Ontologische Kritik der Genia-Ontologie

Stefan Schulz, Elena Beißwanger, Anand Kumar

Genia Ontologie und Korpus

- Entwickelt am Tsuji-Lab, Tokio
- Anspruch (WWW): “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, e.g.,
  - Information retrieval (IR) & filtering (IF)
  - Information extraction (IE)
  - Document and term classification & categorization
  - Summarization, etc. “
- GENIA Corpus Ver. 3.0x: 2000 MEDLINE Abstracts. (MeSH terms: *Human*, *Blood Cells*, and *Transcription Factors)*.
- Genia-Korpus ist annotiert mit Termen der Genia-Ontologie

- +---+-source-+-natural-+-organism-+-multi-cell organism
  | | | +-mono-cell organism
  | | | +-virus
  | | +-body part
  | | +-tissue
  | | +-cell type
  | | +-cell component
  | | +-other (natural source)
  | +-artificial-+-cell line
  | +-other (artificial source)
  +-substance-+-compound-+-organic-+-amino acid-+-protein-+-protein family or group
  | | | | | +-protein complex
  | | | | | +-individual protein molecule
  | | | | | +-subunit of protein complex
  | | | | | +-substructure of protein
  | | | | | +-domain or region of protein
  | | | | +-peptide
  | | | | +-amino acid monomer
  | | | +-nucleic acid-+-DNA-+-DNA family or group
  | | | | | +-individual DNA molecule
  | | | | | +-domain or region of DNA
  | | | | +-RNA-+-RNA family or group
  | | | | | +-individual RNA molecule
  | | | | | +-domain or region of RNA
  | | | | +-polynucletotide
  | | | | +-nucleotide
  | | | +-lipid-+-steroid
  | | | +-carbohydrate
  | | | +-other (organic compounds)
  | | +-inorganic
  | +-atom
  +-other

Genia Ontologie

Genia Ontologie als Annotationsvokabular

UI - 85146267

TI - Characterization of <NE ti="3" class="protein" nm="aldosterone binding site" mt="SV" subclass="family_or_group" unsure="Class" cmt="">aldosterone binding sites</NE ti="3"> in circulating <NE ti="2" class="cell_type" nm="human mononuclear leukocyte" mt="SV" unsure="OK" cmt="">human mononuclear leukocytes</NE ti="2">.

AB - <NE ti="4" class="protein" nm="Aldosterone binding sites" mt="SV" subclass="family_or_group" unsure="Class" cmt="">Aldosterone binding sites</NE ti="4"> in <NE ti="1" class="cell_type" nm="human mononuclear leukocyte" mt="SV" unsure="OK" cmt="">human mononuclear leukocytes</NE ti="1"> were characterized after separation of cells from blood by a Percoll gradient. After washing and resuspension in <NE ti="5" class="other_organic_compounds" nm="RPMI-1640 medium" mt="SV" unsure="OK" cmt="">RPMI-1640 medium</NE ti="5">, cells were incubated at 37 degrees C for 1 h with different concentrations of <NE ti="6" class="other_organic_compounds" nm="[3H]aldosterone" mt="SV" unsure="OK" cmt="">[3H]aldosterone</NE ti="6"> plus a 100-fold concentration of <NE ti="7" class="other_organic_compounds" nm="RU-26988" mt="SV" unsure="OK" cmt="">RU-26988 </NE ti="7">(<NE ti=“17" class="other_organic_compounds" nm="11 alpha, 17 alpha-dihydroxy-17 beta-propynylandrost-1,4,6-trien-3-one" mt="SV" unsure="OK" cmt="">11 alpha, 17 alpha-dihydroxy-17 beta-propynylandrost-1,4,6-trien-3-one</NE ti=“17">), with or without an excess of unlabeled <NE ti="8" class="other_organic_compounds" nm="aldosterone" mt="SV" unsure="OK" cmt="">aldosterone</NE ti="8">. <NE ti="9" class="other_organic_compounds" nm="Aldosterone" mt="SV" unsure="OK" cmt="">Aldosterone</NE ti="9"> binds to a single class of <NE ti="10" class="protein" nm="receptor" mt="SV" subclass="family_or_group" unsure="OK" cmt="">receptors</NE ti="10"> with an affinity of 2.7 +/- 0.5 nM (means +/- SD, n = 14) and a capacity of 290 +/- 108 sites/cell (n = 14). The specificity data show a hierarchy of affinity of <NE ti="11" class="other_organic_compounds" nm="desoxycorticosterone" mt="SV" unsure="OK" cmt="">desoxycorticosterone</NE ti="11"> = <NE ti="12"

Unser Verständnis einer formalen Ontologie

- Klare Festlegung des Diskursbereichs, im Fall von Genia: konkrete physikalische Entitäten aus der Molekularbiologie (z.B. Nukleotide, Zellen, Gewebe)
- Eindeutige Charakterisierung der ontologischen Natur der Entitäten (Klassen, Konzepte, Individuen)
- Eindeutige Semantik von Relationen, Operatoren und Quantoren
- Anbindung an domänenunabhängige “Upper Ontology” wünschenswert
- Soweit möglich, Angabe von hinreichenden und notwendigen Bedingungen, also vollständige Definitionen (Aristoteles: genus + differentia)

Taxonomie als Rückgrat formaler Ontologien

- Taxonomischer Link “Is-A” (ist ein)
  - Leber Is-A Organ: für alle Instanzen von Klasse/Konzept/Typ Leber gilt, dass sie auch Instanzen von Klasse/Konzept/Typ Organ sind
  - Normalerweise mengentheoretische Deutung, daher klare Semantik
  - Klassen werden verstanden als Mengen, die über die Zeit persistieren und dabei Elemente gewinnen und verlieren können.

Wofür stehen die “Knoten” einer Ontologie ??

