Ontology Alignment

Valentina Ivanova

most slides from Patrick Lambrix

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

Ontologies in Biomedical Research

- many biomedical ontologies

  e.g. GO, OBO, SNOMED-CT

- practical use of biomedical ontologies

  e.g. databases annotated with GO

GENE ONTOLOGY (GO)

immune response
  i- acute-phase response
  i- anaphylaxis
  i- antigen presentation
  i- antigen processing
  i- cellular defense response
  i- cytokine metabolism
    i- cytokine biosynthesis
      synonym cytokine production
      …
    p- regulation of cytokine biosynthesis
    …
  …
  i- B-cell activation
    i- B-cell differentiation
    i- B-cell proliferation
  i- cellular defense response
  …
  i- T-cell activation
    i- activation of natural killer cell activity
    …

Ontologies with Overlapping Information

GENE ONTOLOGY (GO)

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    - synonym cytokine production
    - …
  - p- regulation of cytokine biosynthesis
  - …
- …
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response
- …
- i- T-cell activation
  - i- activation of natural killer cell activity
  - …

SIGNAL-ONTOLOGY (SigO)

Immune Response
- i- Allergic Response
- i- Antigen Processing and Presentation
- i- B Cell Activation
- i- B Cell Development
- i- Complement Signaling
  - synonym complement activation
- i- Cytokine Response
- i- Immune Suppression
- i- Inflammation
- i- Intestinal Immunity
- i- Leukotriene Response
  - i- Leukotriene Metabolism
- i- Natural Killer Cell Response
- i- T Cell Activation
- i- T Cell Development
- i- T Cell Selection in Thymus

Ontologies with Overlapping Information

- Use of multiple ontologies
  - custom-specific ontology + standard ontology
  - different views over same domain
  - overlapping domains

- Bottom-up creation of ontologies

  experts can focus on their domain of expertise

→ important to know the inter-ontology relationships

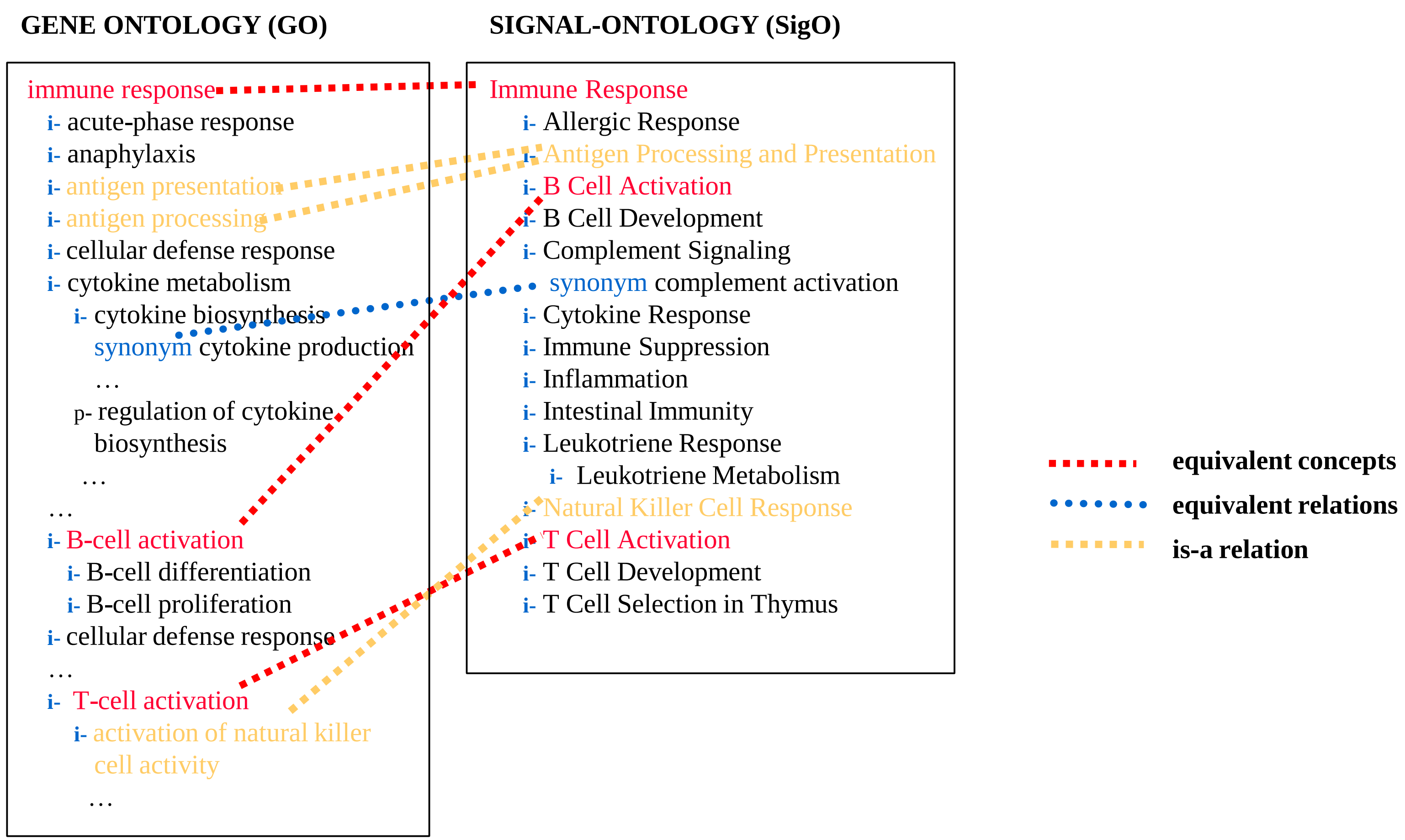
GENE ONTOLOGY (GO)

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  - …
- …
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response
- …
- i- T-cell activation
  - i- activation of natural killer cell activity
  - …

SIGNAL-ONTOLOGY (SigO)

Immune Response
- i- Allergic Response
- i- Antigen Processing and Presentation
- i- B Cell Activation
- i- B Cell Development
- i- Complement Signaling
  synonym complement activation
- i- Cytokine Response
- i- Immune Suppression
- i- Inflammation
- i- Intestinal Immunity
- i- Leukotriene Response
  - i- Leukotriene Metabolism
- i- Natural Killer Cell Response
- i- T Cell Activation
- i- T Cell Development
- i- T Cell Selection in Thymus

