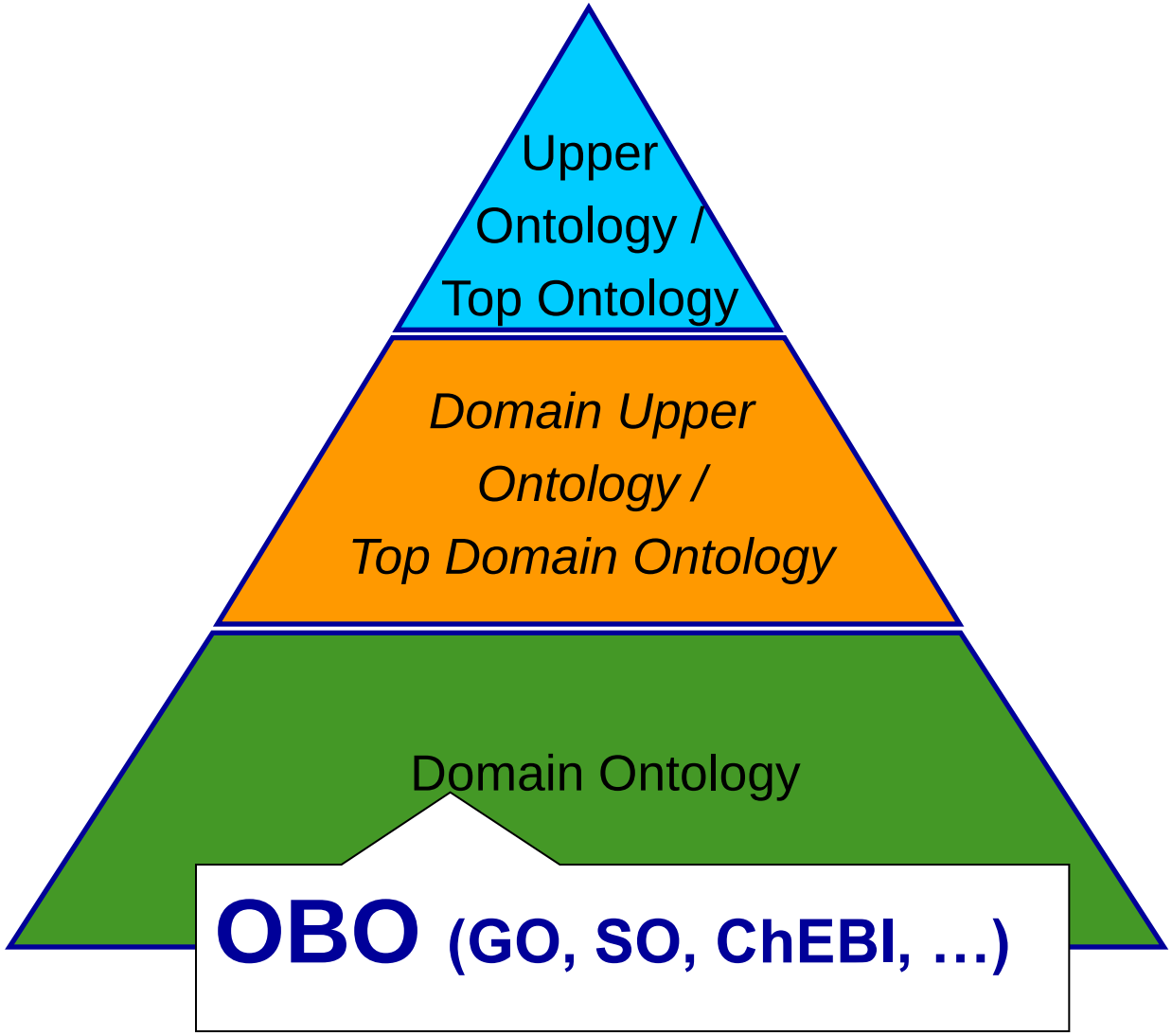
[1] Freiburg University Hospital, Department of Medical Informatics, Germany
[2] Jena University Language and Information Engineering (JULIE) Lab

Boot Strep
Information Society Technologies
StemNet
Semantic Mining
—— Semantic Interoperability and Data Mining in Biomedicine