names

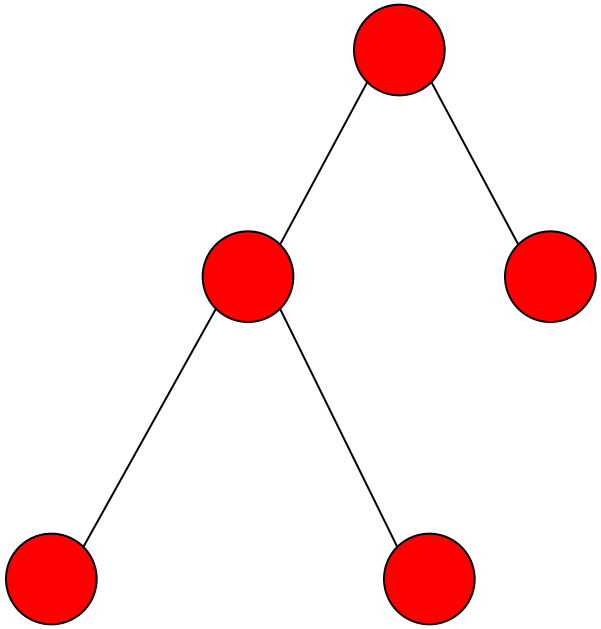
types

sets

universals

sorts

categories



descriptors

entities

synsets

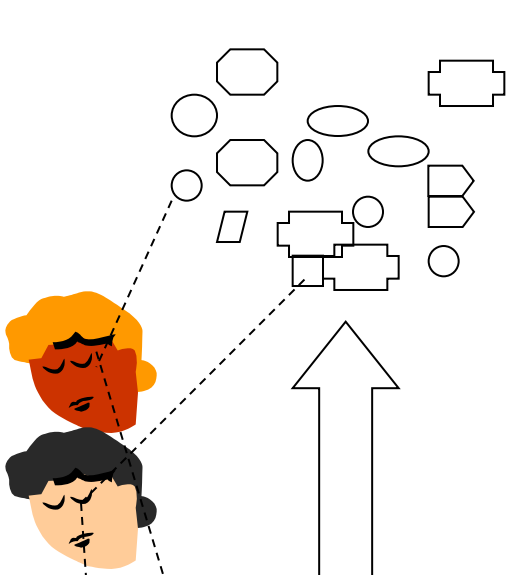
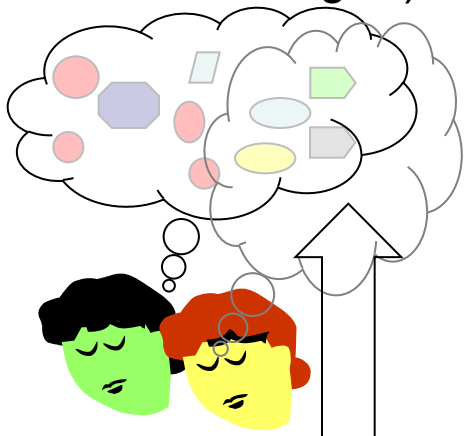
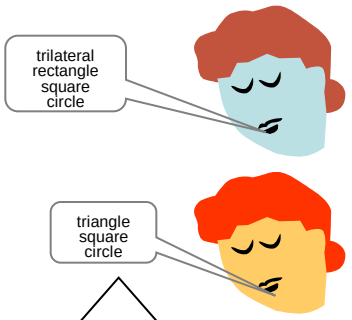
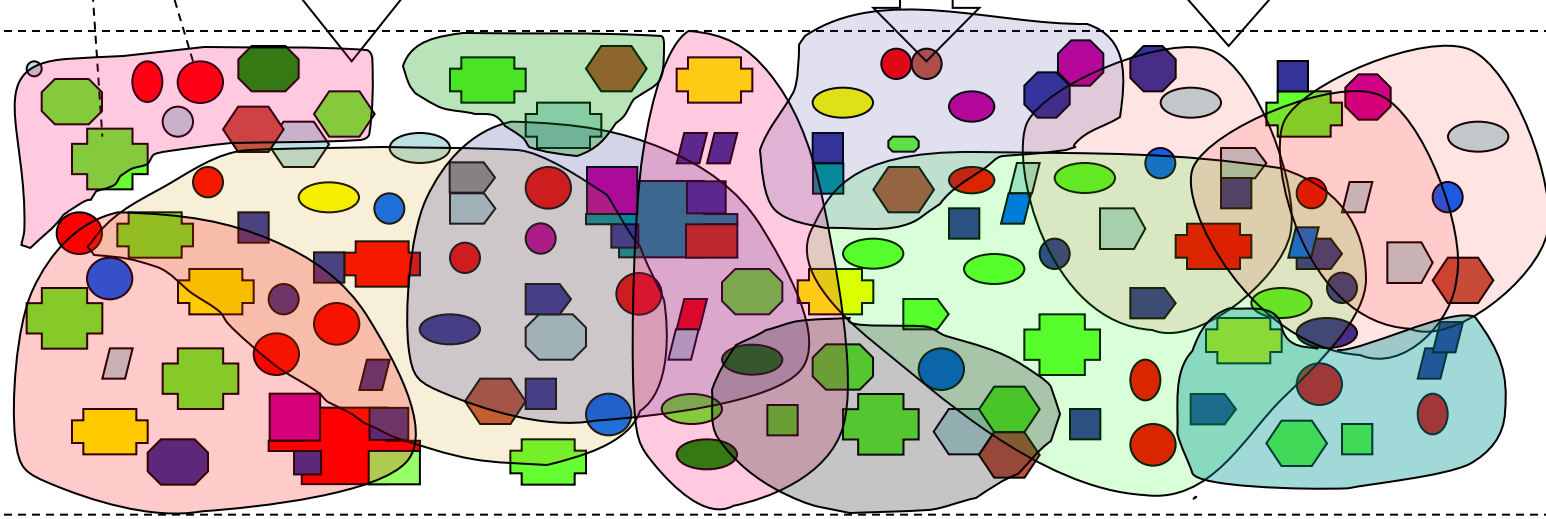
classes