Ontology Alignment

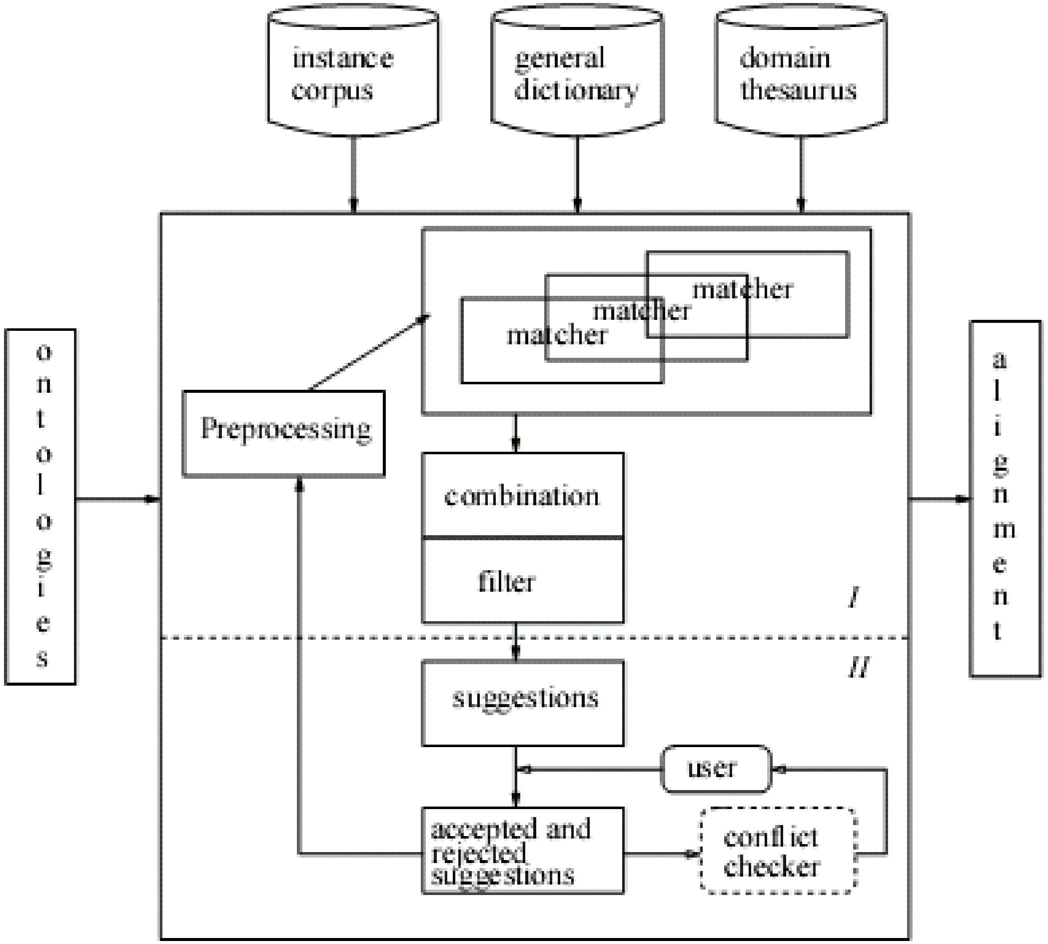


Defining the relations between the terms in different ontologies

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

An Alignment Framework



alignment = a set of mappings

candidate mappings = mapping suggestions = mappings which have not been validated by a user

Classification

- According to input
  - KR: OWL, UML, EER, XML, RDF, …
  - components: concepts, relations, instance, axioms
- According to process
  - What information is used and how?
- According to output
  - 1-1, m-n
  - Similarity vs explicit relations (equivalence, is-a)
  - Confidence (similarity value)

Preprocessing

Preprocessing

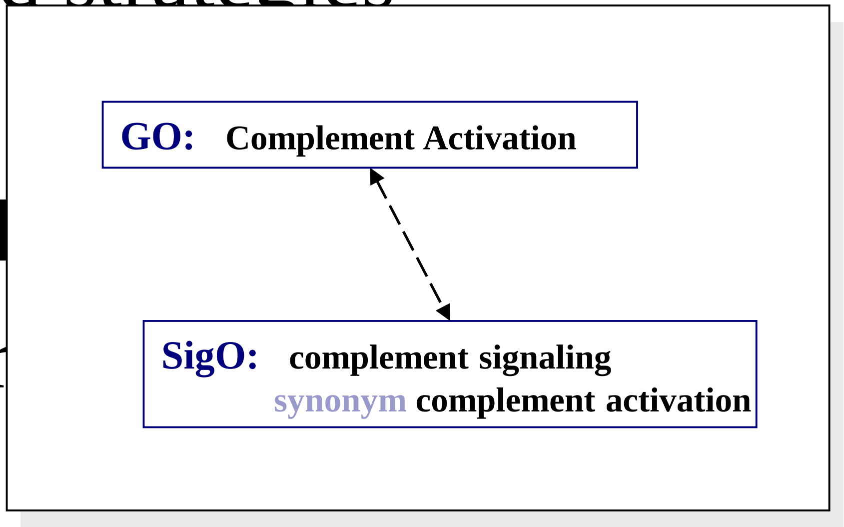
For example,

- Selection of features
- Selection of search space

Matchers

Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-bas
- Instance-based
- Use of auxiliar

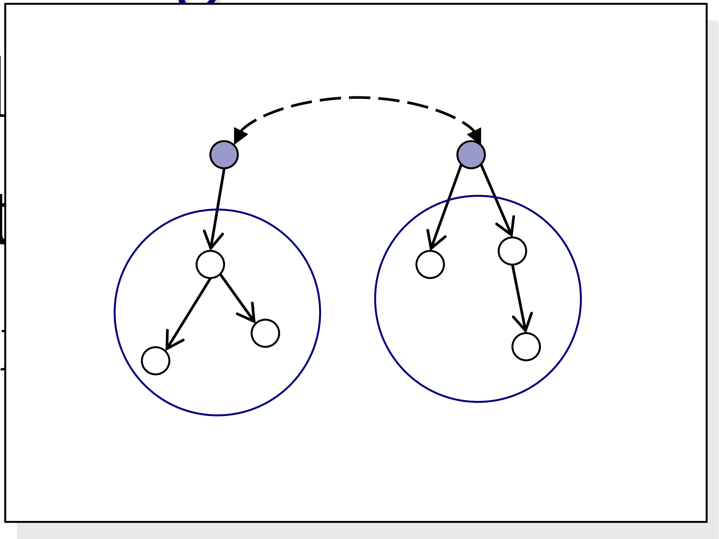


Example Matchers

- Edit distance
  - Number of deletions, insertions, substitutions required to transform one string into another
  - aaaa → baab: edit distance 2
- N-gram
  - N-gram : N consecutive characters in a string
  - Similarity based on set comparison of n-grams
  - aaaa : {aa, aa, aa}; baab : {ba, aa, ab}

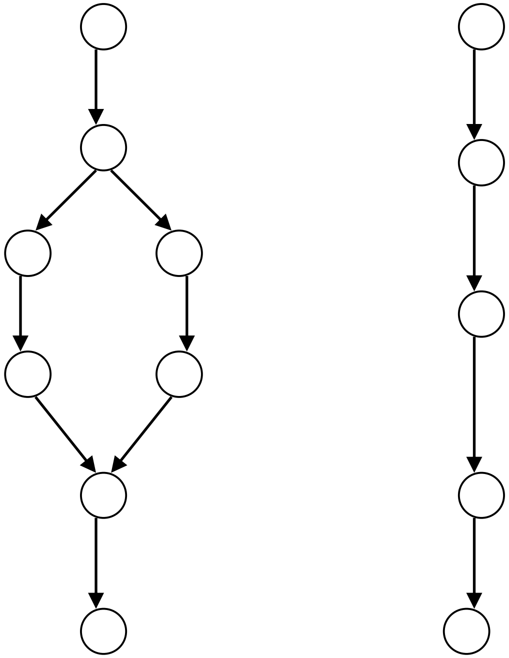
Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based
- Instance-based st
- Use of auxiliary



