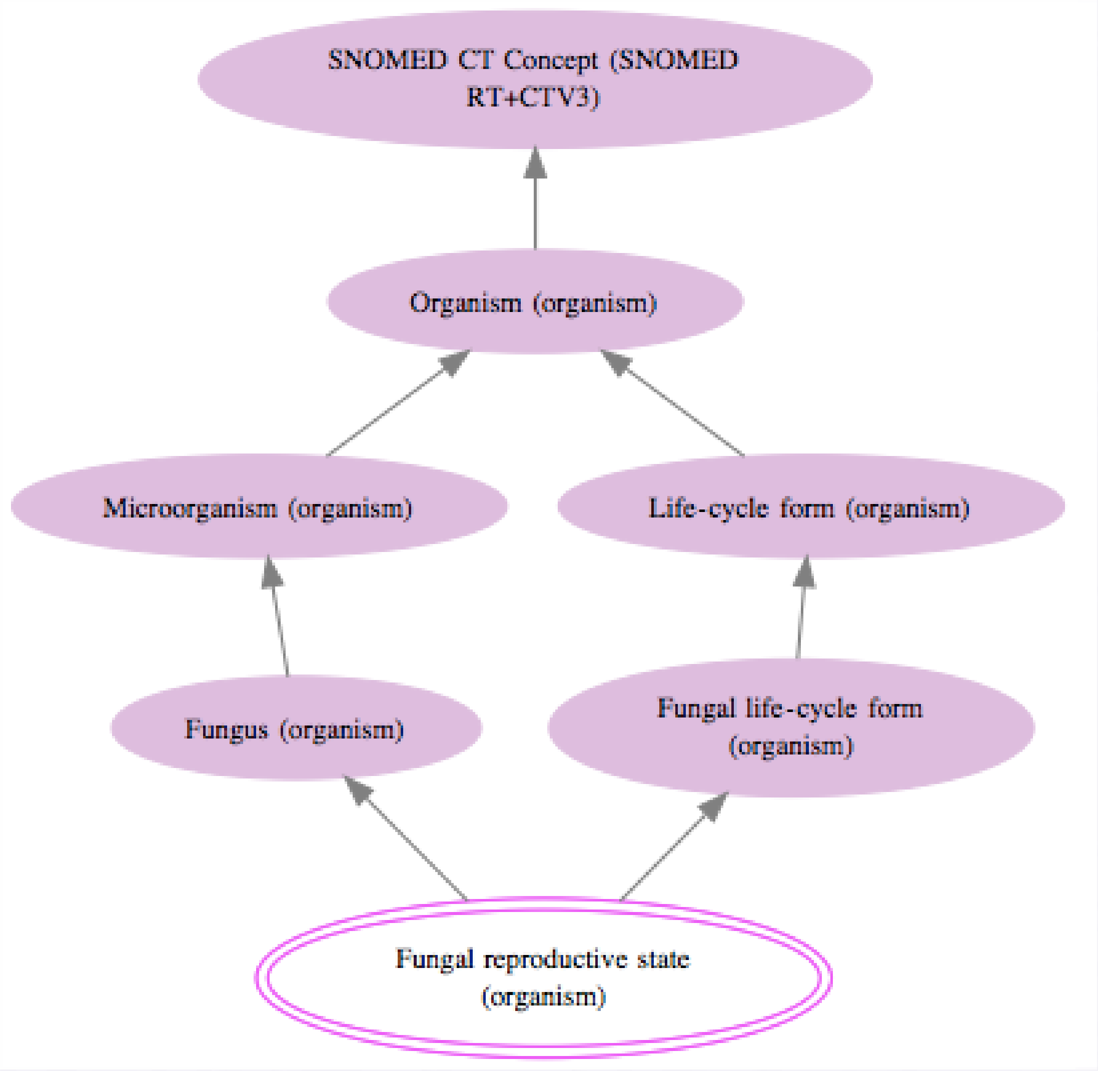
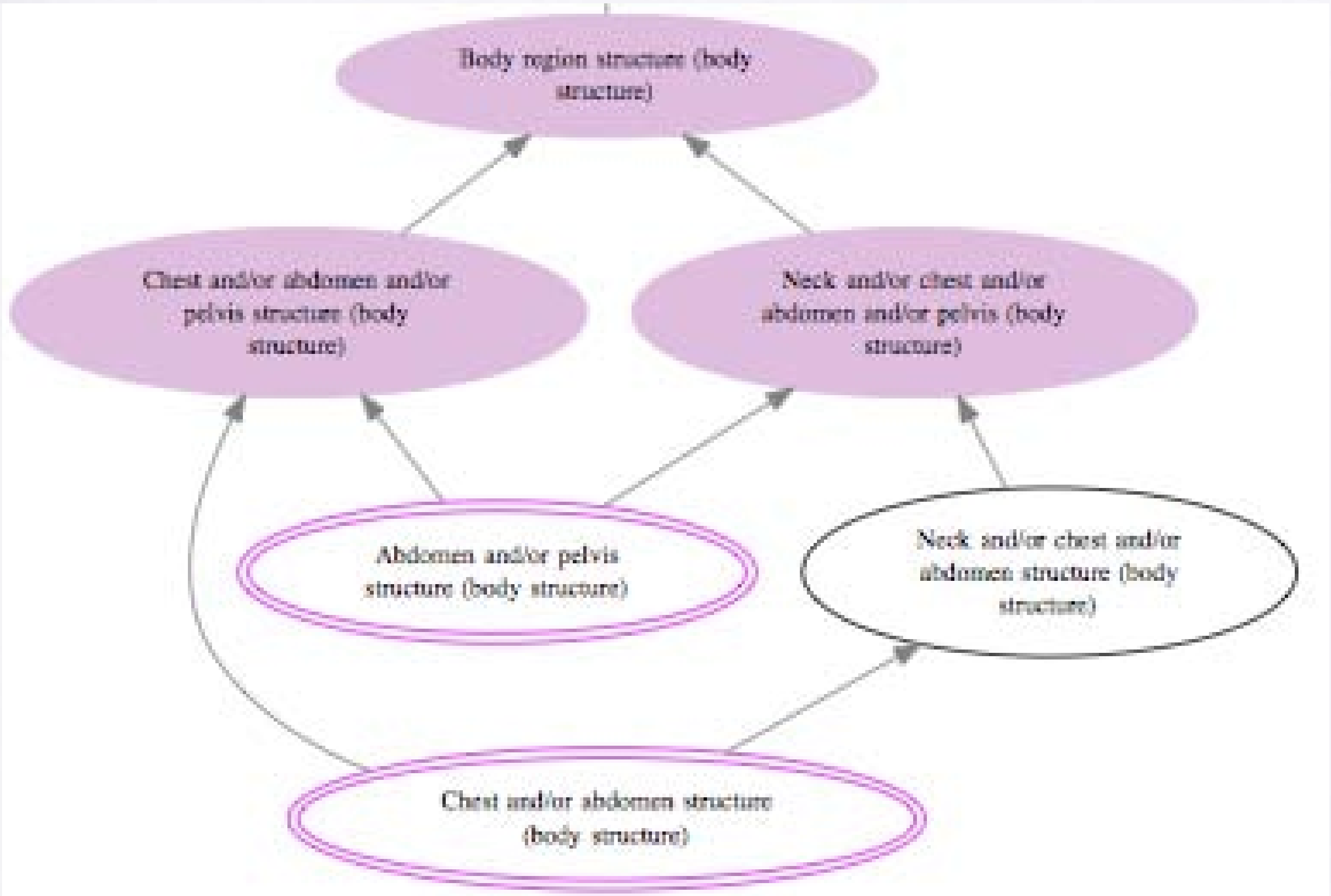
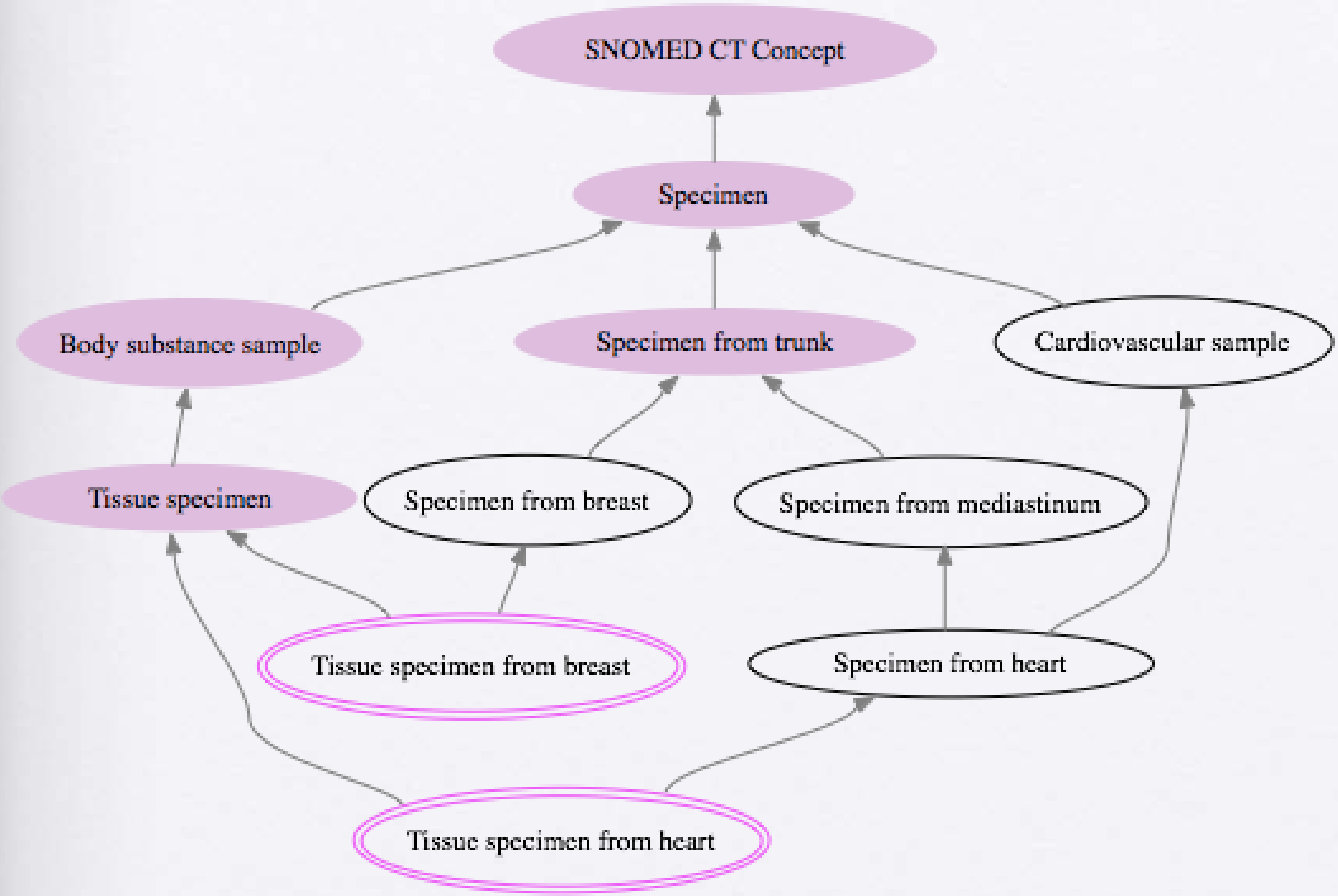
Lattice-based Audit of Ontologies