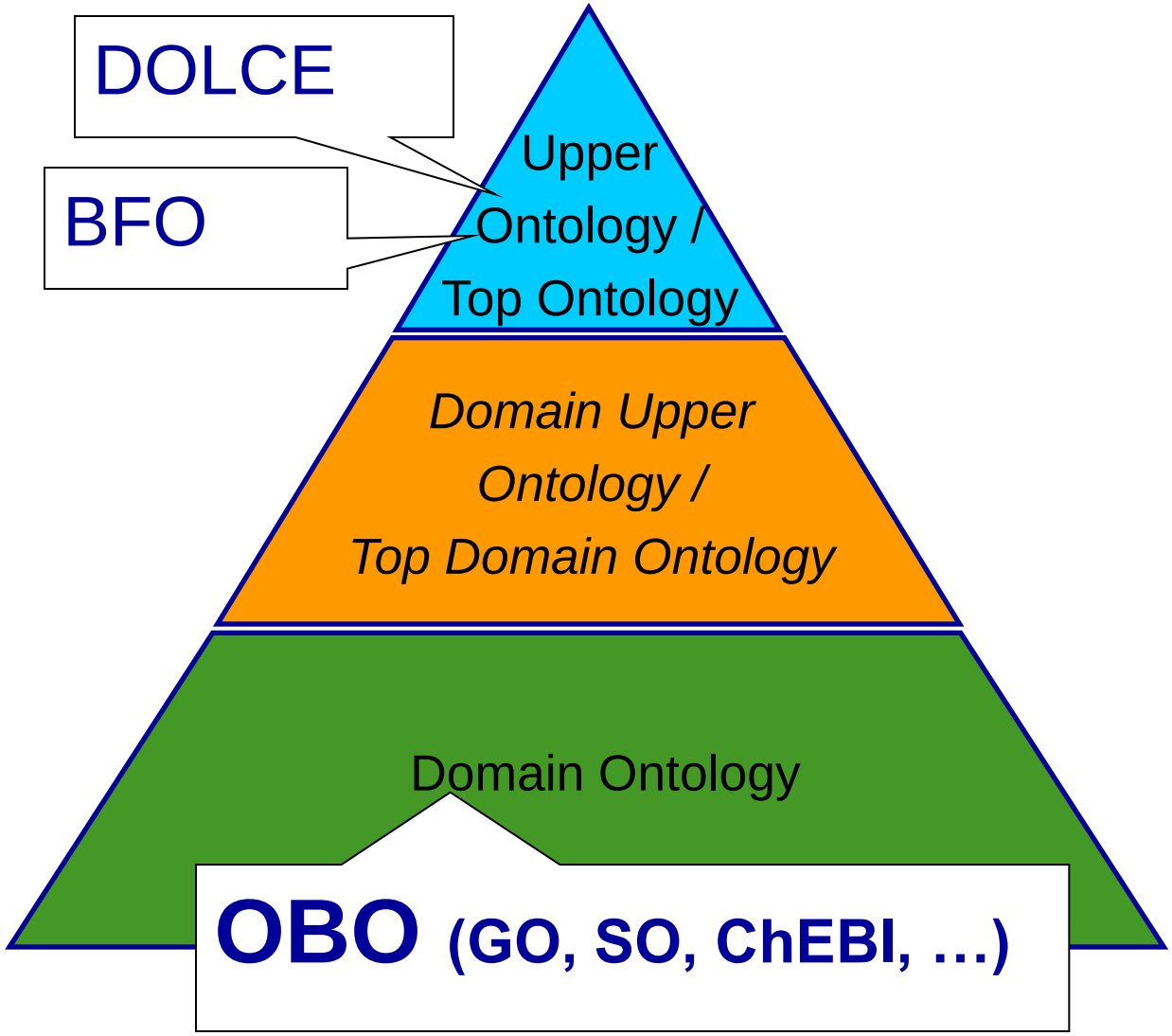
Ontology Layers



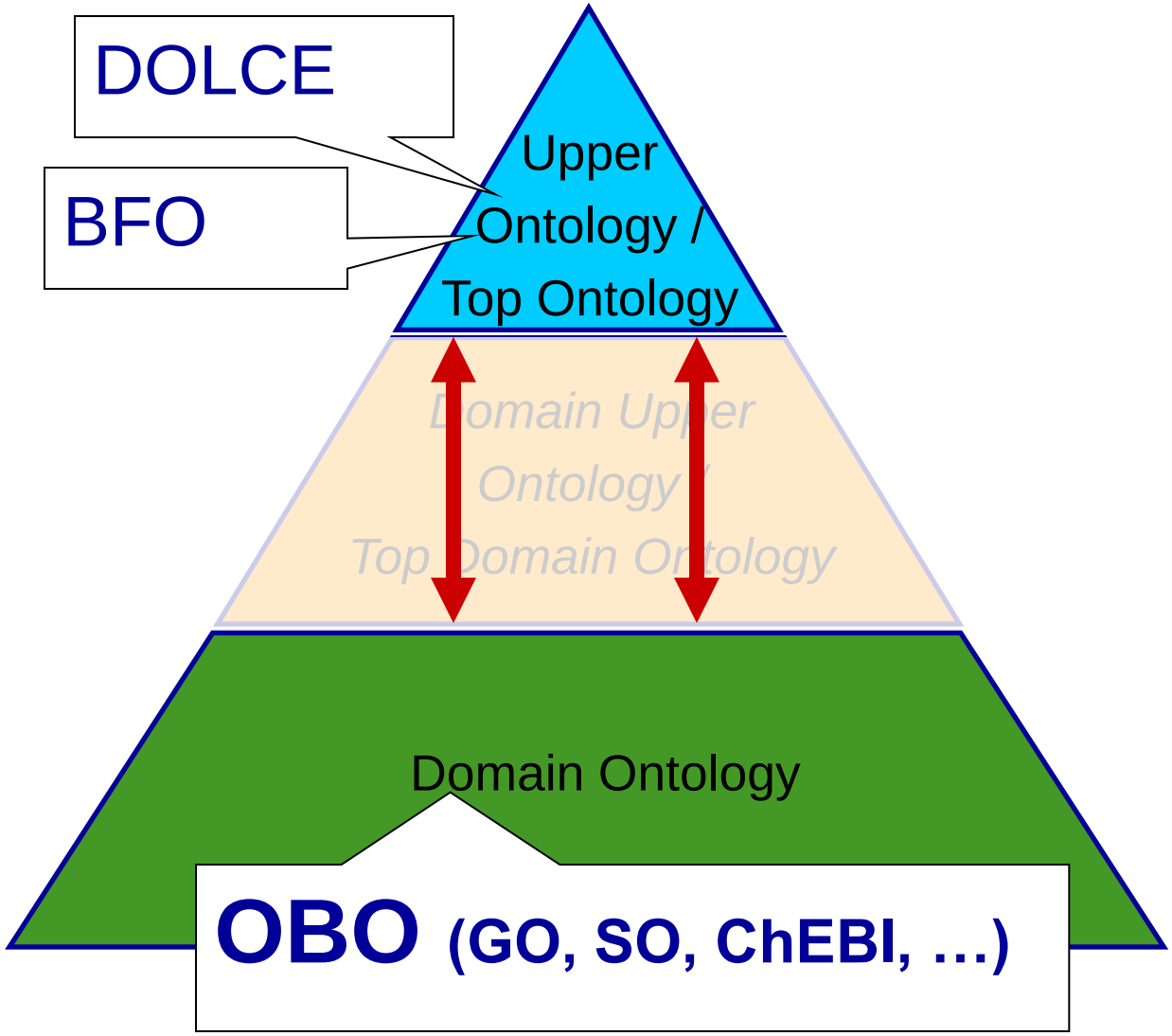
Ontology Layers



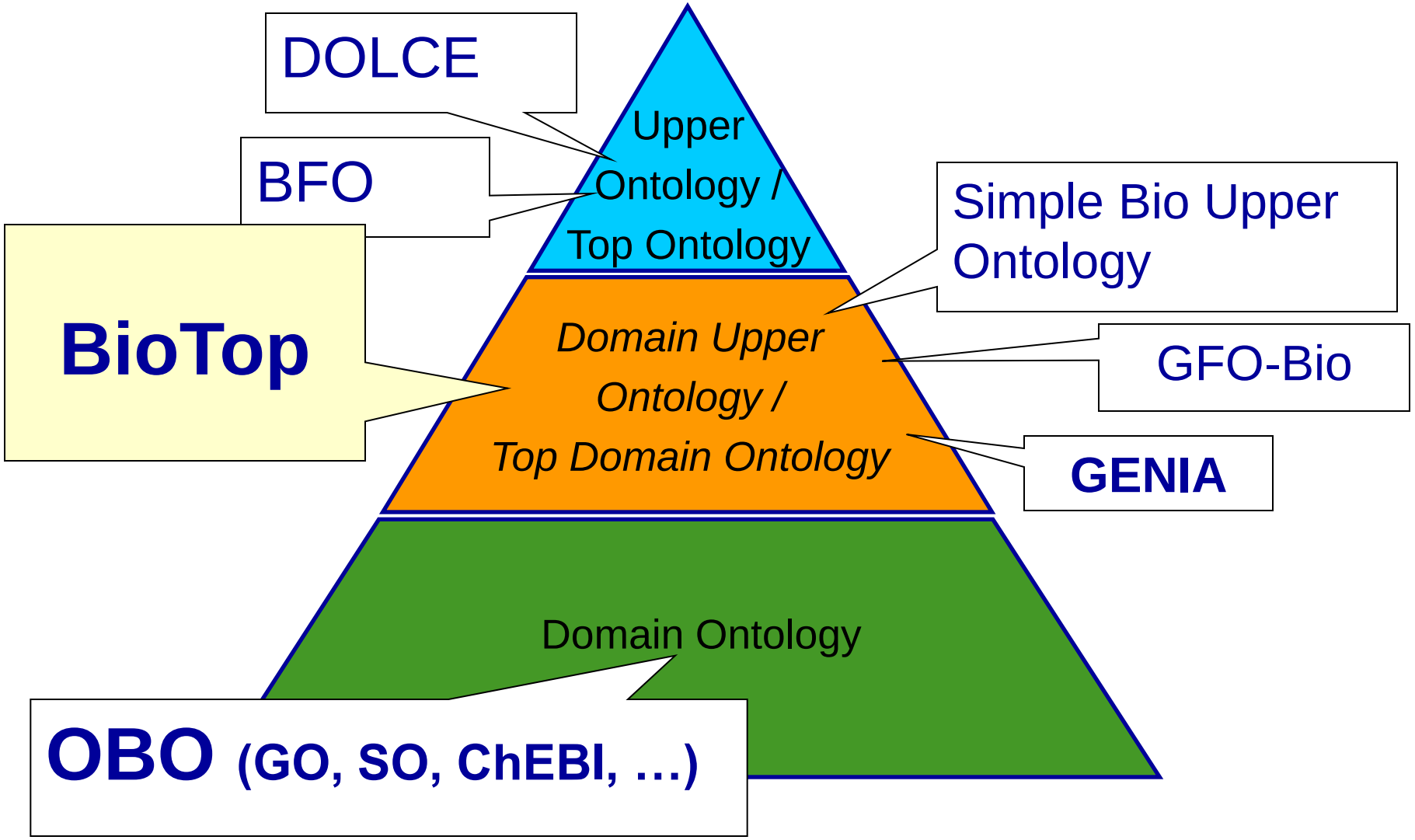
Ontology Layers



Ontology Layers

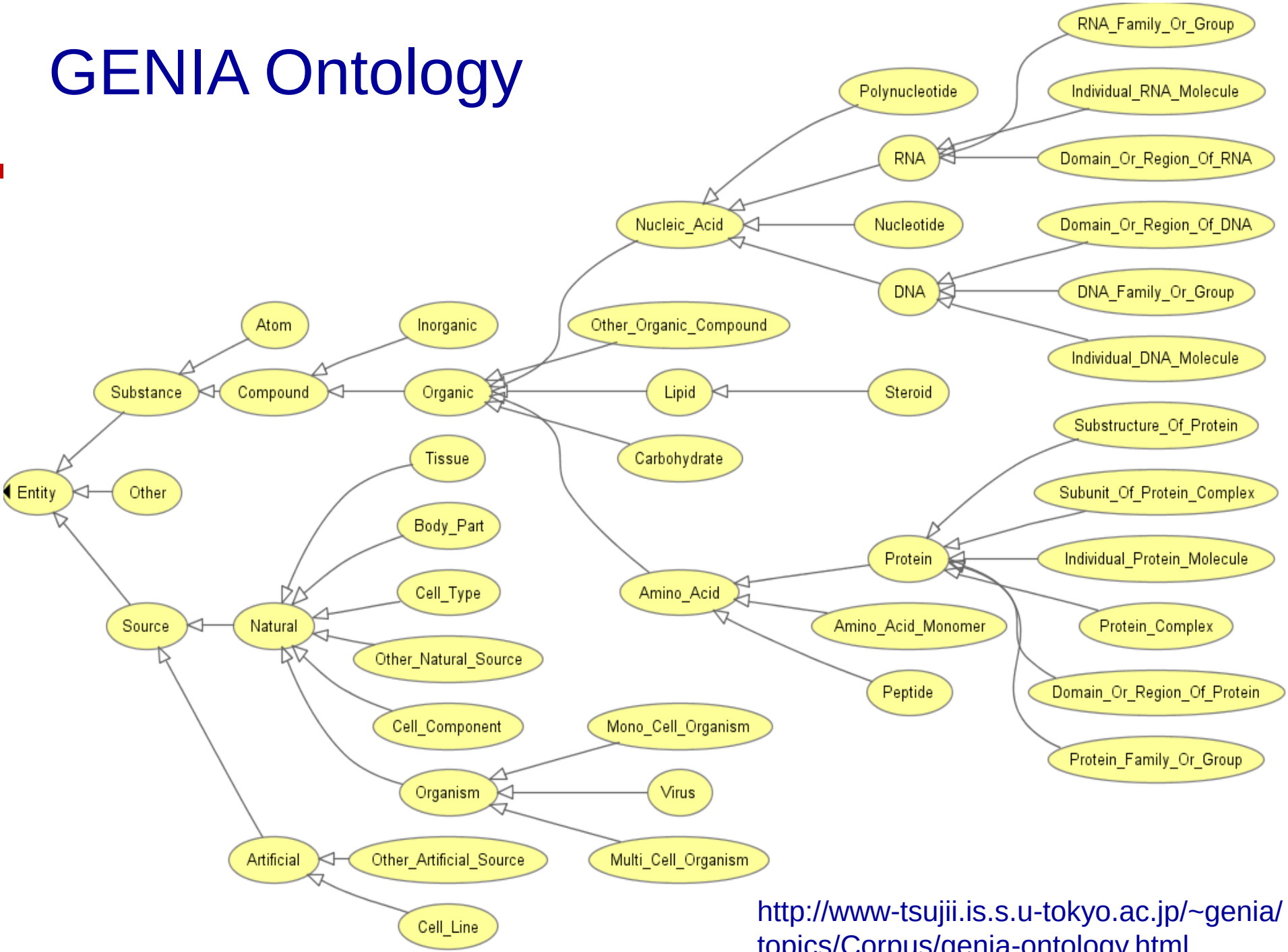


Ontology Layers



GENIA Ontology

GENIA Ontology



http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html

GENIA Ontology

"The GENIA ontology is intended to be a formal model of cell signaling reactions in human.