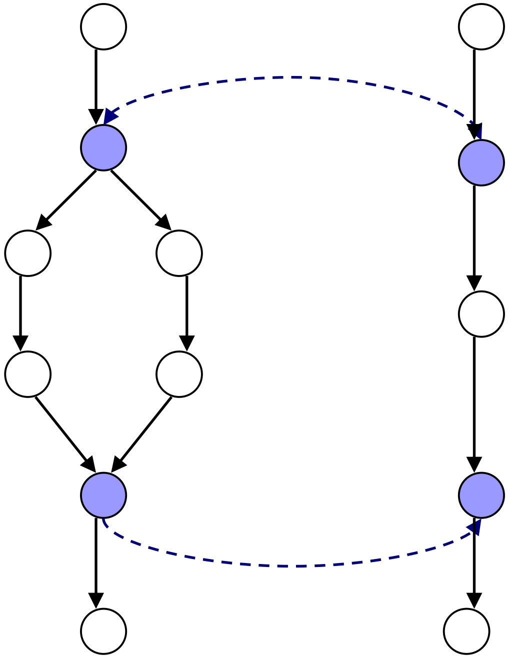
Example Matchers

- Propagation of similarity values
- Anchored matching



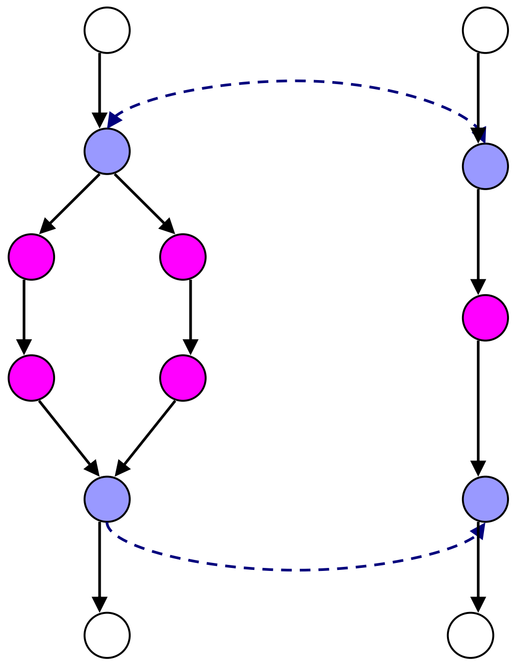
Example Matchers

- Propagation of similarity values
- Anchored matching



Example Matchers

- Propagation of similarity values
- Anchored matching

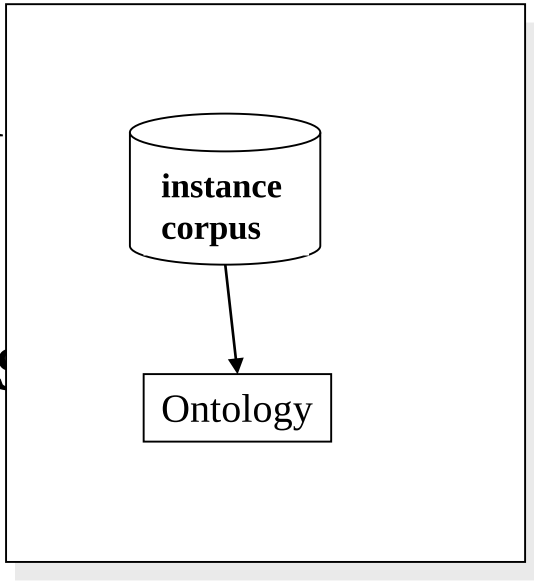


Matcher Strategies

- Strategies based on linguistic matching
- Structure-based strategies
- Constraint-based approaches
- Instance-based strategies
- Use of auxiliary information

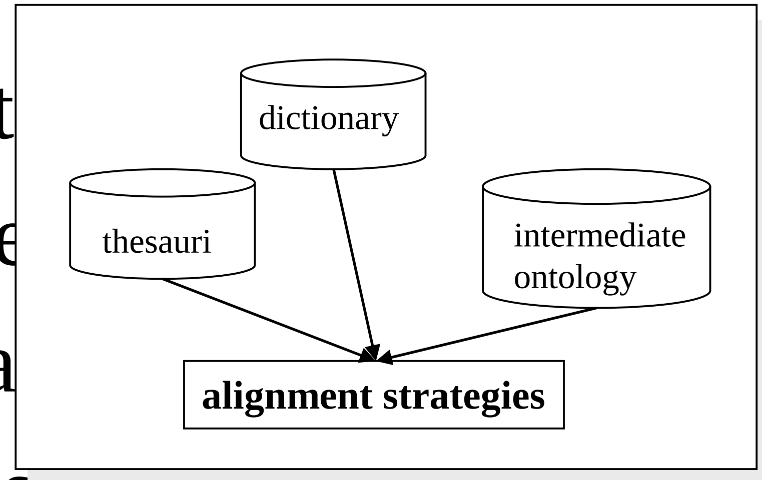
Matcher Strategies

- Strategies based on linguisti
- Structure-based strategies
- Constraint-based approaches
- Instance-based strategies
- Use of auxiliary information



Matcher Strategies

- Strategies based linguist
- Structure-based strategie
- Constraint-based approa
- Instance-based strategies
- Use of auxiliary information



Example Matchers

- Use of WordNet
  - Use WordNet to find synonyms
  - Use WordNet to find ancestors and descendants in the is-a hierarchy
- Use of Unified Medical Language System (UMLS)
  - Includes many ontologies
  - Includes many alignments (not complete)
  - Use UMLS alignments in the computation of the similarity values

Combinations

Combination Strategies

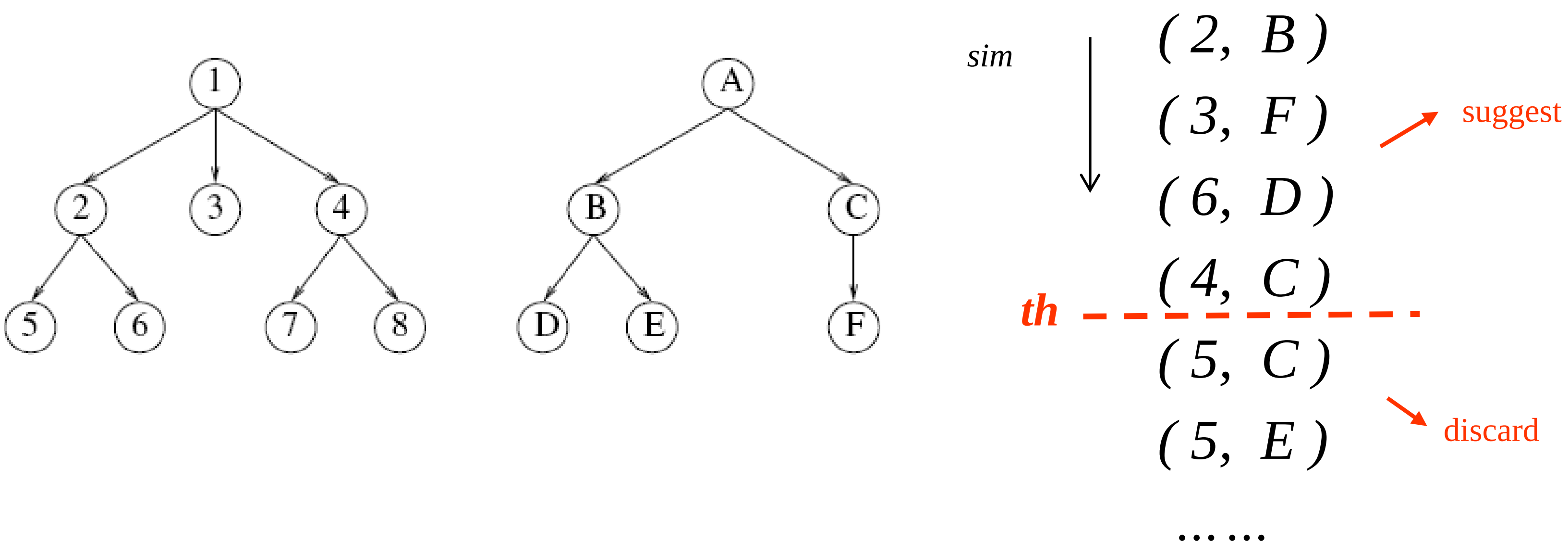
- Usually weighted sum of similarity values of different matchers
- Maximum of similarity values of different matchers

Filtering

Filtering Techniques

- Threshold filtering

  Pairs of concepts with similarity higher or equal than threshold are alignment suggestions