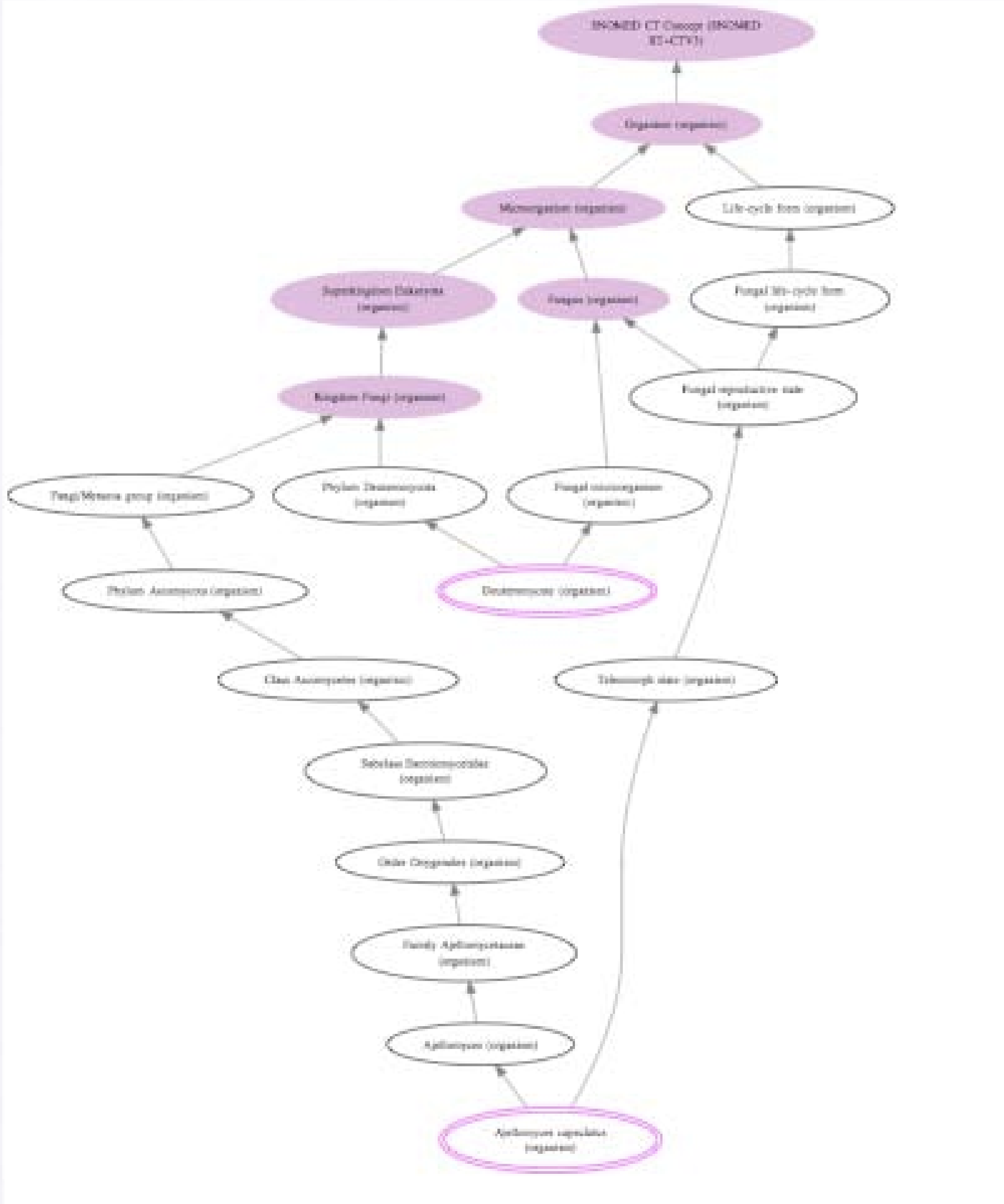
GQ Zhang
Case Western Reserve University
Joint work with Olivier Bodenreider, NLM