... use of the GENIA ontology is to provide a basis for integrated view of multiple databases"

- 45 classes
- Taxonomy (taxonomic relations not clearly defined)
- Text definitions ("scope notes")
- Developed for semantic annotation of biological papers

GENIA Critique

1. Incomplete Textual Definitions

2. Modeling Errors

GENIA Scope Notes

1. Amino Acid Monomer:

   "An amino acid monomer, e.g. tyrosine, serine, tyr, ser"

2. DNA:

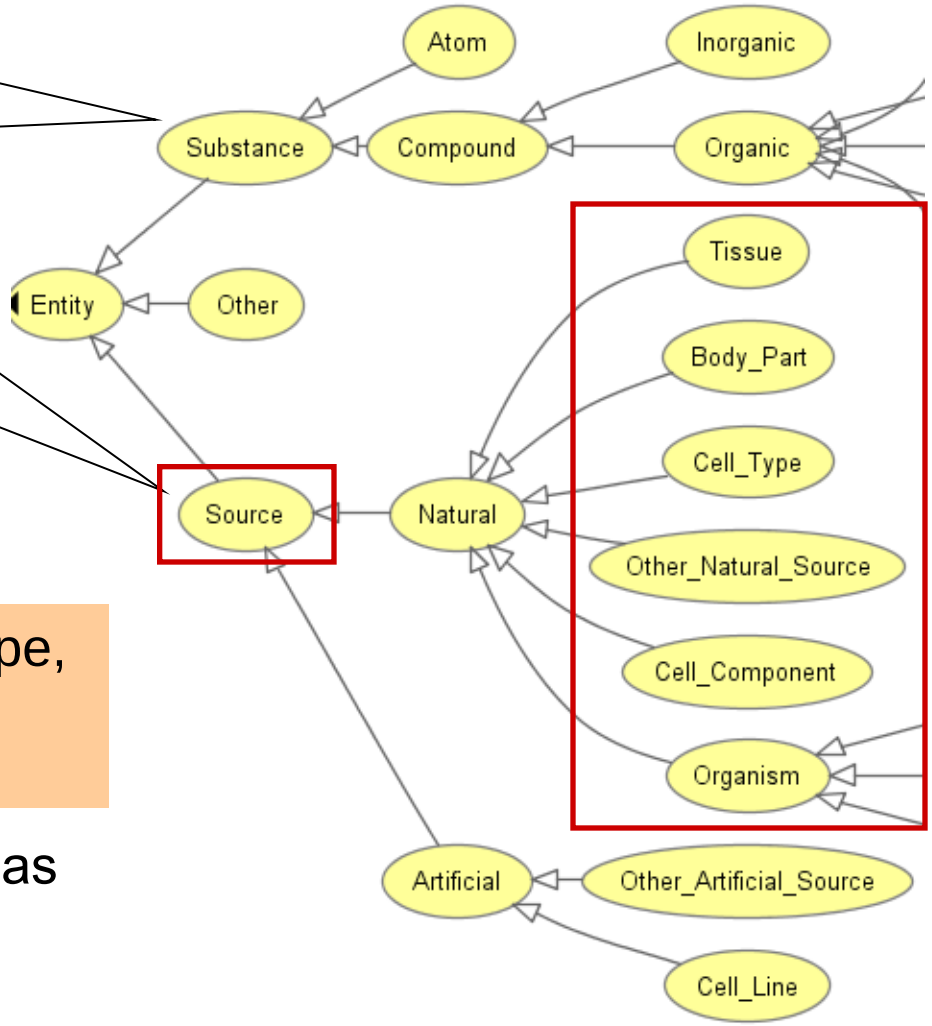
   "*DNAs include DNA groups, families, molecules, domains, and regions"*

False Taxonomic Parent

GENIA: Source

"… substances involved in biochemical reactions."

"Sources are biological locations where substances are found and their reactions take place, …"



Critique: Organism, body part, cell type, are not primarily 'sources'…

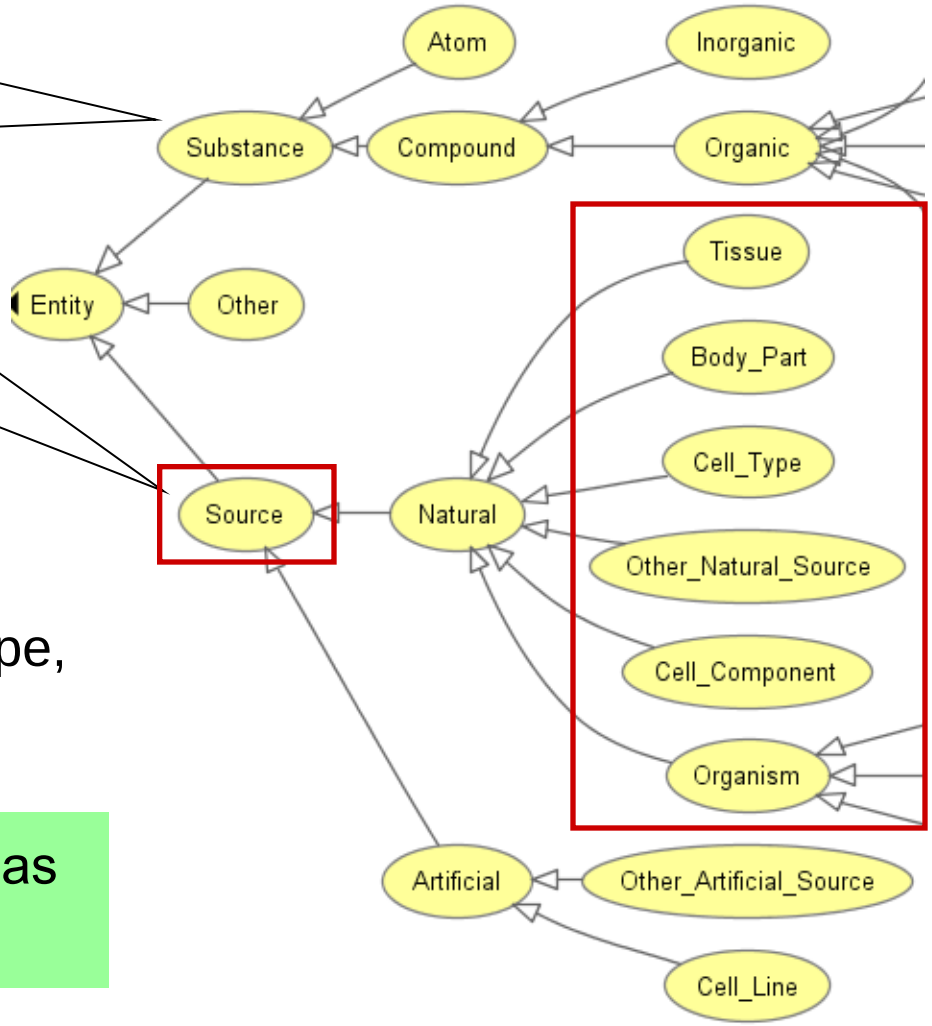
Suggestion: 'Source' as role, 'object' as superclass instead

False Ontological Parent

GENIA: Source

"… substances involved in biochemical reactions."

"Sources are biological locations where substances are found and their reactions take place, …"