terms

properties

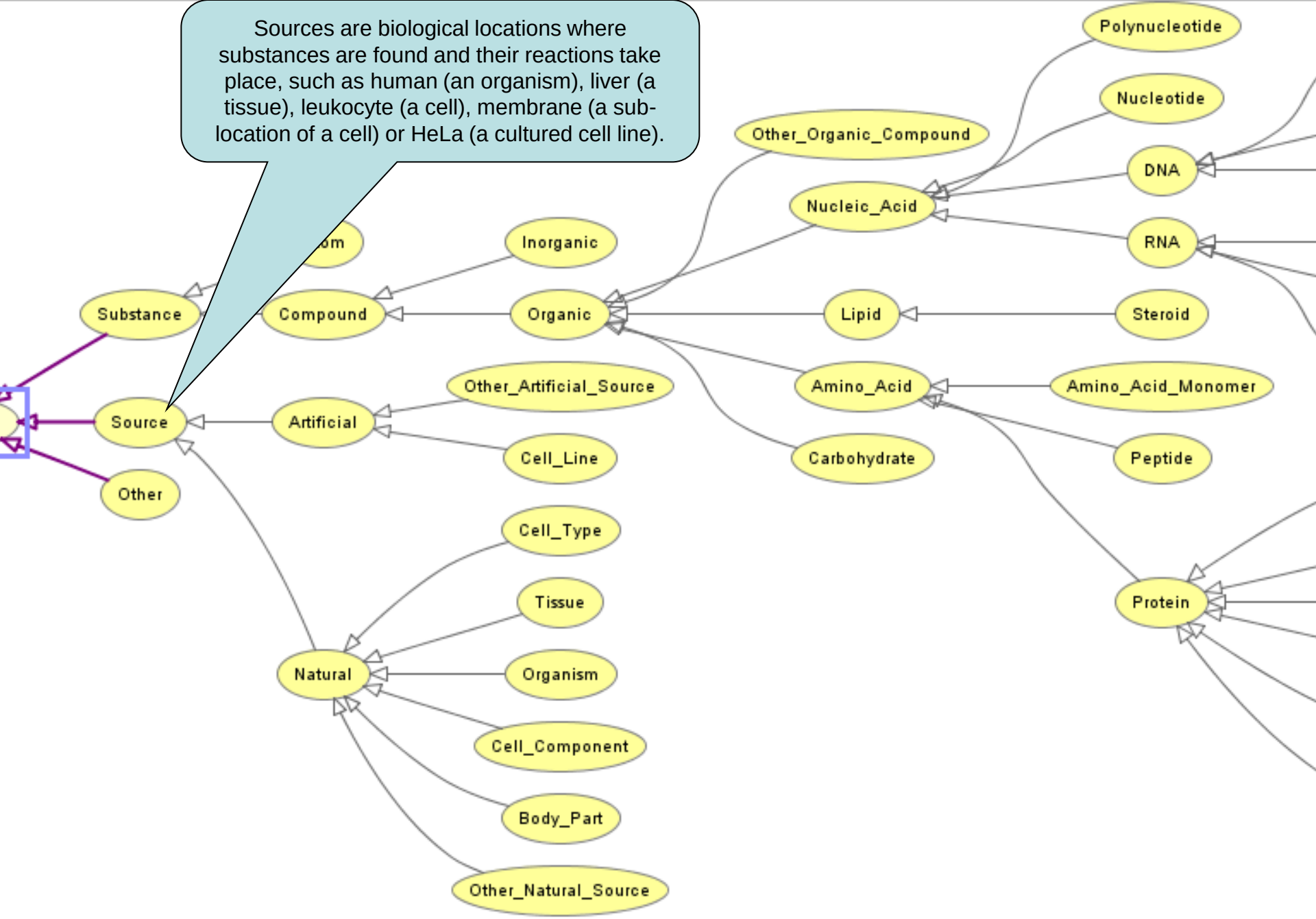
concepts

descriptors

Phil. Tradition	Realisten	Konzeptualisten	Nominalisten
Kategorien	Universalien ,Typen	Konzepte („entities of thought“)	Namen, (logische) Prädikate
Bezug			
Klassen von Dingen der Realität (die nicht von unserer Kognition abhängen)			
	Individuen, Instanzen, Partikularien		

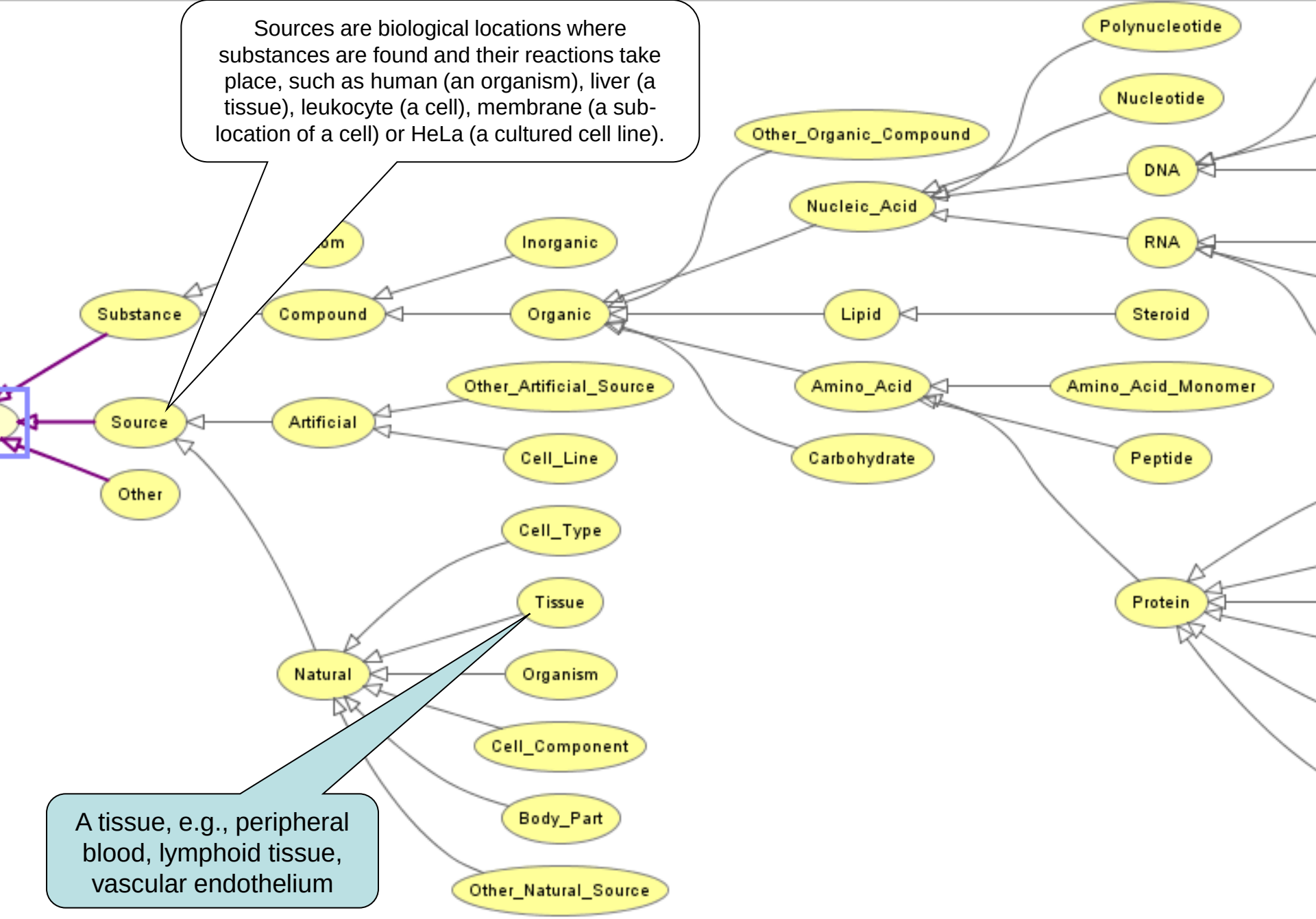
Probleme der Genia-Ontologie

- Taxonomie (Begriffshierachie), keinerlei Anbindung an domänenunabhängige “Upper Ontology”
- Keine Relationen außer Klasseninklusion (Is-A)
- Definitionen nur in natürlichsprachliche Ausdrücken, meist unscharf, teils rein extensional


Sources are biological locations where substances are found and their reactions take place, such as human (an organism), liver (a tissue), leukocyte (a cell), membrane (a sub-location of a cell) or HeLa (a cultured cell line).
Polynucleotide
Nucleotide
Other_Organic_Compound
DNA
Nucleic_Acid
RNA
Inorganic
Substance
Compound
Organic
Lipid
Steroid
Other_Artificial_Source
Amino_Acid
Amino_Acid_Monomer
Source
Artificial
Cell_Line
Carbohydrate
Peptide
Other
Cell_Type
Protein
Tissue
Natural
Organism
Cell_Component
Body_Part
Other_Natural_Source