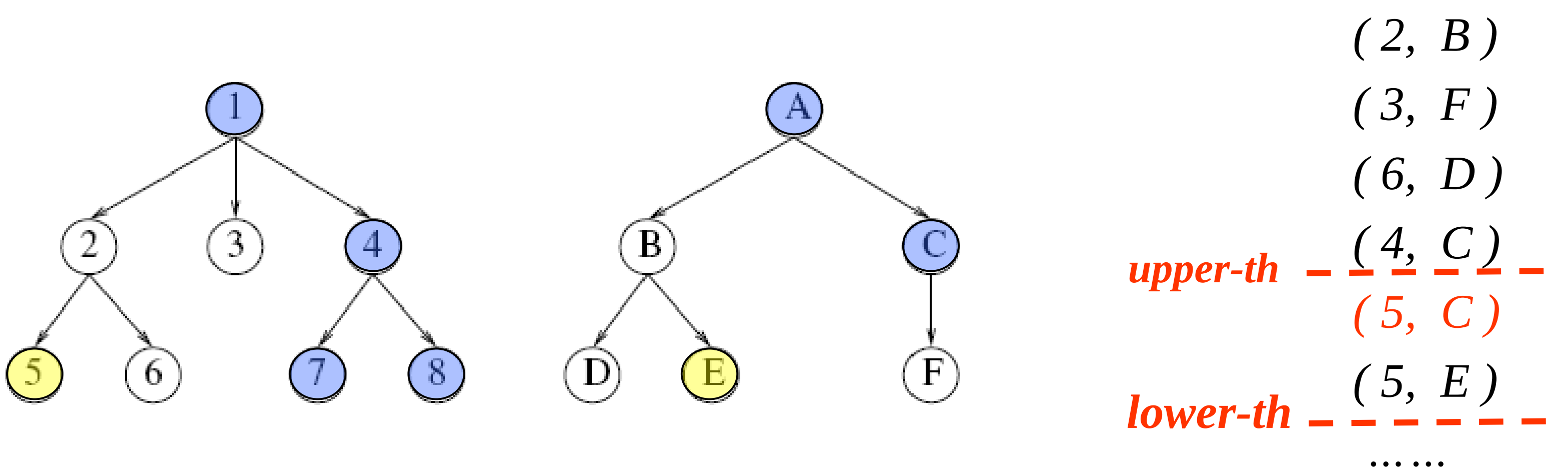


Filtering Techniques

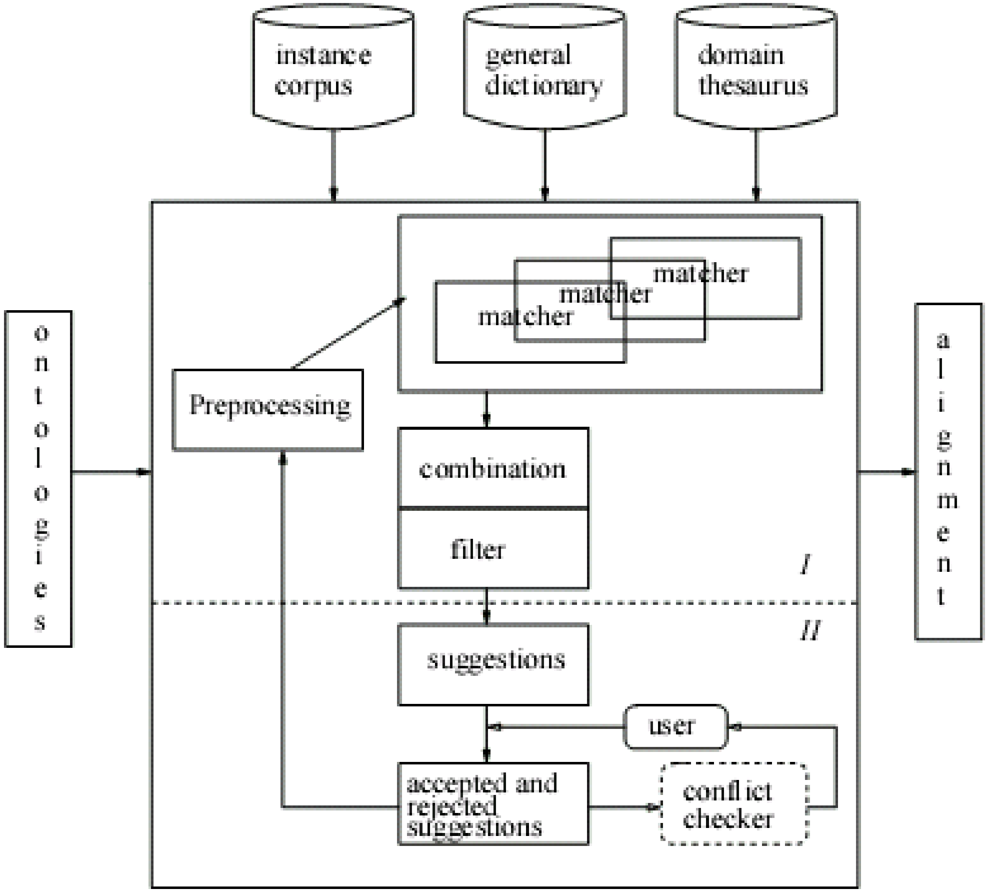
- Double threshold filtering

  (1) Pairs of concepts with similarity higher than or equal to **upper** threshold are alignment suggestions

  (2) Pairs of concepts with similarity between **lower** and **upper** thresholds are alignment suggestions if they make sense with respect to the structure of the ontologies and the suggestions according to (1)



An Alignment Framework



Ontology Alignment and Merging Systems

	linguistic	structure	constraints	instances	auxiliary
ArtGen	name	parents, children		domain specific documents	WordNet
ASCO	name, label description	parents, children, siblings, path from root			WordNet
Chimaera	name	parents, children			
FCA-Merge	name			domain specific documents	
FOAM	name, label	parents, children	equivalence		
GLUE	name	neighborhood		instances	
HCONE	name	parents, children			WordNet
IF-Map				instances	a reference ontology
iMapper		leaf, non-leaf, children, related node	domain, range	instances	WordNet
OntoMapper		parents, children		documents	
(Anchor-) PROMPT	name	direct graphs			
SAMBO	name, synonym	is-a and part-of, descendants and ancestors		domain specific documents	WordNet, UMLS
S-Match	label	path from root	semantic relations codified in labels		WordNet