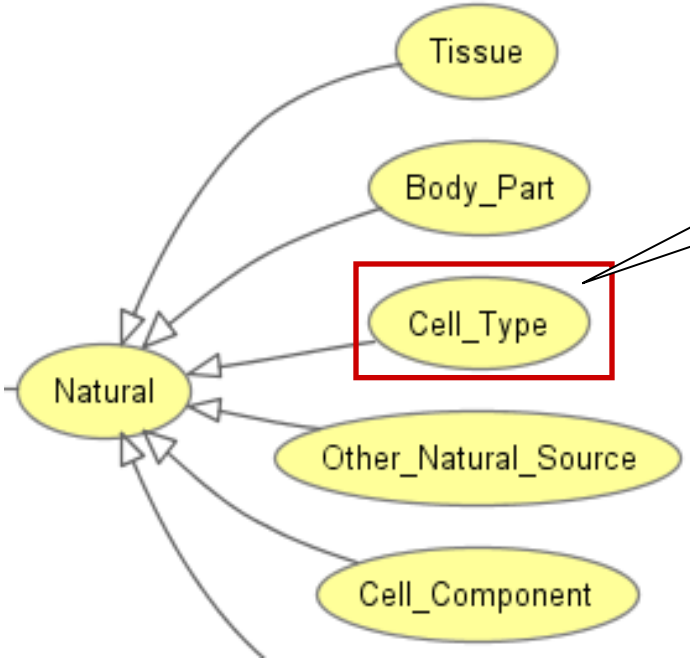


Critique: Organism, body part, cell type, are not primarily 'sources'…

Suggestion: 'Source' as role, 'object' as superclass instead

Misconception of *Type*

GENIA: Cell Type



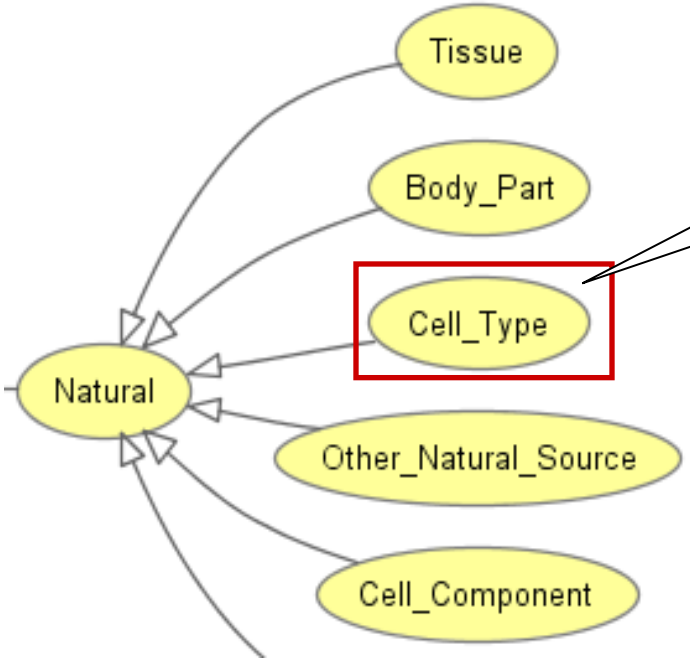
"A cell type, e.g. T-Lymphocyte, T-cell, astrocyte, fibroblast [sic]"

<u>Critique:</u> What is meant by 'Cell *<u>Type</u>*' ?

1. The universal cell (instantiated by e.g. a particular T cell)
2. A meta level category (instantiated by e.g. the universal T cell)

Misconception of *Type*

GENIA: Cell Type



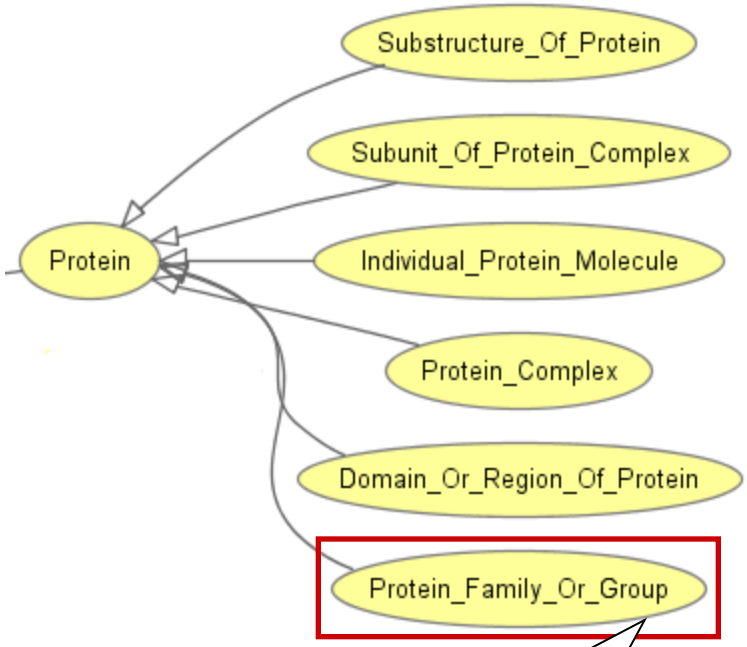
"A cell type, e.g. T-Lymphocyte, T-cell, astrocyte, fibroblast [sic]"

Critique: What is meant by 'Cell *Type*' ?

1. The universal cell (instantiated by e.g. a particular T cell)
2. A meta level category (instantiated by e.g. the universal T cell)

Misclassification

GENIA: Protein Family or Group



"A protein family or group, e.g. STATs"

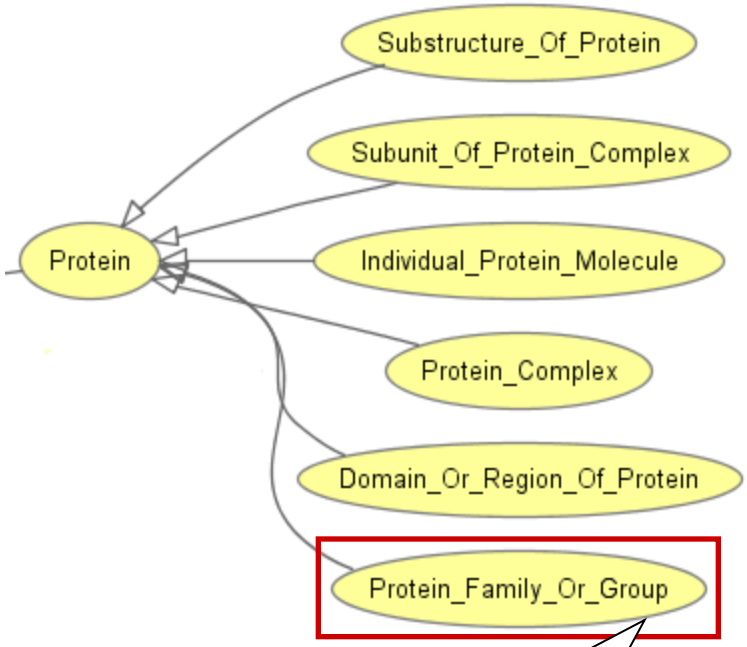
Critique:

1. A protein family is not a 'protein'
2. Refers to a human-made classification grouping proteins of the same function / location / structure.

Suggestion: Introduce classes 'Protein Function', 'Protein Role' and subclasses to classify proteins

Misclassification

GENIA: Protein Family or Group



"A protein family or group, e.g. STATs"

Critique:

1. A protein family is not a 'protein'
2. Refers to a human-made classification grouping proteins of the same function / location / structure.

Suggestion: Introduce classes 'Protein Function', 'Protein Role' and subclasses to classify proteins

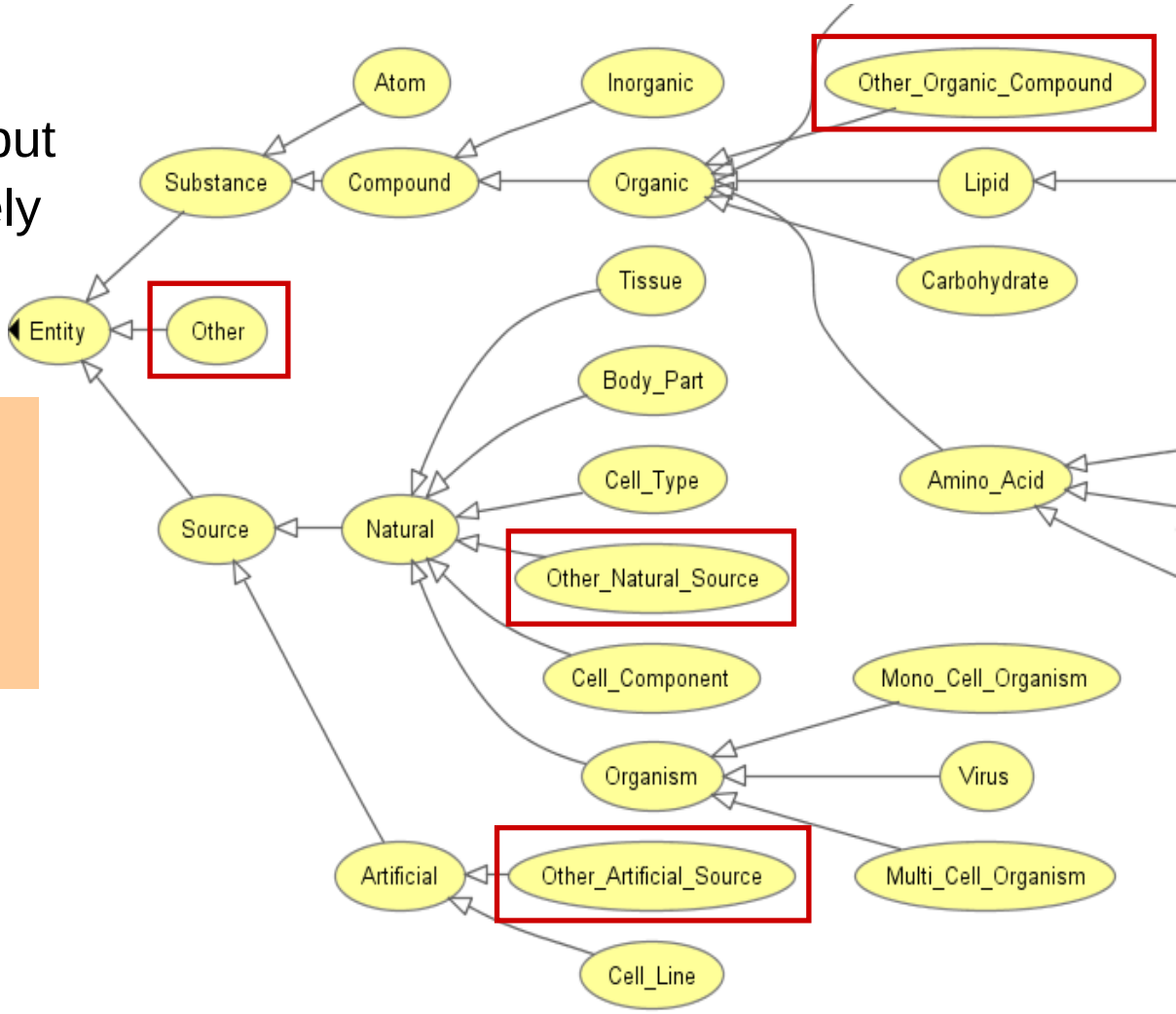
Inconsistent Use of Residual Classes

Ontologically irrelevant, but necessary for exhaustively partitioning a domain

Critique:

- Inconsequent use
- Missing definitions

Suggestion: Consequent use of defined residual classes



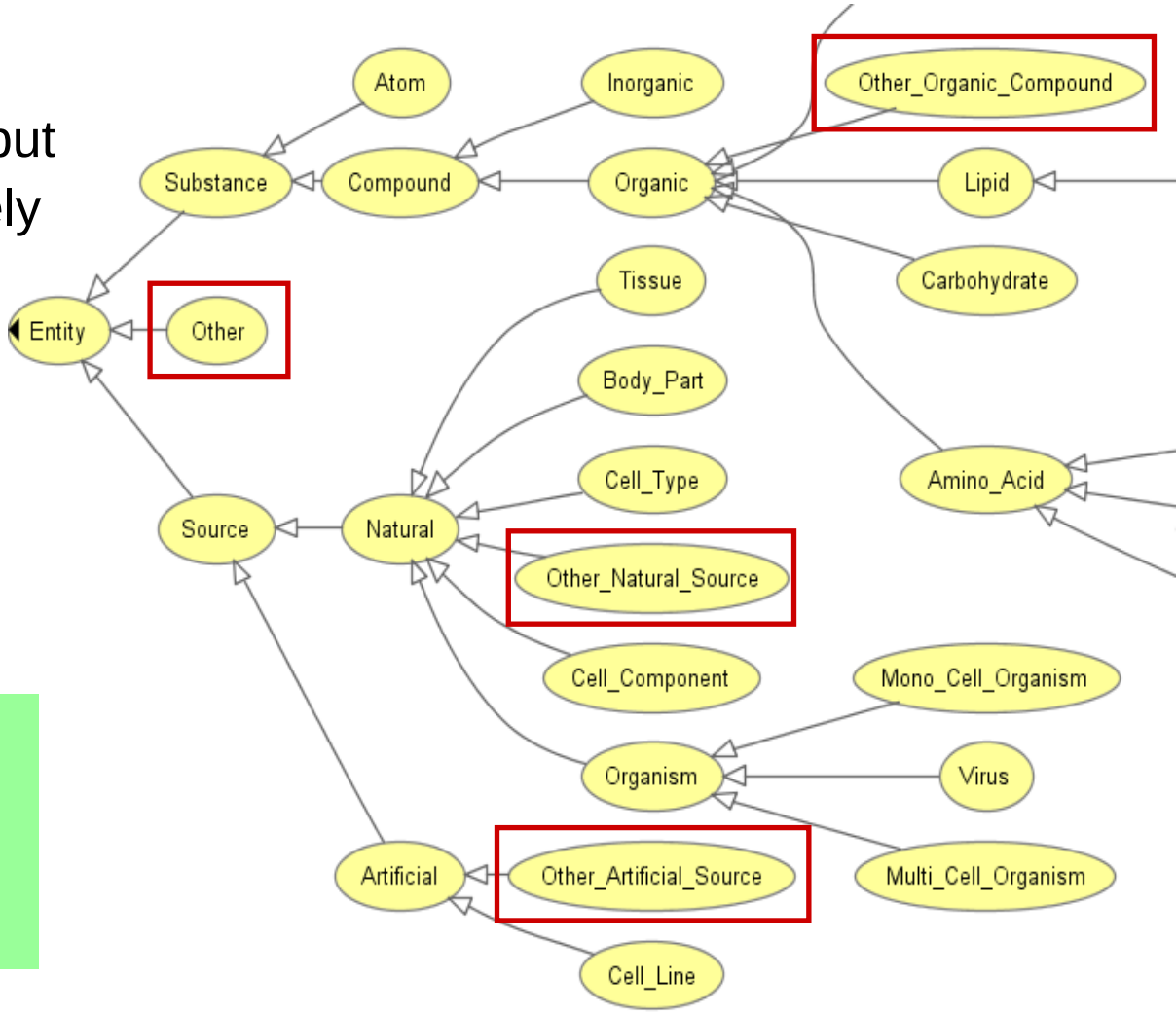
Inconsistent Use of Residual Classes

Ontologically irrelevant, but necessary for exhaustively partitioning a domain

Critique:

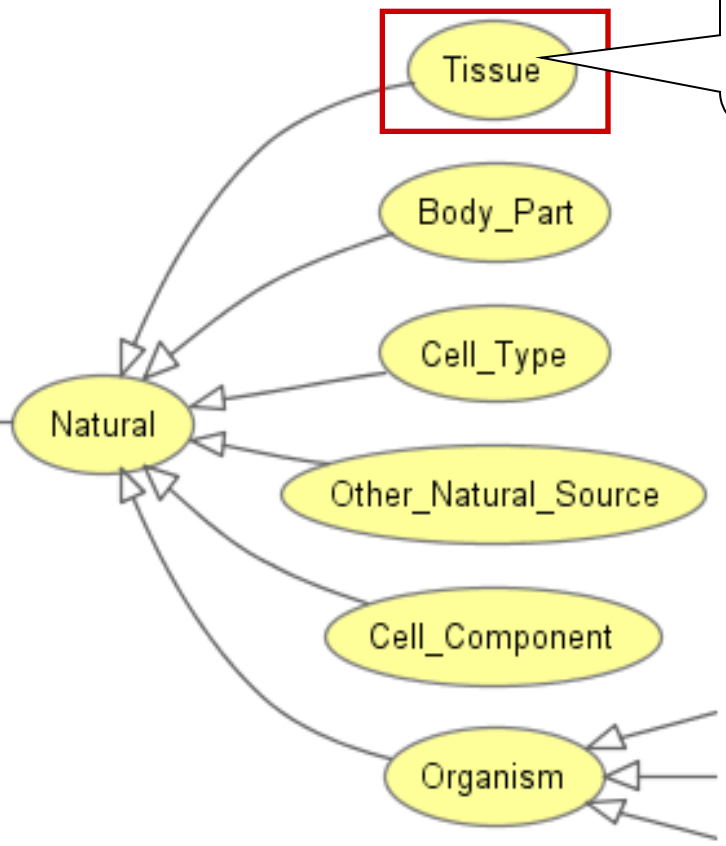
- Inconsequent use
- Missing definitions

Suggestion: Consequent use of defined residual classes



Collectives vs. their Elements

GENIA: Tissue



"Tissue, e.g. peripheral blood, lymphoid tissue, vascular endothelium [sic]"

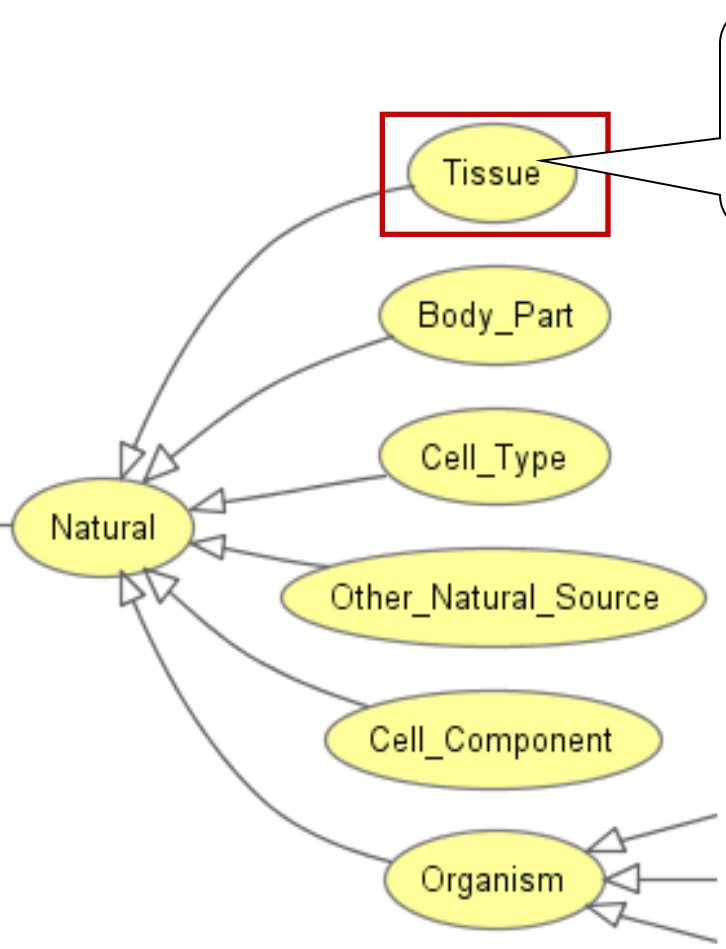
Critique: Not clear what is an instance of tissue.