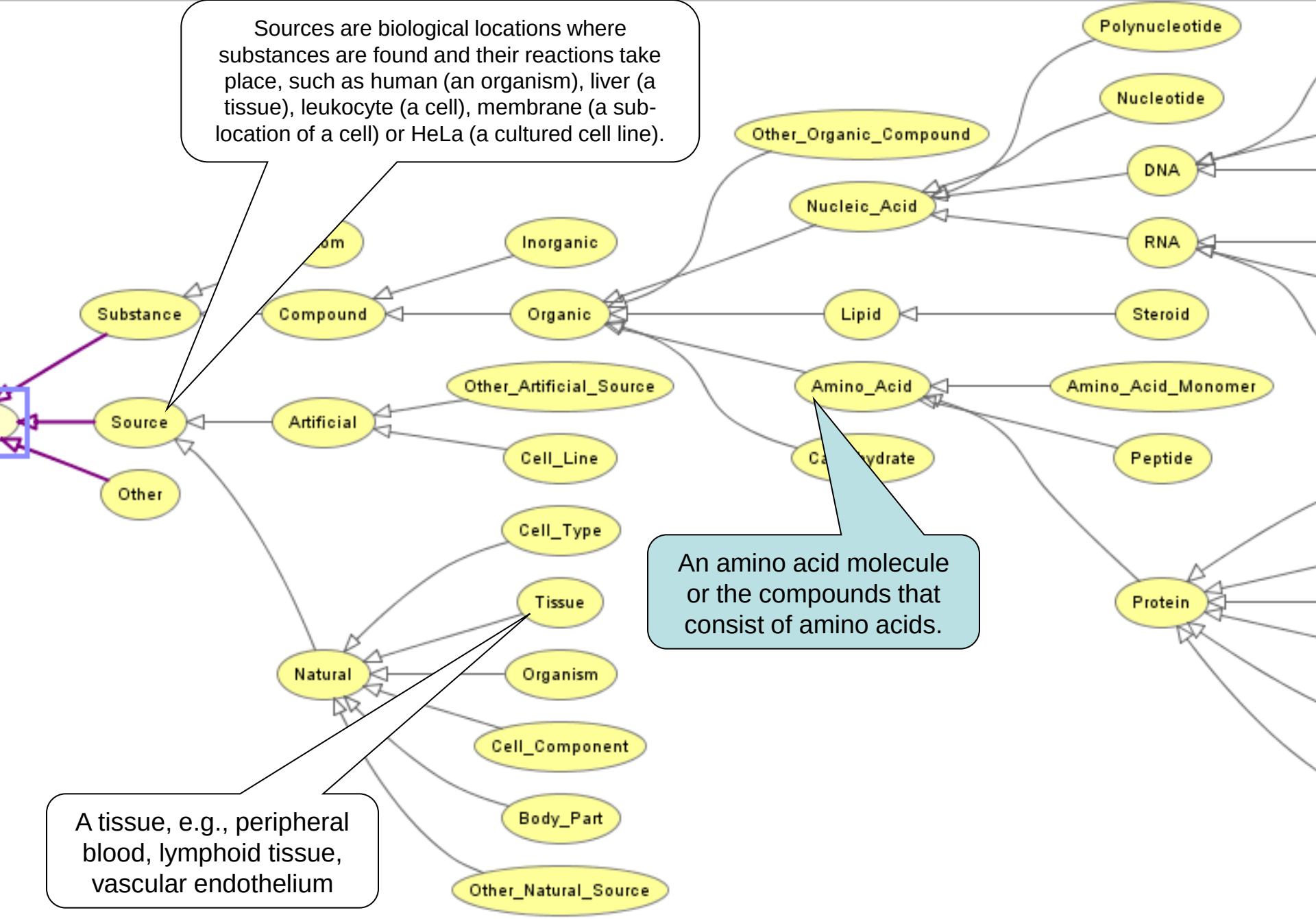
***Sources**: "Sources are biological locations where substances are found and their reactions take place, such as human (an organism), liver (a tissue), leukocyte (a cell), membrane (a sub-location of a cell) or HeLa (a cultured cell line)".*

- Klasseneinteilung sollte gemäß stabiler Merkmale der zu klassifizierenden Entitäten erfolgen. (Zellen können sowohl in Organismen als auch in Gewebekulturen vorkommen)
- "Source" ist eine Rolle, kein diskriminierendes Merkmal


Sources are biological locations where substances are found and their reactions take place, such as human (an organism), liver (a tissue), leukocyte (a cell), membrane (a sub-location of a cell) or HeLa (a cultured cell line).
Polynucleotide
Nucleotide
Other_Organic_Compound
DNA
Nucleic_Acid
RNA
Inorganic
Substance
Compound
Organic
Lipid
Steroid
Other_Artificial_Source
Amino_Acid
Amino_Acid_Monomer
Source
Artificial
Cell_Line
Carbohydrate
Peptide
Other
Cell_Type
Tissue
Protein
Natural
Organism
Cell_Component
Body_Part
Other_Natural_Source
A tissue, e.g., peripheral blood, lymphoid tissue, vascular endothelium

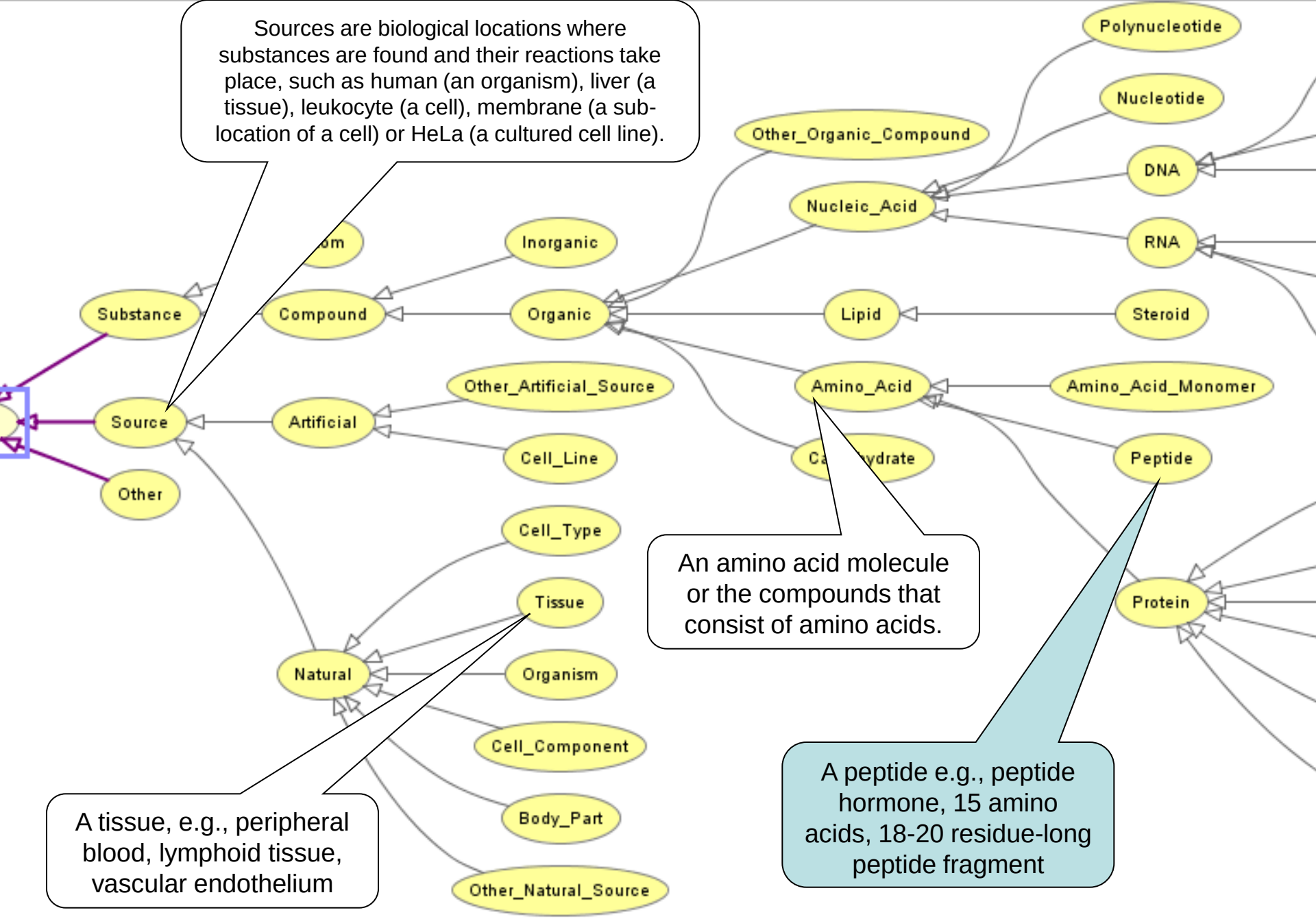
***Tissue**: "A tissue, e.g., peripheral blood, lymphoid tissue, vascular endothelium"*