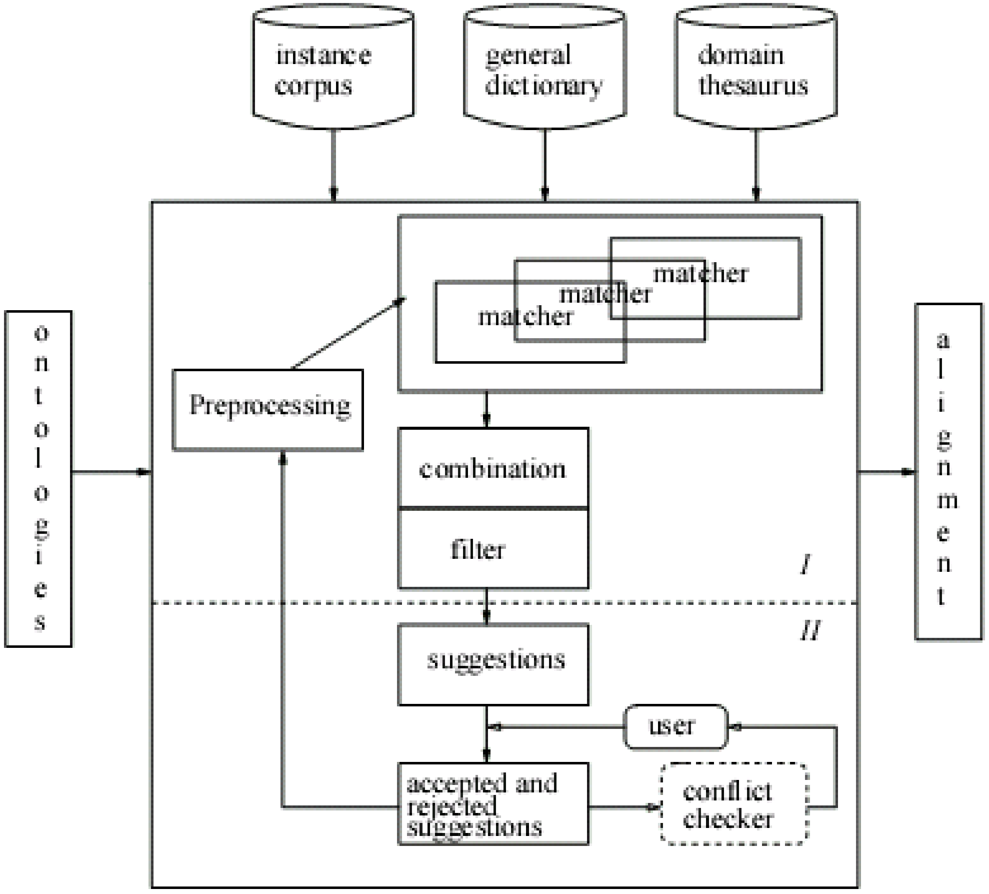
2006

OAEI Anatomy Track 2007-2016*

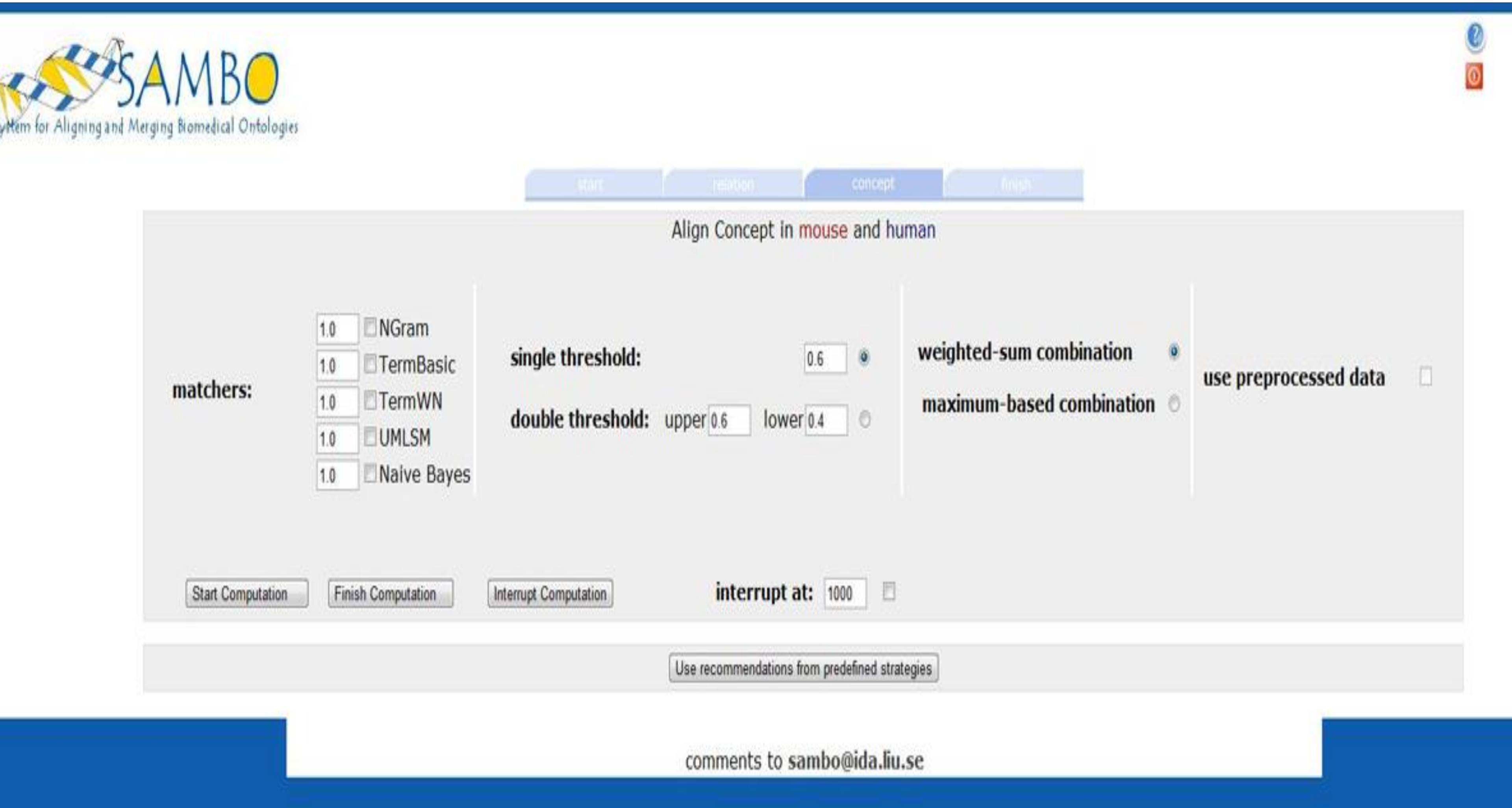
- Components
  - Almost all systems implement preprocessing, matchers, combination, filtering components
  - Debugging component and GUI rarely implemented
- Matching strategies:
  - Variety of string-based strategies
  - Most often string and structured-based strategies
- Use of background knowledge
  - Almost all systems use sources of background knowledge