SNOMED CT Concept (SNOMED RT+CTV3)
Organism (organism)
Microorganism (organism)
Life-cycle form (organism)
Fungus (organism)
Fungal life-cycle form (organism)
Fungal reproductive state (organism)


Body region structure (body structure)
Chest and/or abdomen and/or pelvis structure (body structure)
Neck and/or chest and/or abdomen and/or pelvis (body structure)
Abdomen and/or pelvis structure (body structure)
Neck and/or chest and/or abdomen structure (body structure)
Chest and/or abdomen structure (body structure)



SNOMED CT Concept
Specimen
Body substance sample
Specimen from trunk
Cardiovascular sample
Tissue specimen
Specimen from breast
Specimen from mediastinum
Tissue specimen from breast
Specimen from heart
Tissue specimen from heart



SNOMED CT Concept (SNOMED RT+CTV3)
Organism (organism)
Microorganism (organism)
Life-cycle form (organism)
Superkingdom Eukaryota (organism)
Fungus (organism)
Fungal life-cycle form (organism)
Kingdom Fungi (organism)
Fungal reproductive state (organism)
Fungi/Metazoa group (organism)
Phylum Deuteromycota (organism)
Fungal microorganism (organism)
Phylum Ascomycota (organism)
Deuteromycete (organism)
Class Ascomycetes (organism)
Teleomorph state (organism)

Extension and Intension

The extension of a concept is just the collection of individual things to which it is correctly applied. Thus, the extension of the word “chair” includes every chair that is (or ever has been or ever will be) in the world.

The intension of a concept, on the other hand, is the set of features which are shared by everything to which it applies. Thus, the intension of the word “chair” is (something like) “a piece of furniture designed to be sat upon by one person at a time.”

Formal Concept Analysis

Mathematical modeling of intension and extension leads to Formal Concept Analysis (FCA). The starting point (input data) for FCA is a binary relation called a formal context, consisting of the following components:

- a set G of objects,
- a set M of attributes, and
- a relation I from G to M.

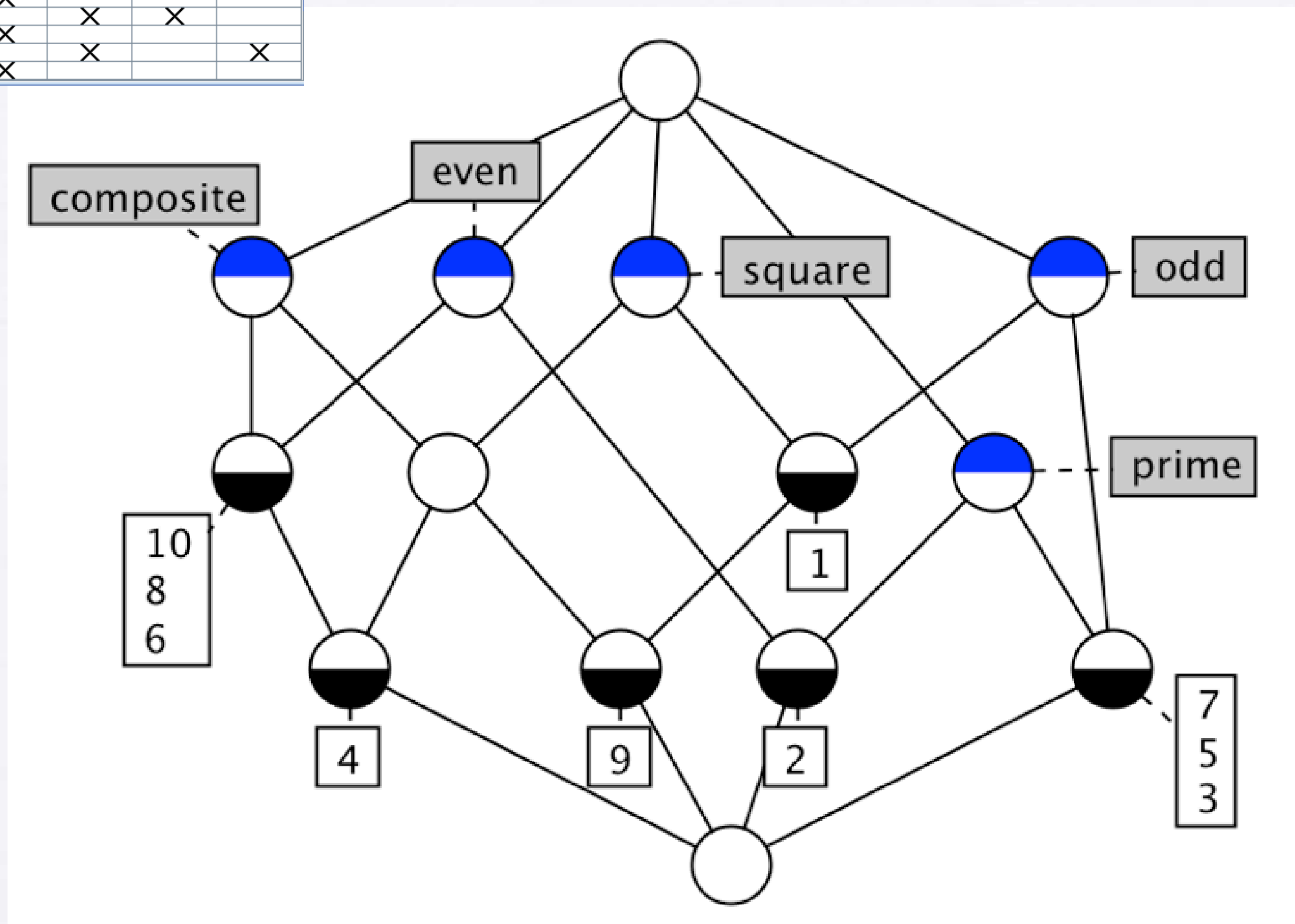
A formal context is often written by collecting these components together as a tuple, (G,M,I).