- Keine Definition
- Rein extensionale Beschreibung: Aufzählung einiger Unterklassen, ohne Angabe differenzierender Kriterien


Sources are biological locations where substances are found and their reactions take place, such as human (an organism), liver (a tissue), leukocyte (a cell), membrane (a sub-location of a cell) or HeLa (a cultured cell line).
Polynucleotide
Nucleotide
Other_Organic_Compound
DNA
Nucleic_Acid
RNA
Inorganic
Substance
Compound
Organic
Lipid
Steroid
Other_Artificial_Source
Amino_Acid
Amino_Acid_Monomer
Source
Artificial
Cell_Line
Peptide
Other
Cell_Type
An amino acid molecule or the compounds that consist of amino acids.
Tissue
Protein
Natural
Organism
Cell_Component
Body_Part
A tissue, e.g., peripheral blood, lymphoid tissue, vascular endothelium
Other_Natural_Source

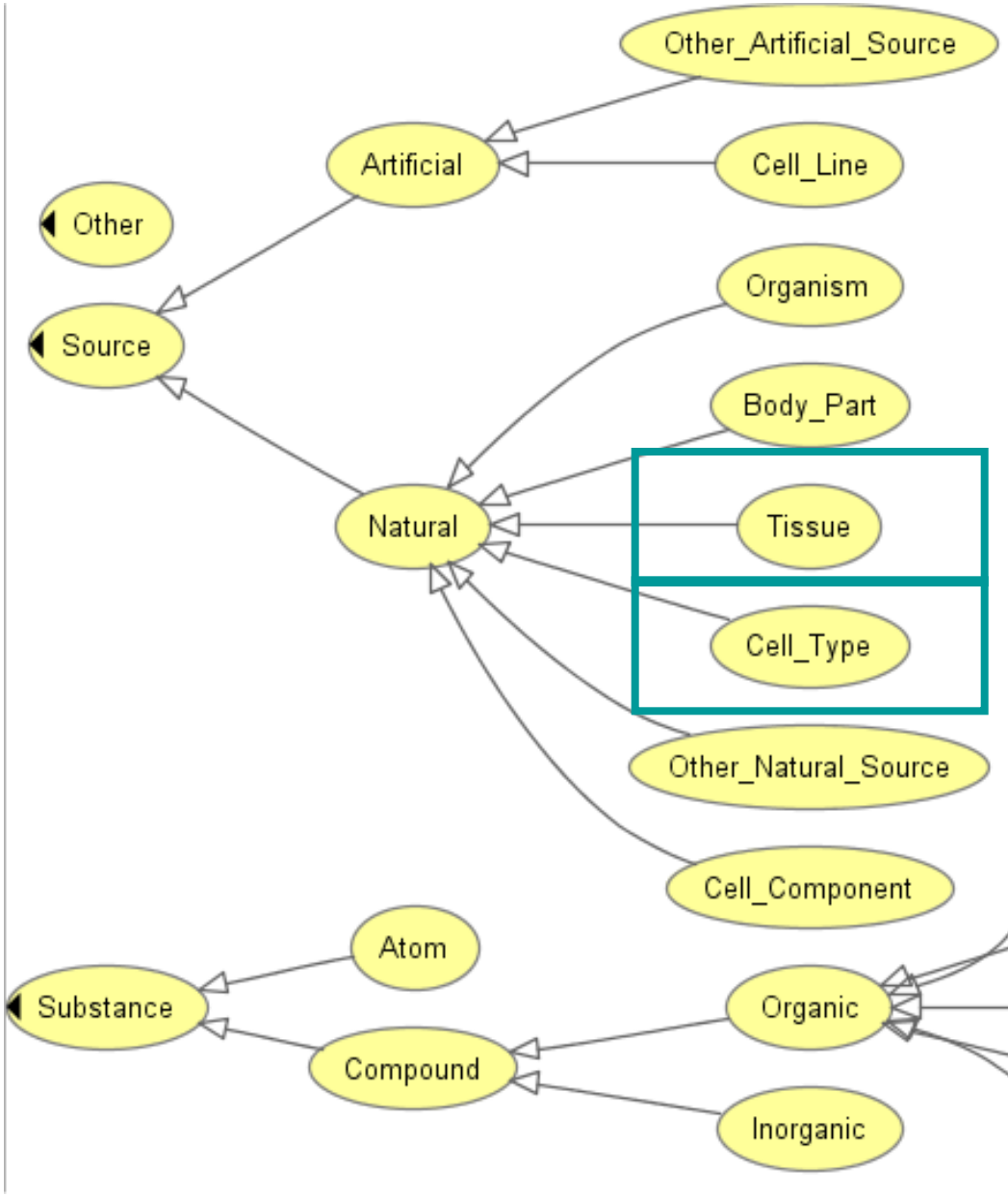
***Amino Acid**: An amino acid molecule or the compounds that consist of amino acids.*

- Sprachlich exakte logische Definition, die jedoch nicht der üblichen Bedeutung von Aminosäure entspricht
- Richtig wäre z.B.
  “Amino_acid_or_amino_acid_containing_biomolecule”


Sources are biological locations where substances are found and their reactions take place, such as human (an organism), liver (a tissue), leukocyte (a cell), membrane (a sub-location of a cell) or HeLa (a cultured cell line).
Polynucleotide
Nucleotide
Other_Organic_Compound
DNA
Nucleic_Acid
RNA
Inorganic
Substance
Compound
Organic
Lipid
Steroid
Other_Artificial_Source
Amino_Acid
Amino_Acid_Monomer
Source
Artificial
Cell_Line
Peptide
Other
Cell_Type
An amino acid molecule or the compounds that consist of amino acids.
Tissue
Protein
Natural
Organism
Cell_Component
A peptide e.g., peptide hormone, 15 amino acids, 18-20 residue-long peptide fragment
Body_Part
A tissue, e.g., peripheral blood, lymphoid tissue, vascular endothelium
Other_Natural_Source