* https://doi.org/10.1186/s13326-017-0166-5

An Alignment Framework

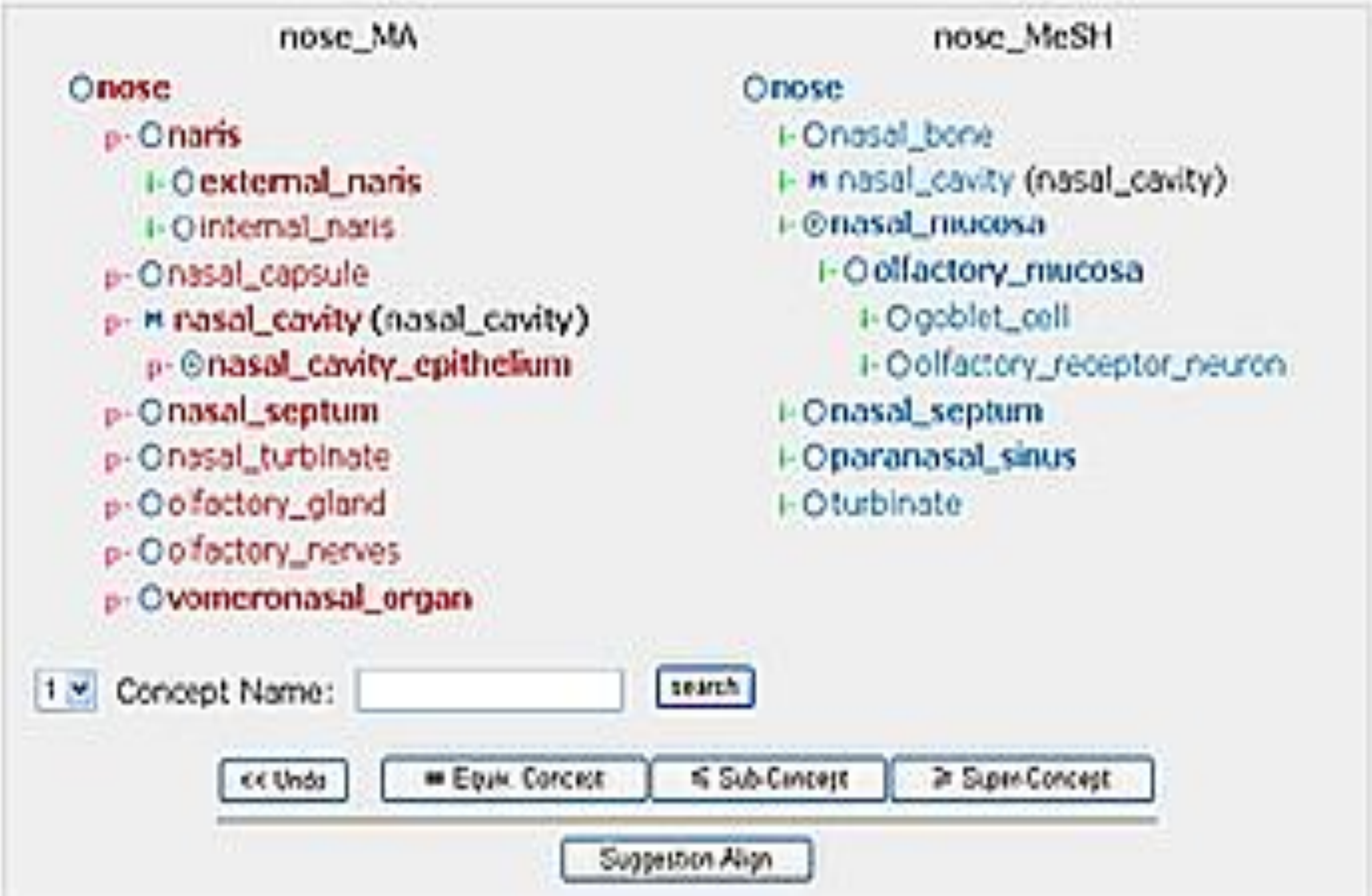


Example: Alignment System SAMBO – matchers, combination, filter



Example: Alignment System SAMBO – suggestion mode

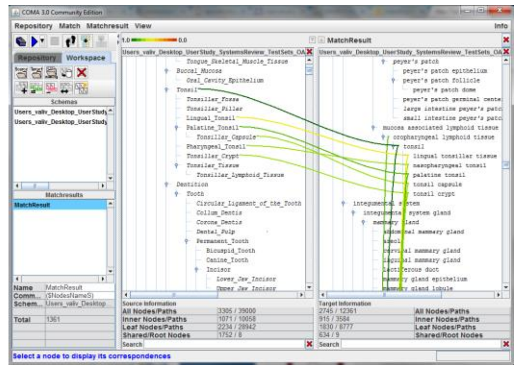
Example: Alignment System SAMBO – manual mode

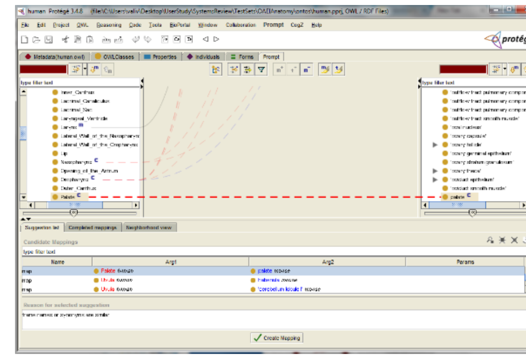


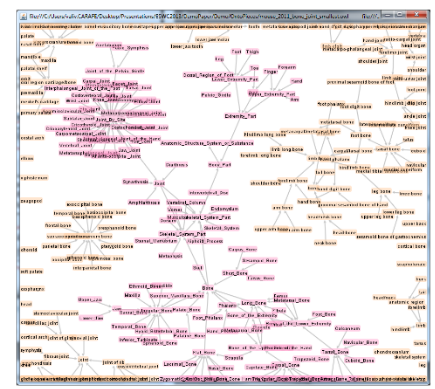
Ontology Alignment User Tasks

Navigation & exploration

- ◇Panning
- ◇Zooming
- ◇Scrolling
- ◇Expanding
- ◇Collapsing



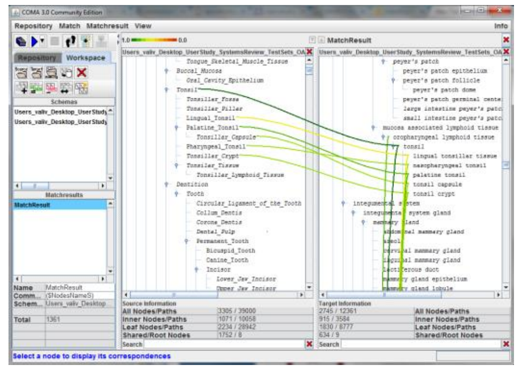


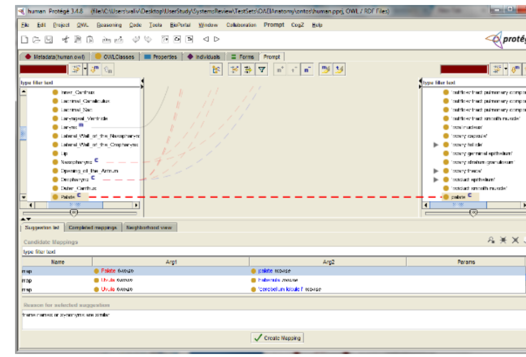


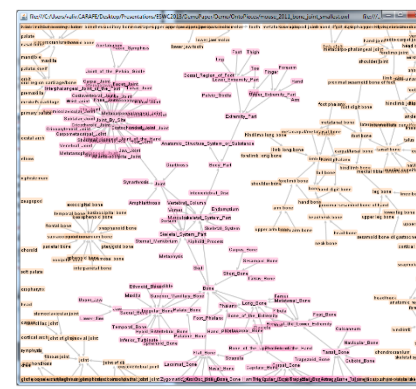
Ontology Alignment User Tasks

Align

- ⟐ Validating
- ⟐ Evaluating
- ⟐ Revising
- ⟐ Debugging
- ⟐ Sharing







Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

Ontology Alignment Evaluation Initiative (OAEI)

- Since 2004
- Evaluation of systems
- Different tracks (2017)
  - anatomy, conference, large biomedical ontologies, disease and phenotype
  - multilingual: multifarm (9 languages)
  - process model
  - interactive
  - instance
  - link discovery for spatial data

Evaluating Ontology Alignments

Results for OAEI 2016 - Anatomy track

In the following, we analyze all participating systems that could generate an alignment. The listing comprises of 13 entries. As previous years some of the s
different versions. LogMap participated with LogMap, LogMapBio and a lightweight version LogMapLite that uses only some core components. Similarly
with two versions, DKP-AOM and DKP-AOM-Lite. There are a number of systems which participate in the anatomy track for the first time. These are Ali
LYAM. While every year there are several new participants in this track, there are also systems participating for several years in a row. LogMap is a consta
and XMap joined the track in 2013. DKP-AOM, Lily and CroMatcher participate for the second year in a row in this track. Lily participated in the track ba
participated in 2013 but did not produce an alignment within the given time frame back then. For more details, we refer the reader to the papers presenting
have 10 different systems (not counting different versions) which generated an alignment.

Matcher	Runtime	Size	Precision	F-Measure	Recall	Recall+	Coherent
AML	47	1493	0.95	0.943	0.936	0.832	+
LPHOM	1601	1555	0.709	0.718	0.727	0.497	-
Lily	272	1382	0.87	0.83	0.794	0.515	-
DKP-AOM	379	207	0.99	0.238	0.135	0.0	+
XMap	45	1413	0.929	0.896	0.865	0.647	+
DKP-AOM-Lite	372	207	0.99	0.238	0.135	0.0	+
FCA_Map	117	1361	0.932	0.882	0.837	0.578	-
Alin	306	510	0.996	0.501	0.335	0.0	+
LogMap	24	1397	0.918	0.88	0.846	0.593	+
LogMapBio	758	1531	0.888	0.892	0.896	0.728	+
LYAM	799	1539	0.863	0.869	0.876	0.682	-
LogMapLite	20	1147	0.962	0.828	0.728	0.288	-
CroMatcher	573	1442	0.949	0.925	0.902	0.773	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-

Unlike the last two editions of the track when 6 systems genereated an alignment in less than 100 seconds, this year only 4 of them were able to complete the alignment task in this time frame. These are AML, XMap, LogMap and LogMapLite. Similarly to the last 4 years LogMapLite has the shortest runtime, followed by LogMap, XMap and AML. Depending on the

Evaluation Measures

- Precision:

$$\frac{\#\ \text{correct mapping suggestions}}{\#\ \text{mapping suggestions}}$$