Example Context

A	B	C	D	E	F
	composite	even	odd	prime	square
1			X		X
2		X		X	
3			X	X	
4	X	X			X
5			X	X	
6	X	X			
7			X	X	
8	X	X			
9	X		X		X
10	X	X			

From Contexts to Lattices

A	B	C	D	E	F
	composite	even	odd	prime	square
1			X		X
2		X		X	
3			X	X	
4	X	X			X
5			X	X	
6	X	X			
7			X	X	
8	X	X			
9	X		X		X
10	X	X			



Formal Concept Analysis I

A formal context (or context) (G, M, I) consists of two sets G and M and a relation $I \subseteq G \times M$. The elements of G are called the objects and the elements of M are called the attributes of the context. In order to express that an object g is in a relation I with an attribute m, we write $(g, m) \in I$.
For a set $O \subseteq G$, define

$$O' := \{m \in M \mid (g, m) \in I \text{ for all } g \in O\}$$

(the set of all attributes common to all the objects in O).
Correspondingly, for a set $A \subseteq M$ define

$$A' := \{g \in G \mid (g, m) \in I \text{ for all } m \in A\}$$

(the set of all objects which have all the attributes in A).

Formal Concept Analysis II

A formal concept (or concept) of the context (G, M, I) is a pair (O, A) with $O \subseteq G$, $A \subseteq M$, $O' = A$ and $A' = O$.

O is said to be the extent and A is said to be intent of the concept (O, A).

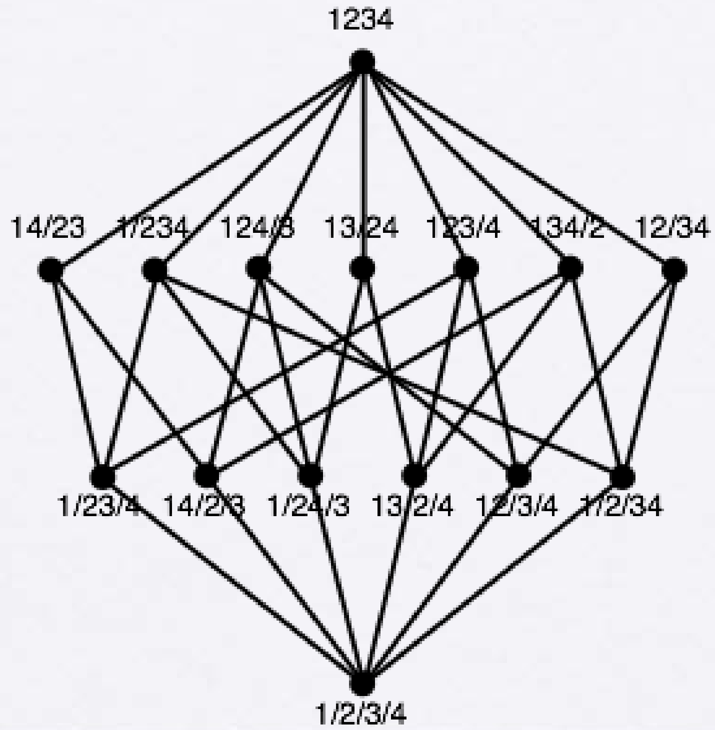
For an object $g \in G$ we write g' instead of $\{g\}'$. Similarly, for $m \in M$, we write m' for $\{m\}' := \{g \in G \mid (g, m) \in I\}$. Example

I		Permission (M)						
		p_1	p_2	p_3	p_4	p_5	p_6	p_7
User (G)	u_1	×	×			×	×	×
	u_2				×	×	×	×
	u_3	×	×		×	×		
	u_4	×	×					

Remark. One can safely remove empty, or repeated rows or columns for RBAC.

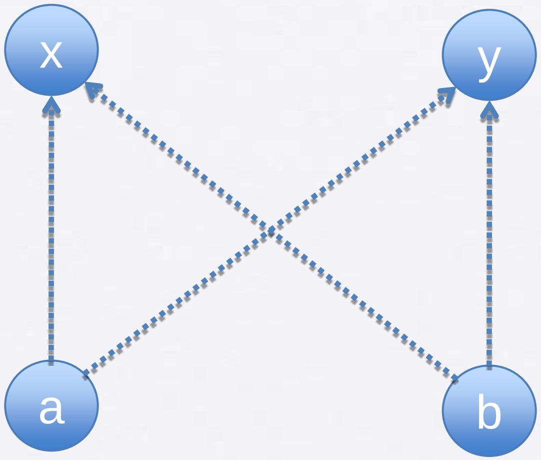
What Is a Lattice?