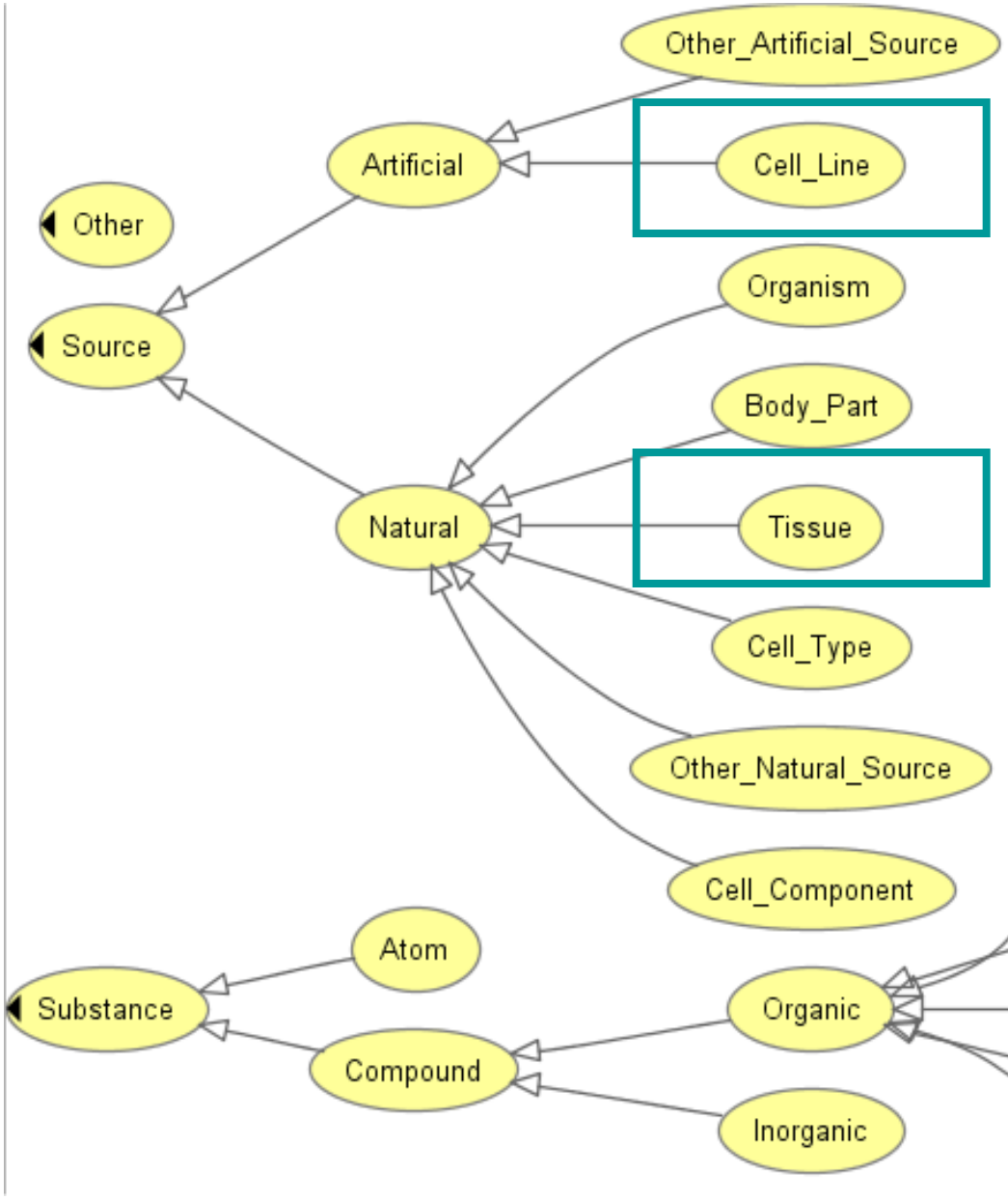
Peptide: A peptide e.g., peptide hormone, 15 amino acids, 18-20 residue-long peptide fragment

- Statt Definition ist eine prototypische Instanz angegeben



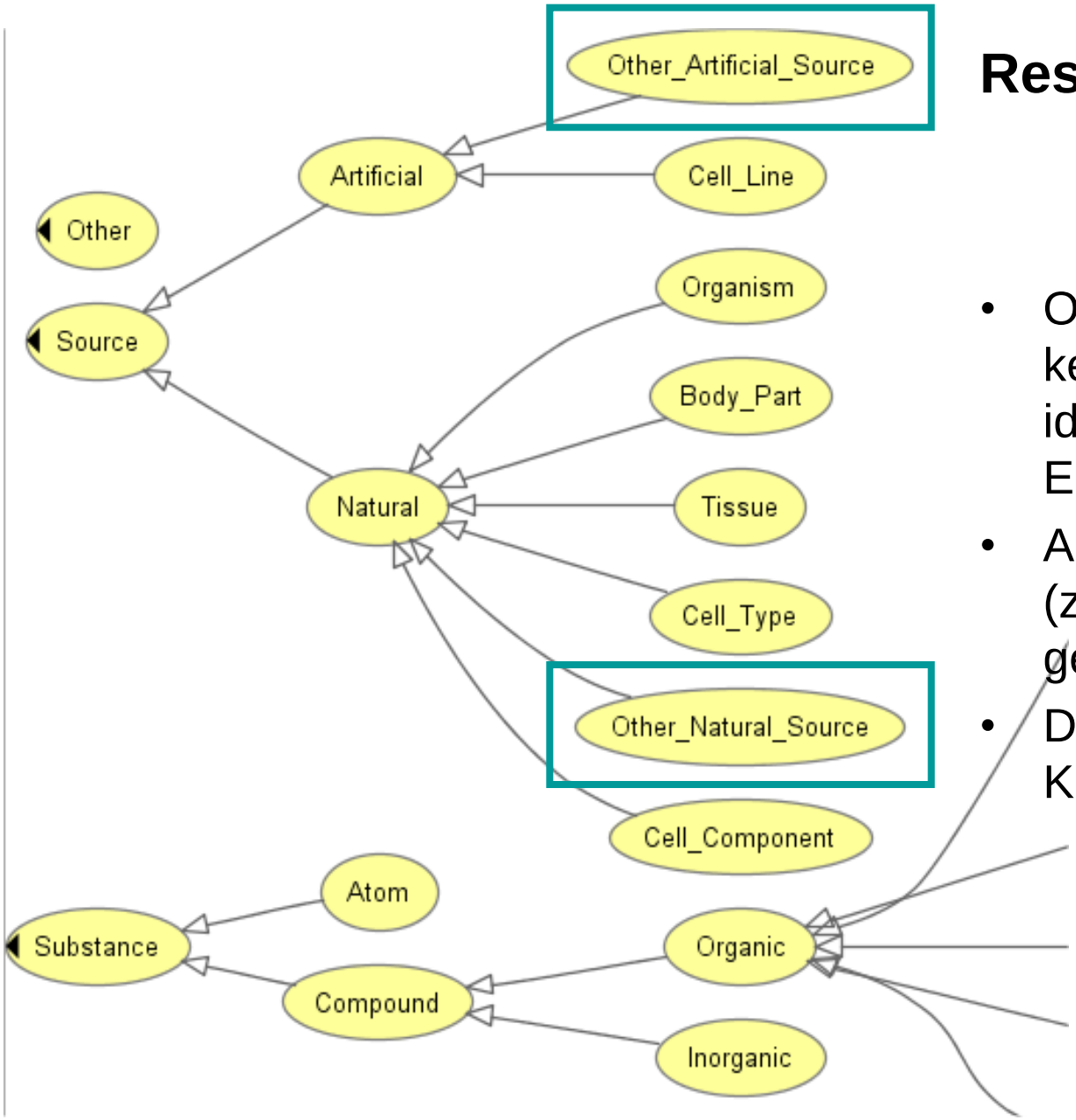
Uneinheitliche Namensgebung: “Cell_Type”, aber warum nicht “Tissue_Type”:

- Verwirrend: Was ist eine Instanz von Cell_Type ?
  - eine Einzelzelle
  - eine Klasse von Zellen?
  - Ein Konzept
- Problem: Die Bezeichnung von Klassen als Typen lässt Meta-Kategorien vermuten. Ist das gewollt ?



Fehlende Anbindung an eine “Upper Ontology” verhindert genaue Charakterisierung.

- Was ist eine Instanz von “Cell_Line”?
  - eine Einzelzelle
  - eine Menge von Einzelzellen
  - eine Zellfamilie
- Was ist eine Instanz von “Tissue”:
  - eine genau umrissene Gewebeprobe ?
  - eine arbiträre Menge von Gewebe
  - die Gesamtheit allen Gewebes

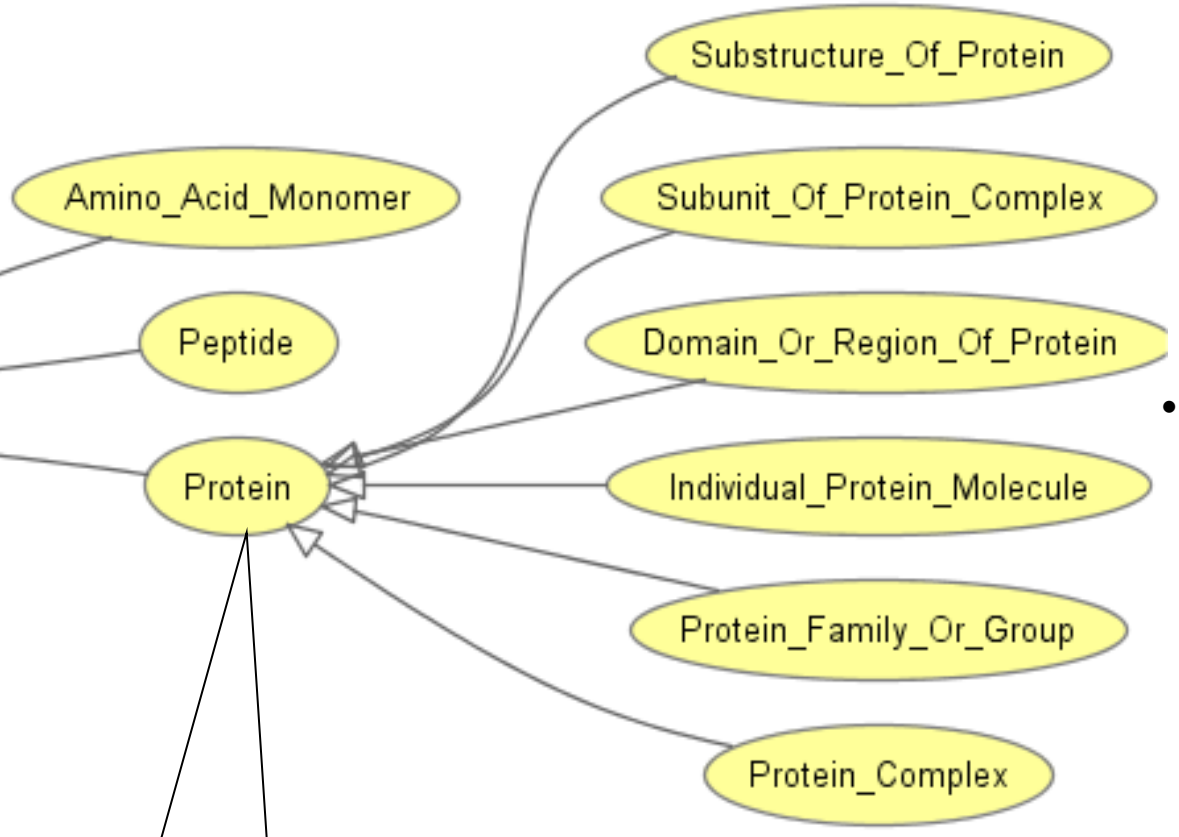


Resteklassen

- Ontologisch irrelevant, da keine gemeinsame, identitätsstiftende Eigenschaft
- Aus praktischen Gründen (zur Annotation) gerechtfertigt.
- Definition als logisches Komplement

Geschwisterklassen (“siblings”)

- In GENIA als taxonomische Unterklassen oft bedenklich
  - ist Substructure of Protein nicht eher part-of Protein ?
  - ist eine Instanz von Protein_Family_Or_Group eine Instanz von Protein ?
- Bilden die Siblings eine komplette Partition, oder gibt es Überlappungen oder Lücken?

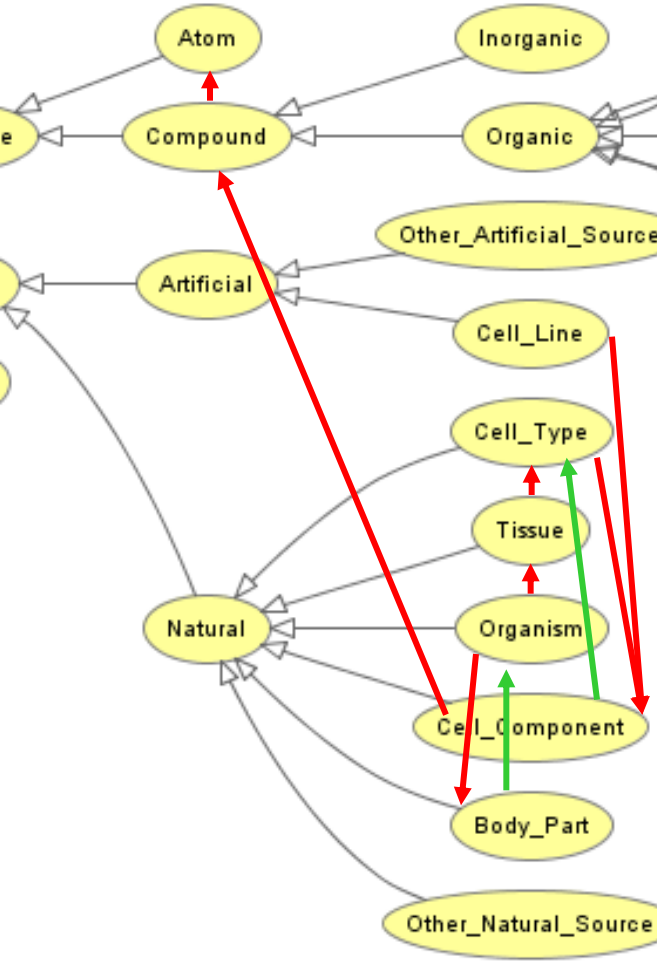


Siblings unvergleichbar

“Proteins include protein groups, families, molecules, complexes, and substructures”