- Recall:

$$\frac{\#\ \text{correct mapping suggestions}}{\#\ \text{correct mappings}}$$

- F-measure: combination of precision and recall

Evaluating Ontology Alignments

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

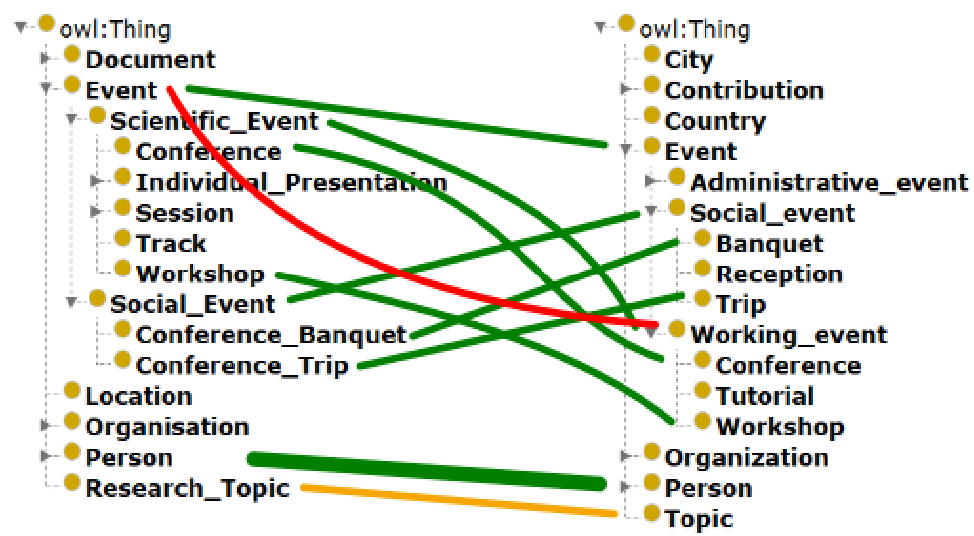
http://oaei.ontologymatching.org/2007/results/anatomy/

	Reference	AOAS	Sambo	ASMOV	RiMOM	LabelEq.
Reference	-	0.756	0.687	0.598	0.316	0.600
AOAS	0.756	-	0.738	0.679	0.308	0.706
Sambo	0.687	0.738	-	0.573	0.302	0.576
ASMOV	0.598	0.679	0.573	-	0.282	0.681
RiMOM	0.316	0.308	0.302	0.282	-	0.277
LabelEq.	0.600	0.706	0.576	0.681	0.277	-
Falcon-AO	0.578	0.657	0.562	0.620	0.272	0.814
TaxoMap	0.464	0.484	0.443	0.461	0.264	0.515
AgreementM.	0.423	0.435	0.401	0.408	0.252	0.444
Prior+	0.420	0.416	0.407	0.376	0.258	0.404
Lily	0.349	0.334	0.326	0.297	0.224	0.308
X-Som	0.205	0.239	0.206	0.232	0.089	0.333
DSSim	0.109	0.115	0.103	0.106	0.065	0.121

Evaluating Ontology Alignments

Precision & recall do not show:

- Often & rarely found mappings
- Correct & incorrect mappings
- (Dis)Agreement between tools
- Regions with the same or different number of mappings
- Fine-grained evaluation
- …



Evaluating Ontology Alignments

In practice:

- Write custom scripts
  - Error-prone
  - New question = new script
  - …
- Explore their results
  - in (sortable) tables …

How to write queries about unexpected findings?