In mathematics, a **lattice** is a partially ordered set (also called a *poset*) in which any *two elements* have a unique supremum (the elements' least upper bound; called their **join**) and an infimum (greatest lower bound; called their **meet**).
-Wikipedia

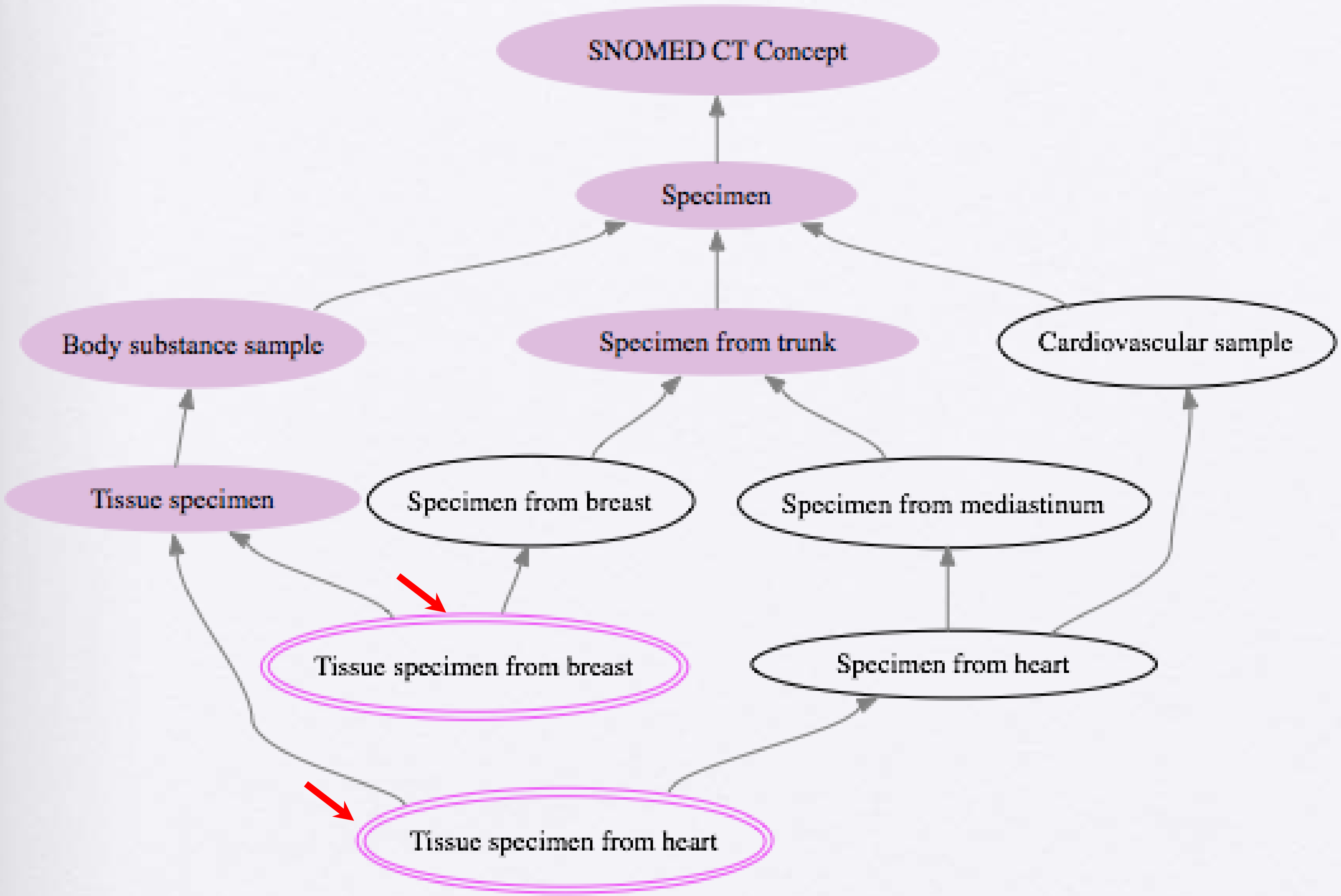


A Pattern Every Finite Non-Lattice Must Contain

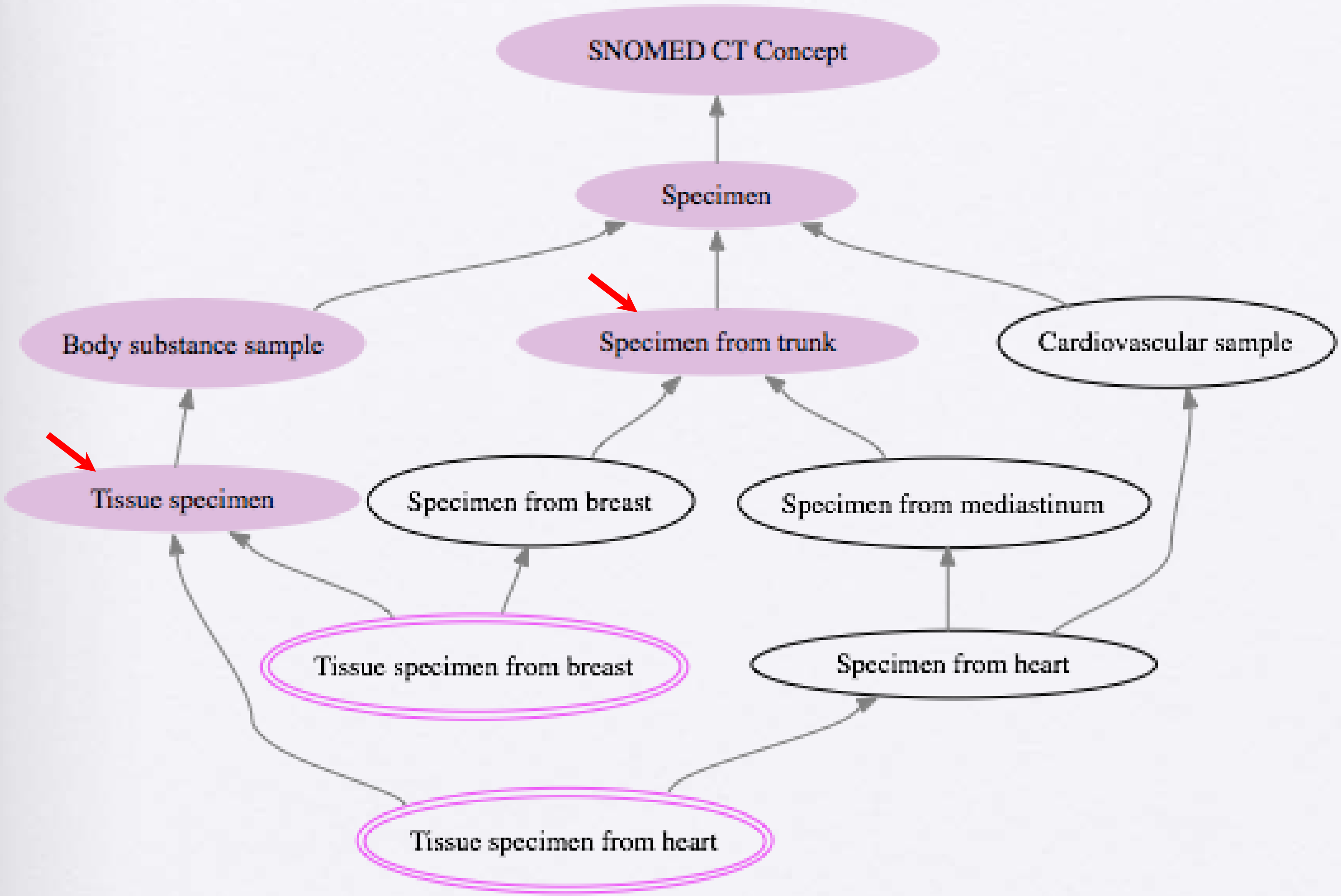
... a **lattice** is ... in which any *two elements* have a unique supremum (the least upper bound, the youngest common ancestor ...