1. Totality of the mass/collective, e.g. all red blood cells (RBCs)
2. Any portion of it, e.g. RBC in a lab sample
3. The minimal constituent, e.g. a single RBC

Suggestion: Strict distinction between singletons and collectives. New class 'collective' and respective sublcasses

Collectives vs. their Elements

GENIA: Tissue



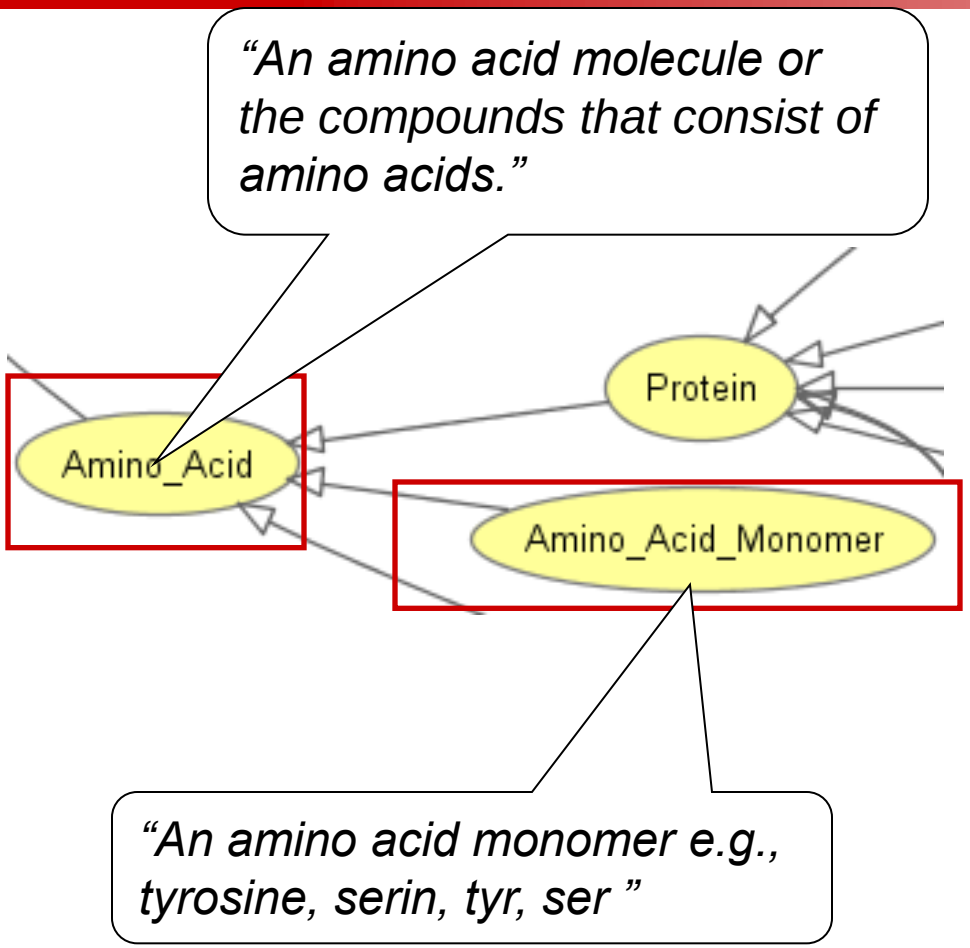
"Tissue, e.g. peripheral blood, lymphoid tissue, vascular endothelium [sic]"

Critique: Not clear what is an instance of tissue.

1. Totality of the mass/collective, e.g. all red blood cells (RBCs)
2. Any portion of it, e.g. RBC in a lab sample
3. The minimal constituent, e.g. a single RBC

Suggestion: Strict distinction between singletons and collectives. New class 'collective' and respective sublcasses

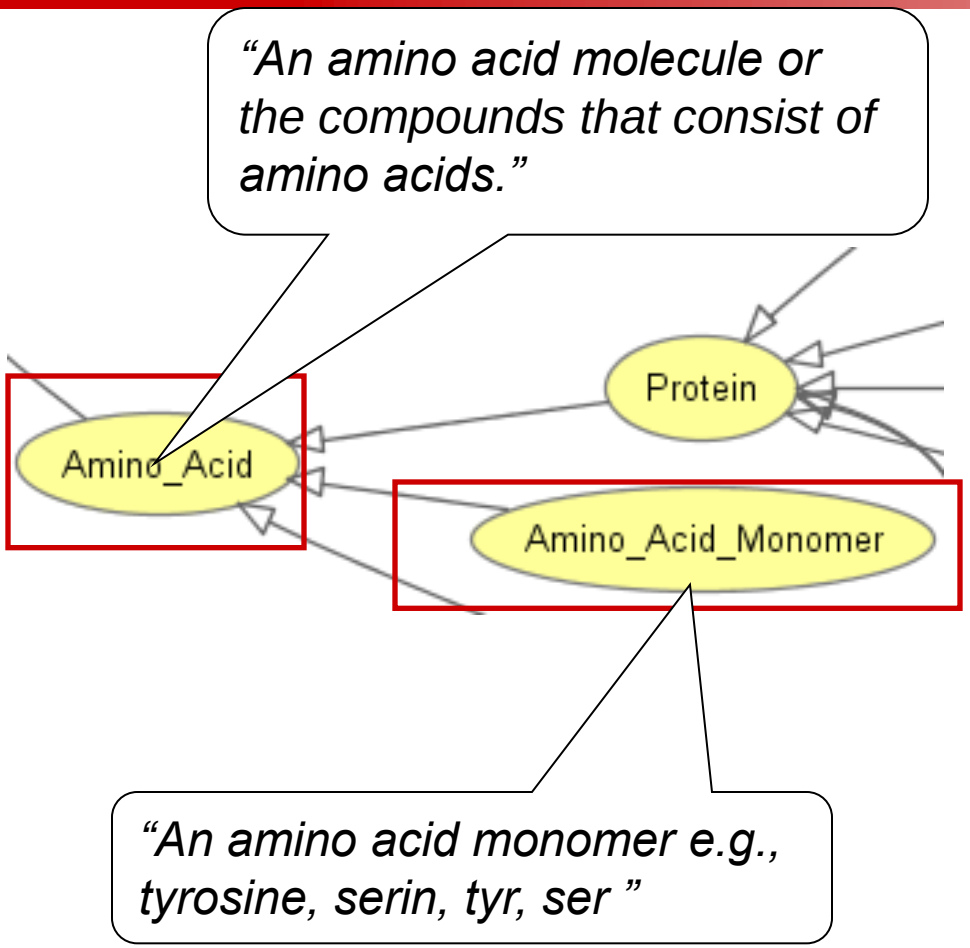
Non-Conformable Naming Policy



Critique: 'Amino Acid ' counterintuitive name (suggests single molecule, not chain)

Suggestion: Remove class 'Amino Acid', use classes 'Amino Acid Monomer' and 'Amino Acid Polymer' with proper definitions

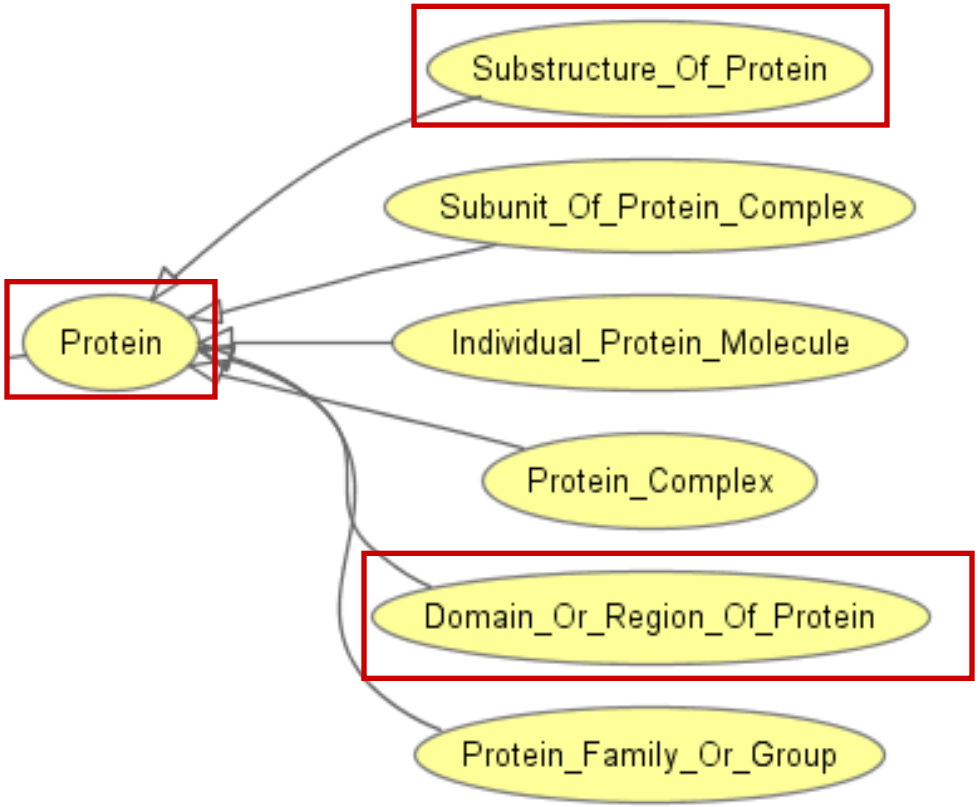
Non-Conformable Naming Policy



Critique: 'Amino Acid ' counterintuitive name (suggests single molecule, not chain)

Suggestion: Remove class 'Amino Acid', use classes 'Amino Acid Monomer' and 'Amino Acid Polymer' with proper definitions