Definition der Oberklasse unscharf

Partonomien als 2. wichtiges Ordnungsprinzip für Ontologien

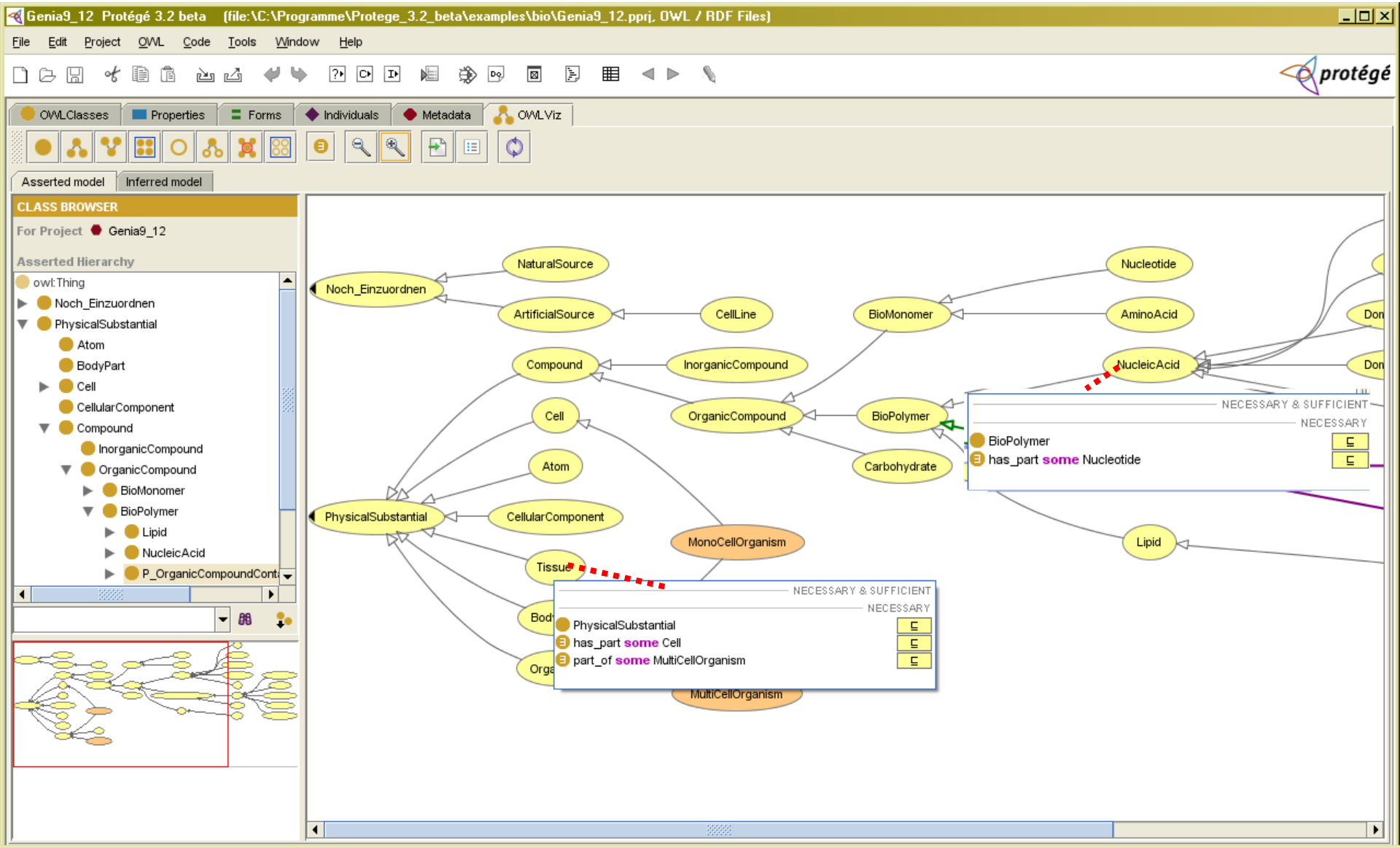


- In OBO gleichberechtigt zu Is-A
- In Genia nicht oder höchstens implizit in Klassennamen (Body_Part) vorhanden
- Part-Of und Has-Part: Transitive Relationen zwischen Klassen
- Definition nach OBO (Smith et al.)
  - A Part-Of B heißt: jede Instanz von A ist Teil einer Instanz von B
  - B Has-Part A heißt: für jede Instanz von B gibt es eine Instanz von A, die davon Teil ist
  - Wichtig: A Part-Of B impliziert nicht B Has-Part A

Von Genia zu Genia-OWL

- OWL (ontology web language): standarisierte, logikbasierte Sprache des Semantic Web
- Genia-OWL: logikbasierte Definition der Genia-Klassen:
  - eindeutige Definitionen
  - weitgehende Abstraktion von natürlicher Sprache
  - höhere Reliabilität bei der Annotation
  - Interface zu anderen formalen Ontologien
  - maschinelles Schließen

Genia-OWL


Genia9_12 Protégé 3.2 beta (file:\C:\Programme\Protege_3.2_beta\examples\bio\Genia9_12.pprj, OWL / RDF Files)
File Edit Project OWL Code Tools Window Help
protégé
OWLClasses Properties Forms Individuals Metadata OWLViz
Asserted model Inferred model
CLASS BROWSER
For Project Genia9_12
Asserted Hierarchy
owl:Thing
Noch_Einzuordnen
PhysicalSubstantial
Atom
BodyPart
Cell
CellularComponent
Compound
InorganicCompound
OrganicCompound
BioMonomer
BioPolymer
Lipid
NucleicAcid
NaturalSource
ArtificialSource
CellLine
Nucleotide
AminoAcid
Carbohydrate
MonoCellOrganism
Tissue
MultiCellOrganism
NECESSARY & SUFFICIENT
NECESSARY
BioPolymer
has_part some Nucleotide
PhysicalSubstantial
has_part some Cell
part_of some MultiCellOrganism
Start
5 Internet ...
2 Windows...
voucher.txt - ...
2 Microsoft...
kr2000 - sch...
Protege.exe
Genia9_1...
19:12

$\forall \exists \leftarrow \rightarrow \leftrightarrow \wedge \vee$

$\forall x{:}P(x) \rightarrow \exists y,z{:}\ x=y+z \wedge (P(y) \vee M(y)) \wedge (P(z) \vee M(z))$