A Non-Lattice Fragment in SNOMED-CT



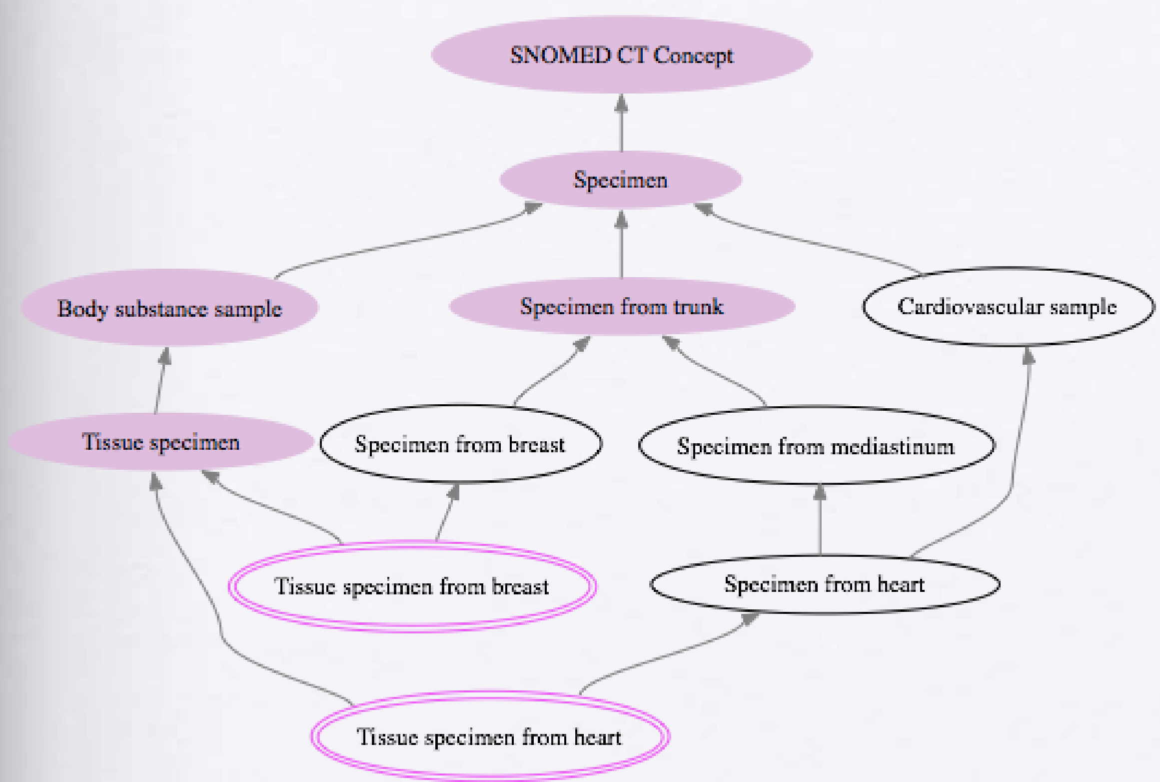
A Non-Lattice Fragment in SNOMED-CT



Research Questions

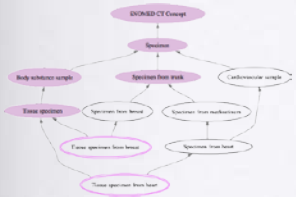
- **Automate**: find all problematic fragments - concept pairs in SNOMED-CT that have at least two youngest ancestors
- **Classify**: group similar problematic fragments together and explain the possible causes
- **Compare**: compare with other existing auditing techniques
- **Resolve**: suggest ways to minimize or eliminate problematic fragments; this will necessarily be a community effort and needs collaborative tools to support ontological evolution
- **Expand**: apply this tool methodology to other ontological systems such as GO, LOINC, ...

Graphs and RDF Triples



"Tissue specimen from breast" **subClassOf** "Tissue specimen"
"Tissue specimen from heart" **subClassOf** "Tissue specimen"
"Tissue specimen from heart" **subClassOf** "Specimen from heart"
"Tissue specimen from breast" **subClassOf** "Specimen from breast"
"Tissue specimen" **subClassOf** "Body substance sample"
…

SPARQL and Virtuoso



"Tissue specimen from breast" **subClassOf** "Tissue specimen"
"Tissue specimen from heart" **subClassOf** "Tissue specimen"
"Tissue specimen from heart" **subClassOf** "Specimen from heart"
"Tissue specimen from breast" **subClassOf** "Specimen from breast"
"Tissue specimen" **subClassOf** "Body substance sample"
...

```
SELECT    ?sbname count(?sb) as ?sb_links

FROM <http://newton.case.edu/TEST>

WHERE {
{
 { :INST-128166000 isof:type  ?s .
   :INST-127457009 isof:type  ?t .
   ?s ?x  ?sb .
   ?t rdfs:subClassOf ?sb .
         }
union
{   :INST-128166000 isof:type  ?s .
   :INST-127457009 isof:type  ?t .
   ?s ?x  ?sa .
   ?t rdfs:subClassOf ?sa .
   ?sa rdfs:subClassOf ?sb .
} }
?sb rdfs:label ?sbname .
}
ORDER BY ASC (?sb_links)
LIMIT 2
```

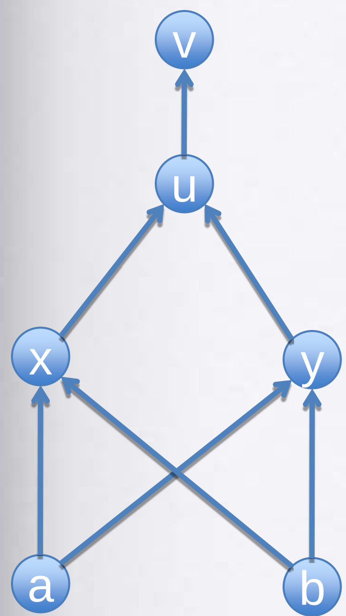
sbname	sb_links
Tissue specimen	1
Specimen from trunk	1

How to make it work

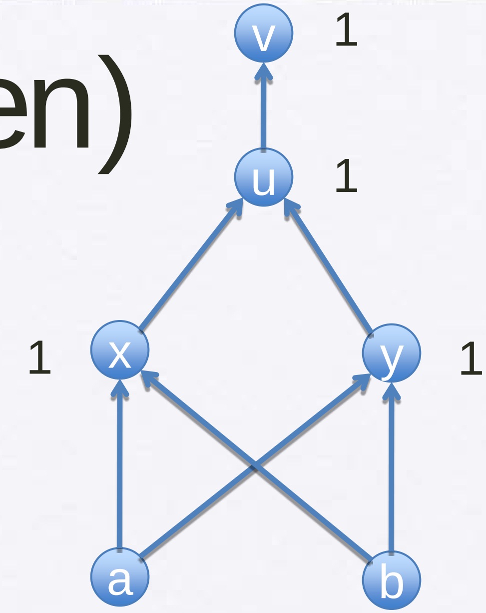
- In principle ...
- In practice ...