Lack of Non-Taxonomic Relations

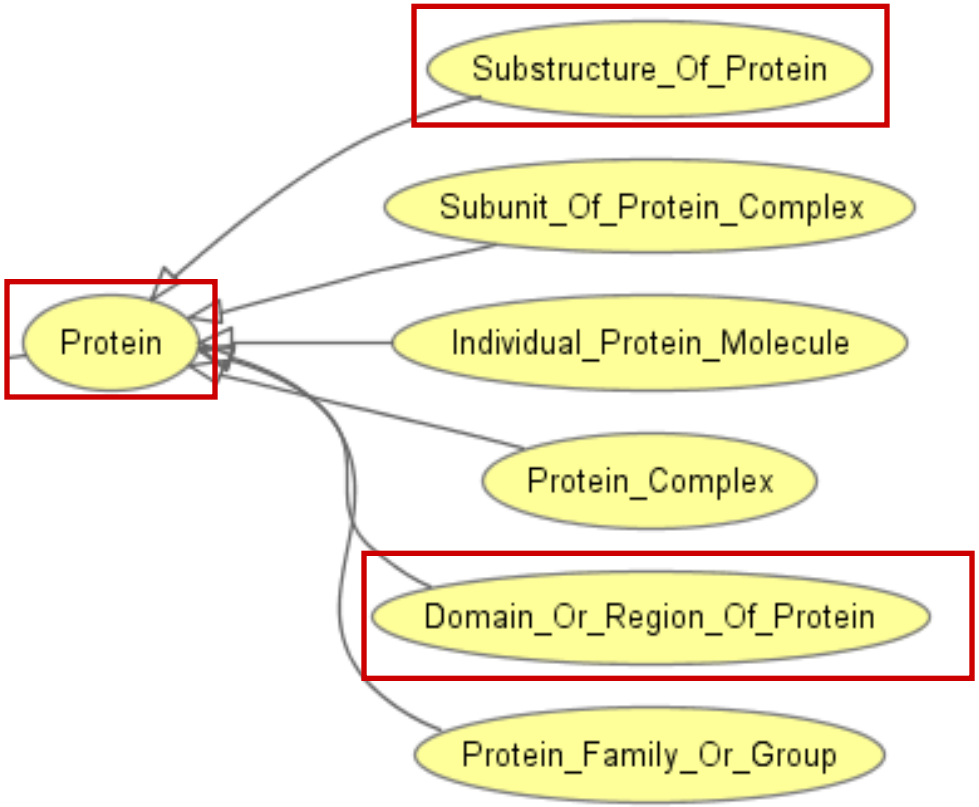


Critique: Many taxonomic links ontologically incorrect.

Example: part-of mistaken for is-a

Suggestion: Reduce is-a to its proper meaning (subclass-of) introduce part-of and other relations

Lack of Non-Taxonomic Relations



Critique: Many taxonomic links ontologically incorrect.

Example: part-of mistaken for is-a

Suggestion: Reduce is-a to its proper meaning (subclass-of) introduce part-of and other relations

From GENIA to BioTop

BioTop

- Upper level ontology for biology
  (focusing on biomedicine and molecular biology)
- Ontologically founded classes and relations
- Adding formal semantics to only verbally defined classes

- Language: OWL-DL
- Editing environment: Protégé

http://morphine.coling.uni-freiburg.de/~schulz/BioTop/BioTop.html

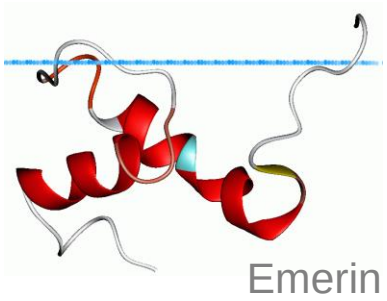
BioTop Relations

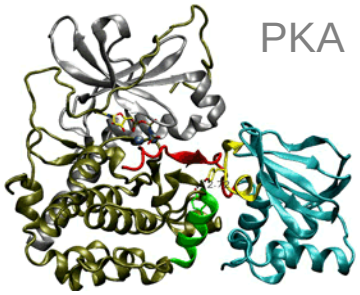
- Basis: OBO Relation Ontology
  - *properPartOf* and *hasProperPart*
  - *locatedIn* and *locationOf*
  - *derivesFrom*
  - *hasParticipant* and *participatesIn*
- Refined *partOf*:
  - *properPartOf* and *hasProperPart:* *irreflexive*
  - *componentOf* and *hasComponent:* *nontransitive*
  - *grainOf* and *hasGrain:* *nontransitive*
- Additional:
  - *hasFunction* and *functionOf*

Collectives and Relation *hasGrain*

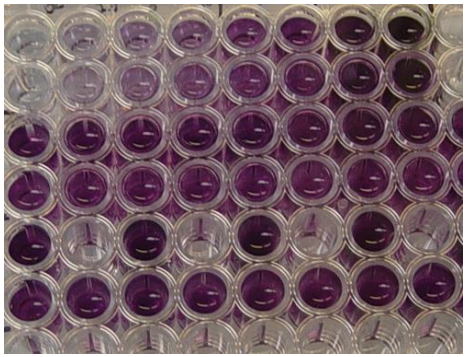
- Controversial: Referents in biological texts: Collectives (pluralities) or count objects?

  "Emerin is phosphorylated on serine 49 by protein kinase A"





← or →

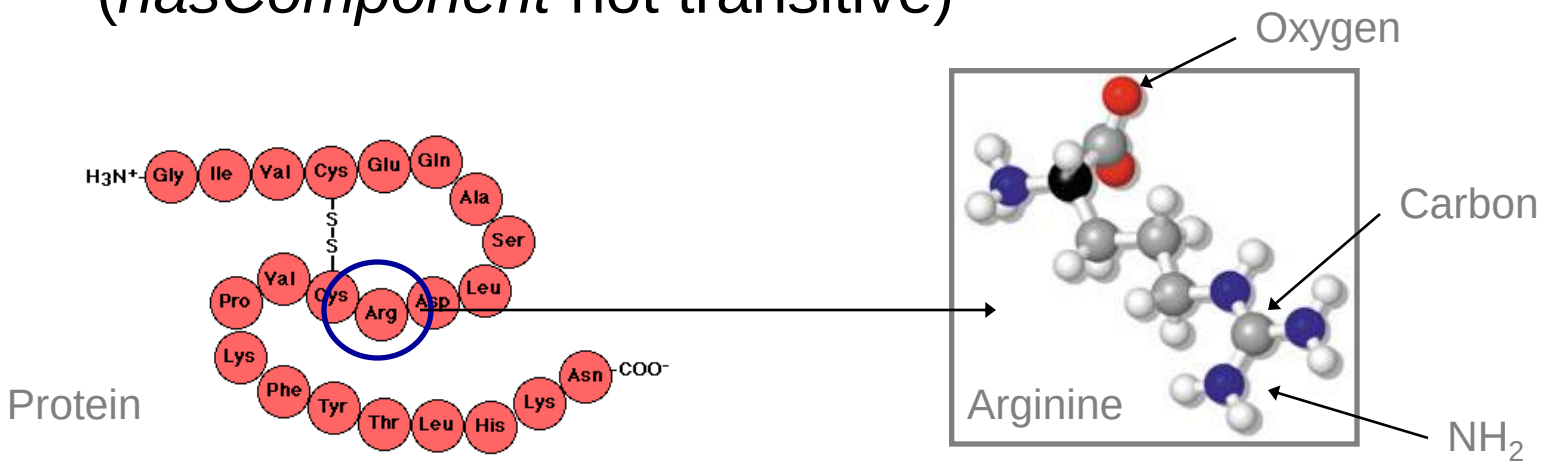


- BioTop: "Collective of x *hasGrain* x"
- *hasGrain* non-transitive subrelation of *hasPart*

cf. Schulz & Jansen, KR-MED 2006

Relation *hasComponent*

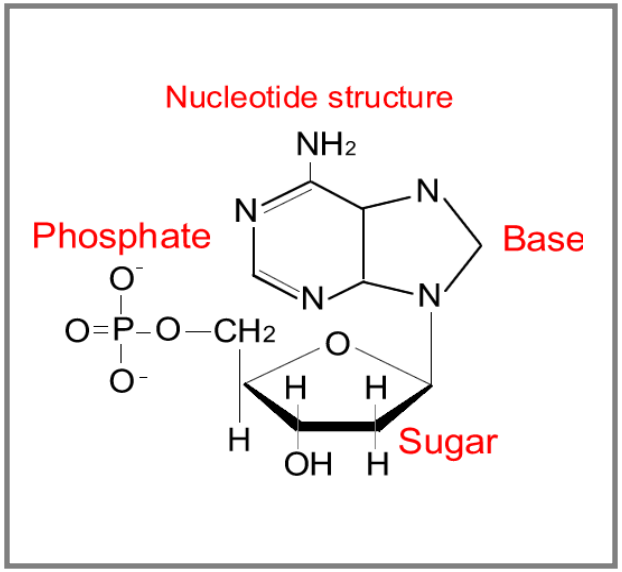
- Non-transitive subrelation of *hasPart*
- Example: Definition of Protein
  - Protein *hasPart* only amino acid is not true (*hasPart* transitive)
  - Protein *hasComponent* only amino acid is true (*hasComponent* not transitive)



Full Definitions

- Classes defined by necessary and sufficient criteria
- Rationales
  - Precise understanding of meaning
  - More rigor in automated validation process (terminological classifier)