In Principle (forbidden)

Forbidden patterns: with >=2 ?sb nodes of count 1

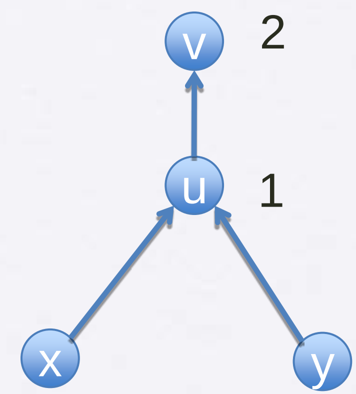


Graph 1:
a rdfs:subClassOf ?sb
b rdfs:subClassOf ?sb



Graph 1 with
counts for ?sb indicated

Graph 2:
a rdfs:subClassOf ?sa
b rdfs:subClassOf ?sa
?sa rdfs:subClassOf ?sb

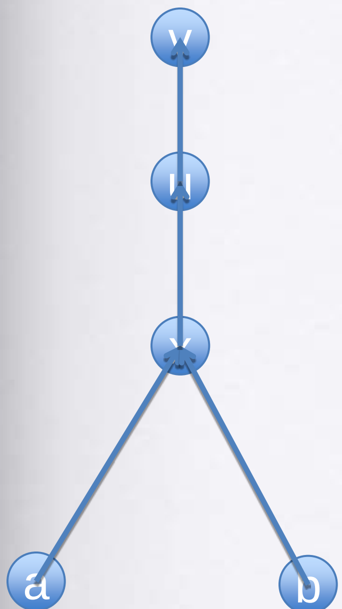


Graph 2 with
counts for ?sb indicated

In Graph 1 Union Graph 2:
count (x)=1, count (y)=1
count (u) = 2, count (v)=3

In Principle (allowed)

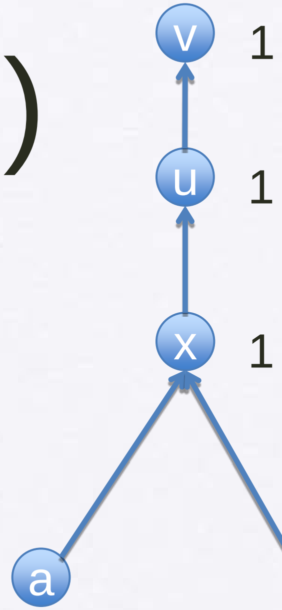
Legal patterns: with a single ?sb node of count 1



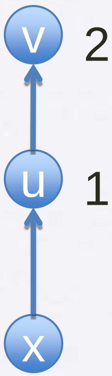
Graph 1:
a rdfs:subClassOf ?sb
b rdfs:subClassOf ?sb