- Introduction of new classes required
- Example Nucleotide



Rearranged Classes

- Snapshot

New Branches in BioTop

- Non-Physical Continuant
  - Biological Function
  - Biological Location

- Process
  - Biological Process

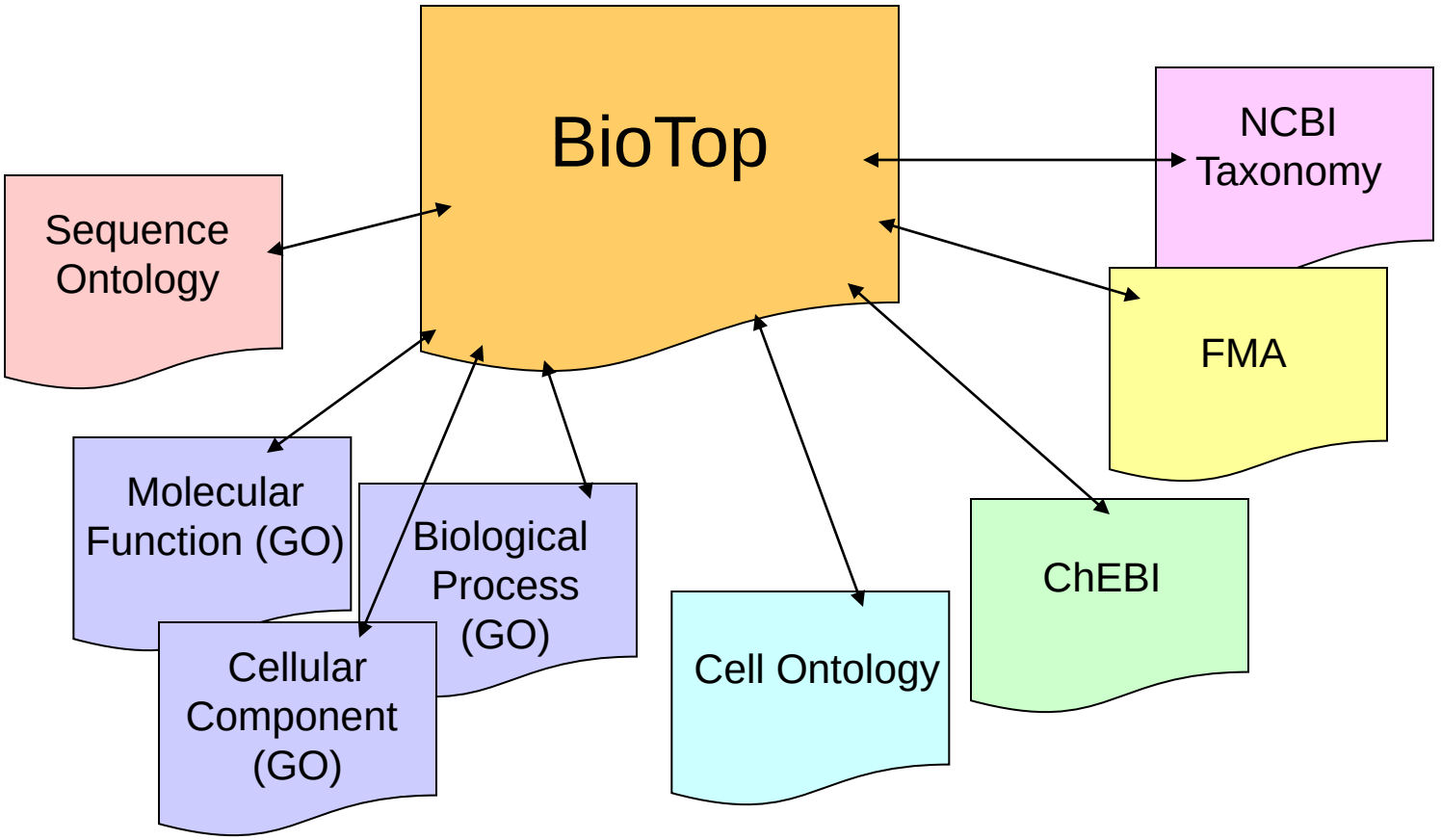
BioTop Results

- ??# Classes, ??# fully defined
- ??# Relations (evtl. Noch aufgeschluesselt)

Ontology Integration

Protein Function$_{BioTop}$	↔	Molecular Function$_{GO}$
Cell Component$_{BioTop}$	↔	Cellular Component$_{GO}$
Biological Process$_{BioTop}$	↔	Biological Process$_{GO}$
Cell$_{BioTop}$	↔	Cell$_{Cell\ Ontology}$ and Cell$_{FMA}$
Atom$_{BioTop}$	↔	Atoms$_{ChEBI}$
Organic Compound$_{BioTo}$	↔	Organic Molecular Entities$_{ChEBI}$
Tissue$_{BioTop}$	↔	Tissue$_{FMA}$
DNA$_{BioTop}$, RNA$_{BioTop}$	↔	DNA$_{SO}$, RNA$_{SO}$
Protein$_{BioTop}$	↔	Protein$_{SO}$
Organism$_{BioTop}$	↔	??$_{NCBI\ Taxonomy}$

Ontology Integration



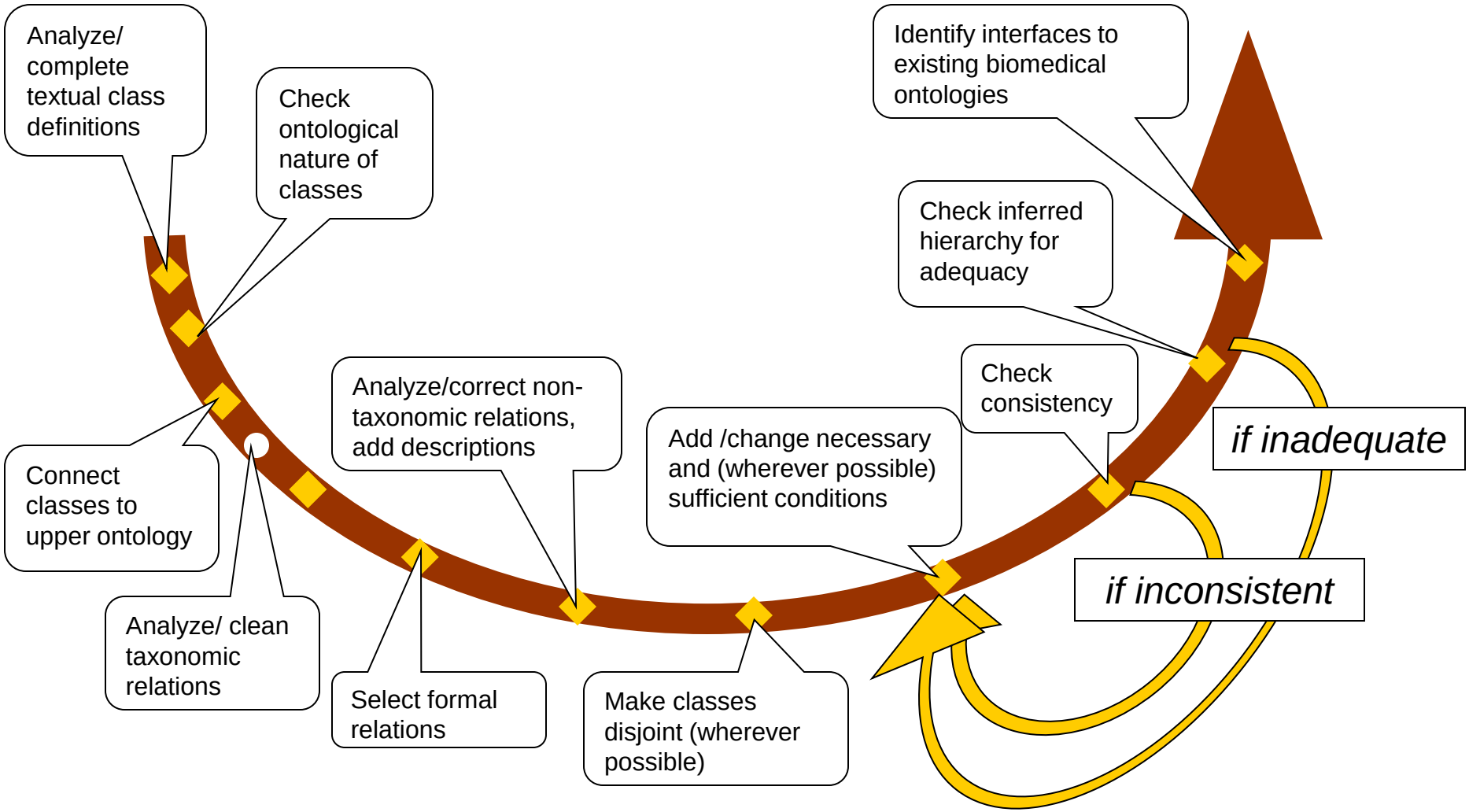
Mapping to GENIA

- **`BioTopGenia.owl`**:
  - Imports GENIA
  - Provides links from BioTop to GENIA classes

Outlook

- Integration with BFO (work in process)
- Completion of textual and formal definitions
- Extension of process and function branch
- Integration of BioTop in OBO foundry
- Use of BioTop in text mining: Experimental validation of added value compared to GENIA / thesaurus approach
- Create evidence of whether
  - Formal Ontologies better serve the needs of knowledge annotation / processing in biomedicine
  - Informal Thesauri are sufficient

Ontology Redesign is going on



Ontology Redesign is going on