Graph 2:
a rdfs:subClassOf ?sa
b rdfs:subClassOf ?sa
?sa rdfs:subClassOf ?sb

In Graph 1 Union Graph 2:
count (x)=1, count (u)=2
count (v) = 3



Graph 1 with
count for ?sb indicated



Graph 2 with
count for ?sb indicated

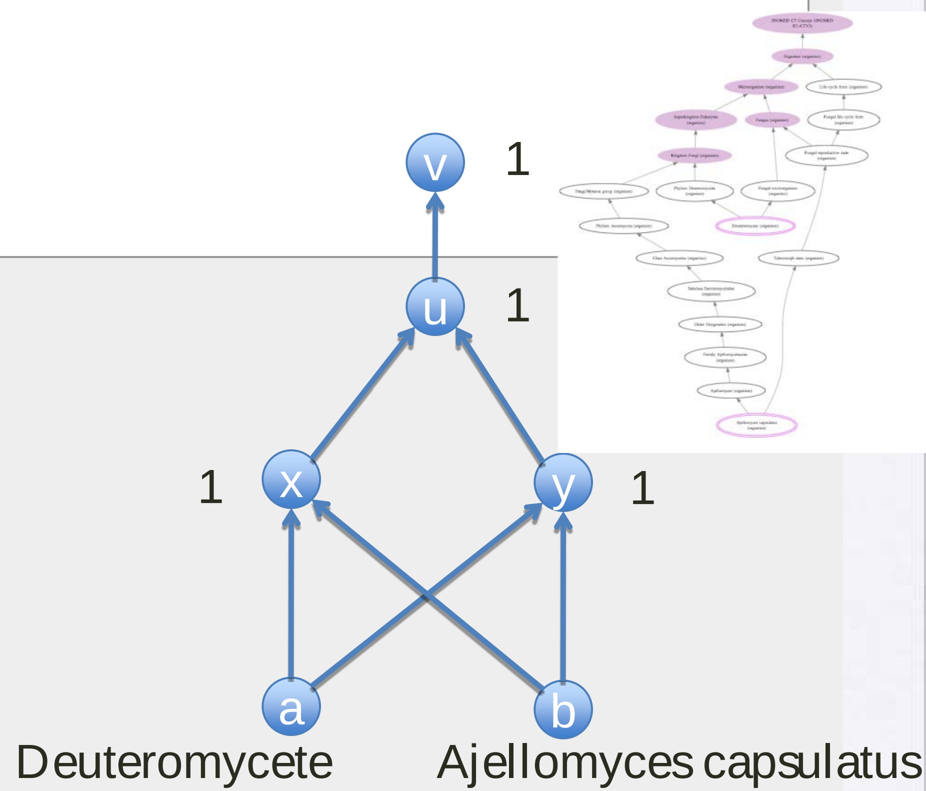
SPARQL Query1-graph1

Query

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix : <http://nlm.nih.gov/SNOMEDCT#>
SELECT  ?sbname  count(?sb) as ?sb_links
FROM <http://newton.case.edu/TEST>
WHERE {
      <http://mor.nlm.nih.gov#SNOMEDCT:45970004> rdfs:subClassOf ?sb .
      <http://mor.nlm.nih.gov#SNOMEDCT:62969002> rdfs:subClassOf ?sb .
       ?sb rdfs:label ?sbname .
}
ORDER BY ASC (?sb_links)
```

Execute Save Load Clear

sbname	sb_links
Superkingdom Eukaryota	1
Microorganism	1
SNOMED CT Concept	1
Organism	1
Kingdom Fungi	1
Fungus	1

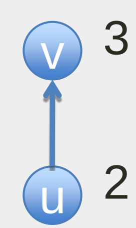


SPARQL Query1-graph2

Query

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix : <http://nlm.nih.gov/SNOMEDCT#>
SELECT  ?sbname  count(?sb) as ?sb_links
FROM <http://newton.case.edu/TEST>
WHERE {
    <http://mor.nlm.nih.gov#SNOMEDCT:45970004> rdfs:subClassOf ?sa .
     <http://mor.nlm.nih.gov#SNOMEDCT:62969002> rdfs:subClassOf ?sa .
     ?sa rdfs:subClassOf ?sb .
     ?sb rdfs:label ?sbname .
}
ORDER BY ASC (?sb_links)
```

Execute   Save   Load   Clear



sbname	sb_links
Superkingdom Eukaryota	1
Microorganism	3
Organism	4
SNOMED CT Concept	5

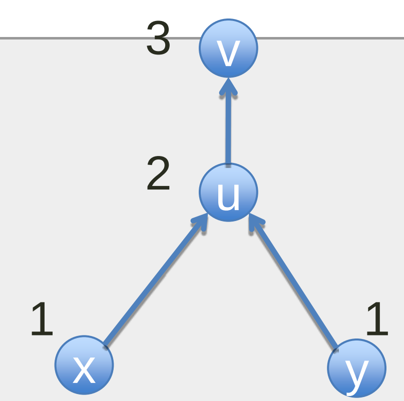
SPARQL Query1-graph1+graph2

Query

```
prefix : <http://nlm.nih.gov/SNOMEDCT#>
SELECT  ?sbname  count(?sb) as ?sb_links
FROM <http://newton.case.edu/TEST>
WHERE { { {
      <http://mor.nlm.nih.gov#SNOMEDCT:45970004> rdfs:subClassOf ?sb .
      <http://mor.nlm.nih.gov#SNOMEDCT:62969002> rdfs:subClassOf ?sb .
  }
  union
  {   <http://mor.nlm.nih.gov#SNOMEDCT:45970004> rdfs:subClassOf ?sa .
      <http://mor.nlm.nih.gov#SNOMEDCT:62969002> rdfs:subClassOf ?sa .
     ?sa rdfs:subClassOf ?sb .  }
?sb rdfs:label ?sbname .
}}
ORDER BY ASC (?sb_links)
LIMIT 2
```

Execute Save Load Clear

sbname	sb_links
Kingdom Fungi	1
Fungus	1



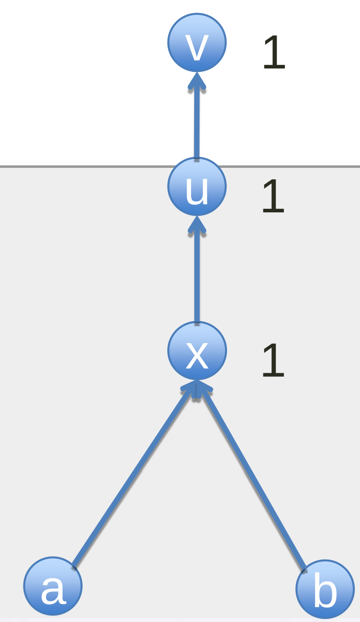
SPARQL Query2-graph1

Query

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix : <http://nlm.nih.gov/SNOMEDCT#>
SELECT  ?sbname  count(?sb) as ?sb_links
FROM <http://newton.case.edu/TEST>
WHERE { { {
      <http://mor.nlm.nih.gov#SNOMEDCT:414561005> rdfs:subClassOf ?sb .
      <http://mor.nlm.nih.gov#SNOMEDCT:23496000> rdfs:subClassOf ?sb .
  }
  union
  {   #<http://mor.nlm.nih.gov#SNOMEDCT:45970004> rdfs:subClassOf ?sa .
      #<http://mor.nlm.nih.gov#SNOMEDCT:62969002> rdfs:subClassOf ?sa .
      # ?sa rdfs:subClassOf ?sb .
}
?sb rdfs:label ?sbname .  }}
ORDER BY ASC (?sb_links)
```

Execute | Save | Load | Clear

sbname	sb_links
Microorganism	1
SNOMED CT Concept	1
Organism	1



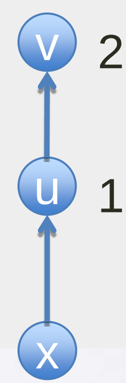
SPARQL Query2-graph2

Query

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix : <http://nlm.nih.gov/SNOMEDCT#>
SELECT  ?sbname  count(?sb) as ?sb_links
FROM <http://newton.case.edu/TEST>
WHERE { { {
#       <http://mor.nlm.nih.gov#SNOMEDCT:414561005> rdfs:subClassOf ?sb .
#       <http://mor.nlm.nih.gov#SNOMEDCT:23496000> rdfs:subClassOf ?sb .
  }
  union
  {    <http://mor.nlm.nih.gov#SNOMEDCT:414561005> rdfs:subClassOf ?sa .
       <http://mor.nlm.nih.gov#SNOMEDCT:23496000> rdfs:subClassOf ?sa .
      ?sa rdfs:subClassOf ?sb .
}
?sb rdfs:label ?sbname .  }}
ORDER BY ASC (?sb_links)
```

Execute Save Load Clear

sbname	sb_links
Organism	1
SNOMED CT Concept	2



SPARQL Query2-graph1+graph2

Query

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix : <http://nlm.nih.gov/SNOMEDCT#>
SELECT  ?sbname  count(?sb) as ?sb_links
FROM <http://newton.case.edu/TEST>
WHERE { { {
      <http://mor.nlm.nih.gov#SNOMEDCT:414561005> rdfs:subClassOf ?sb .
      <http://mor.nlm.nih.gov#SNOMEDCT:23496000> rdfs:subClassOf ?sb .
  }
  union
  {   <http://mor.nlm.nih.gov#SNOMEDCT:414561005> rdfs:subClassOf ?sa .
      <http://mor.nlm.nih.gov#SNOMEDCT:23496000> rdfs:subClassOf ?sa .
     ?sa rdfs:subClassOf ?sb .
}
?sb rdfs:label ?sbname .  }}
ORDER BY ASC (?sb_links)
```

Execute Save Load Clear

sbname	sb_links
Microorganism	1
Organism	2
SNOMED CT Concept	3