List all mappings computed by AML in the Anatomy track of the OAEI?

```
SELECT ma.tool, ma.year, cl1.label, cl2.label, ma.relation,
ma.measure,  m.relation, m.trivial, m.reference
FROM oaei_anatomy_track.matchingsInAlignment ma
INNER JOIN conceptLabels cl1 on ma.source = cl1.name
INNER JOIN conceptLabels cl2 on ma.target = cl2.name
INNER JOIN matching m on m.source=ma.source and m.target=ma.target
where tool ='AML';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

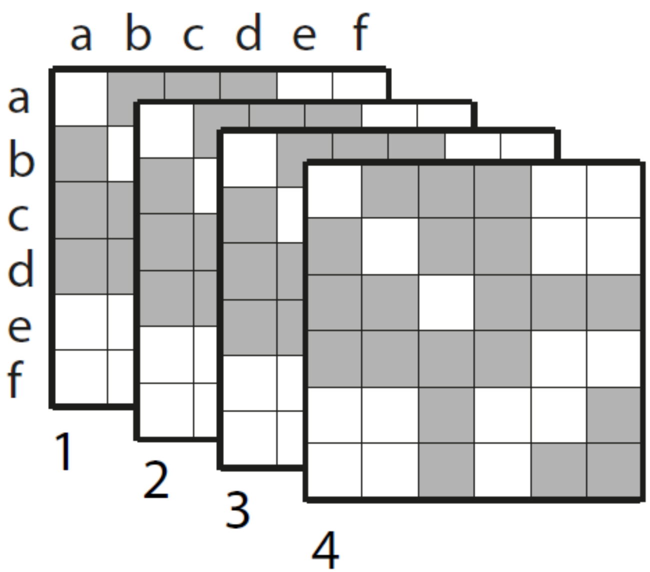
tool	year	label	label	relation	measure	relation	trivial	reference
AML	2013	corpus callosum	Corpus_Callosum	=	0.883599997	=	1	1
AML	2013	cochlea	Cochlea	=	0.883599997	=	1	1
AML	2013	stomach lesser curvature	Lesser_Curvature	=	0.744212985	=	0	1
AML	2013	intercostales internus	Internal_Intercostal_Muscle	=	0.789600015	=	0	0
AML	2013	lateral ventricle ependyma	Lateral_Ventricle_Ependyma	=	0.883599997	=	1	1
AML	2013	mylohyoid	Mylohyoid	=	0.883599997	=	1	1
AML	2013	right lung alveolus	Right_Lung_Alveolus	=	0.971960008	=	1	1
AML	2013	rib 7	Rib_7	=	0.780200005	=	1	1
AML	2013	hand digit 4 phalanx	Hand_Digit_4_Phalanx	=	0.971960008	=	1	1
AML	2013	sphincter pupillae	Sphincter_Pupillae_Muscle	=	0.783222914	=	0	1
AML	2013	hand digit skin	Hand_Digit_Skin	=	0.883599997	=	1	1
AML	2013	urinary bladder wall	Bladder_Wall	=	0.653980553	=	0	1
AML	2013	iliacus	Iliacus	=	0.883599997	=	1	1
AML	2013	thyroid gland follicle	Thyroid_Gland_Follicle	=	0.883599997	=	1	1
AML	2013	metatarsal bone digit 1	Metatarsal_Bone_Digit_1	=	1.000000000	=	1	1
AML	2013	pectoralis minor	Pectoralis_Minor	=	0.971960008	=	1	1
AML	2013	retina layer	Retina_Layer	=	0.883599997	=	1	1
AML	2013	lacrimal sac	Lacrimal_Sac	=	0.883599997	=	1	1
AML	2013	hand digit 5	Hand_Digit_5	=	0.883599997	=	1	1

Evaluation Use Cases

- Selecting, combining & fine tuning tools
- Matchers development
- Ontology alignment evolution
- Validating & debugging of ontology alignments and RA
- Collaborative ontology alignment

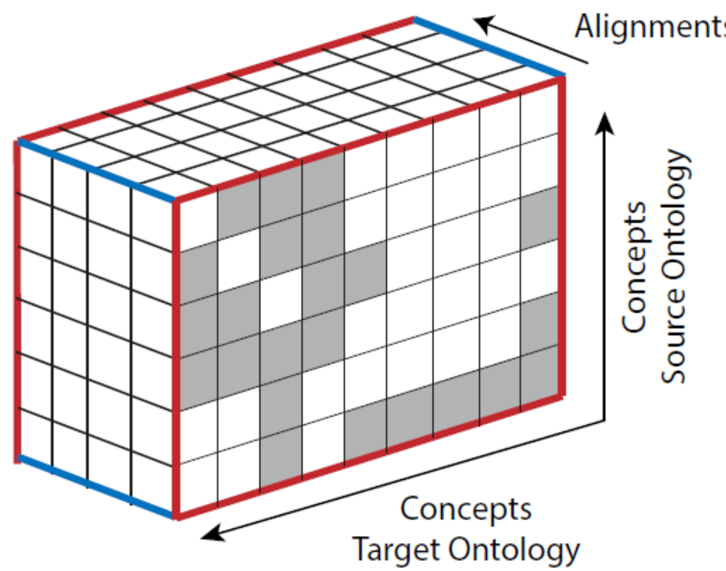
Alignment Cubes

Alignment Cubes

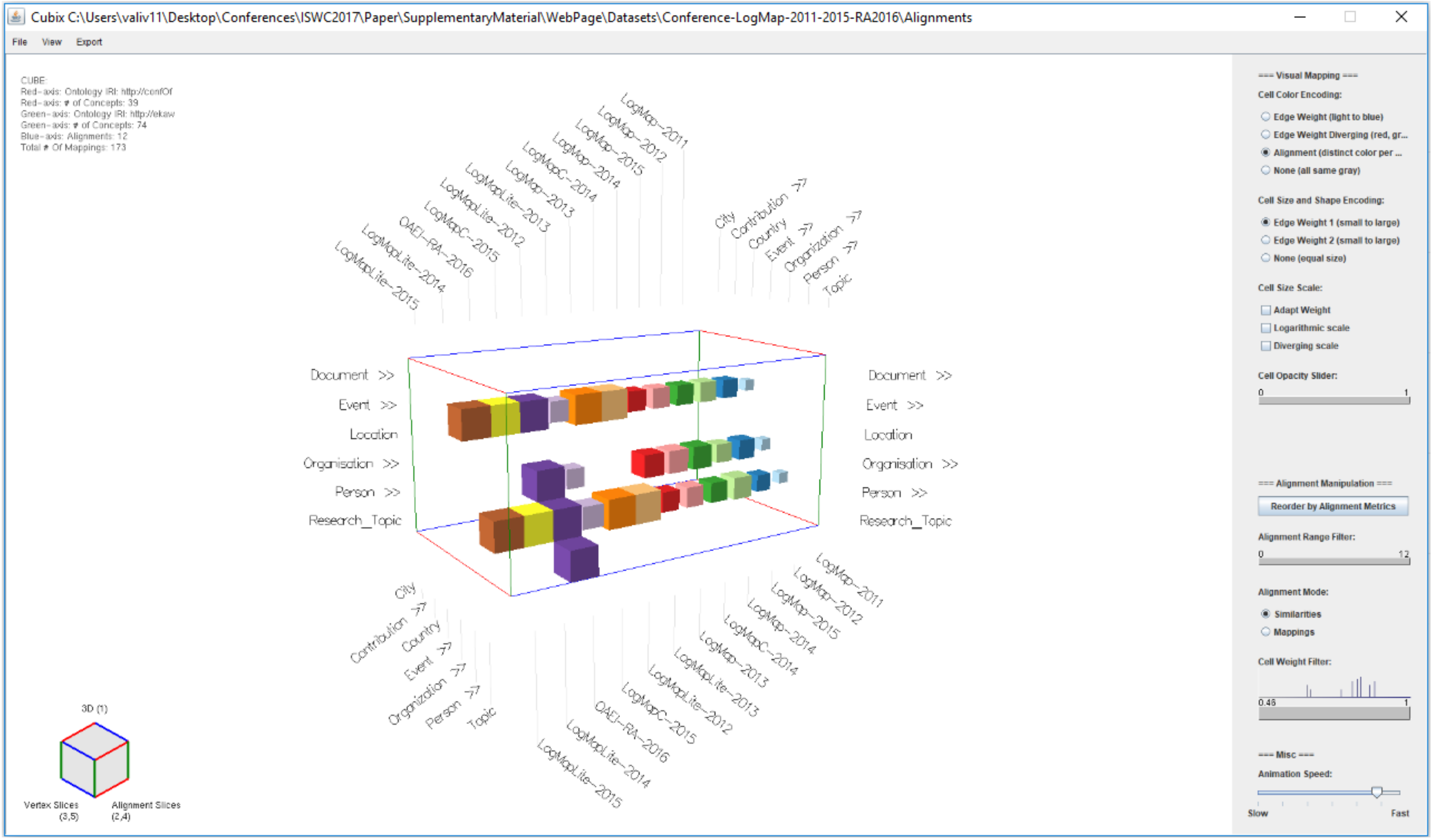


An alignment is represented by a matrix

Stacking several matrices together



Alignment Cubes



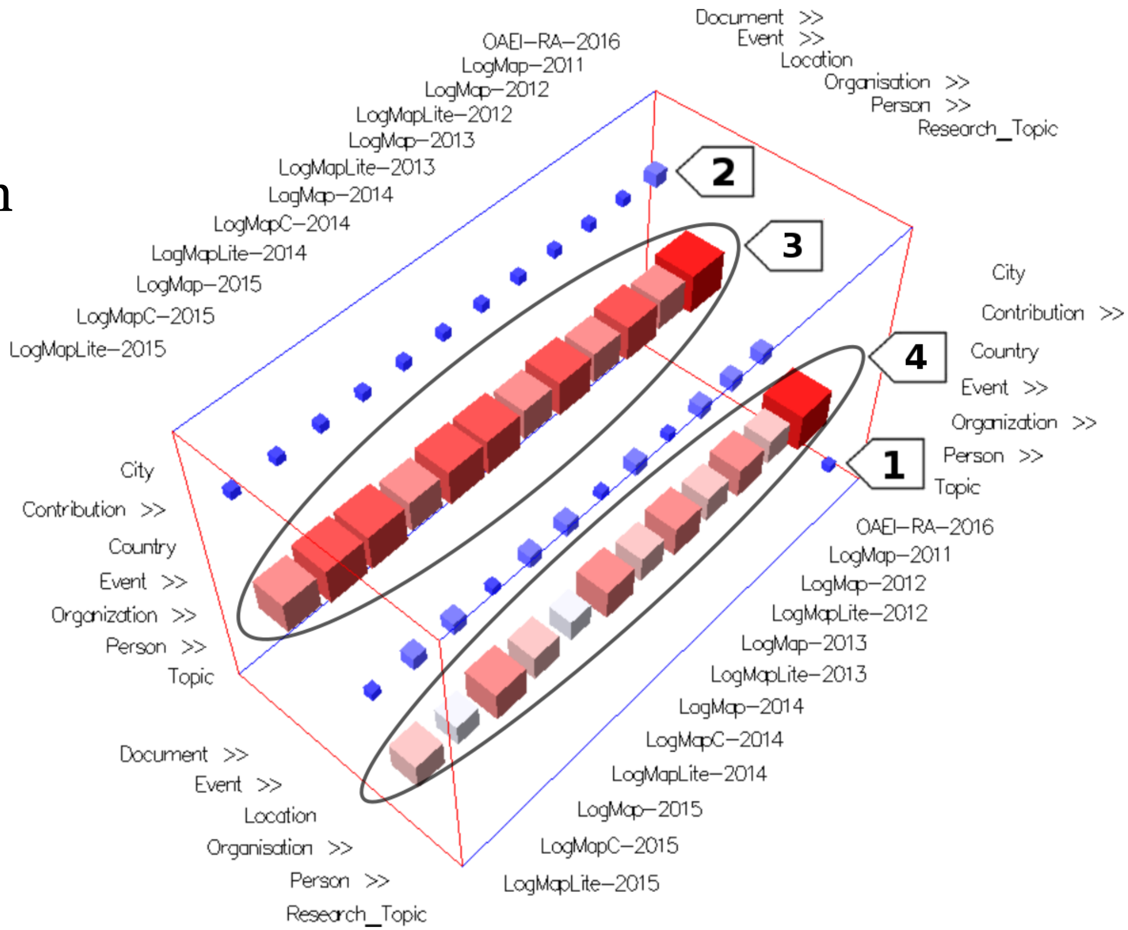
Alignment Cubes

3D view:

- drives *initial exploration*
- supports *identifying regions of interest*
- can possibly yield some *high-level insights*

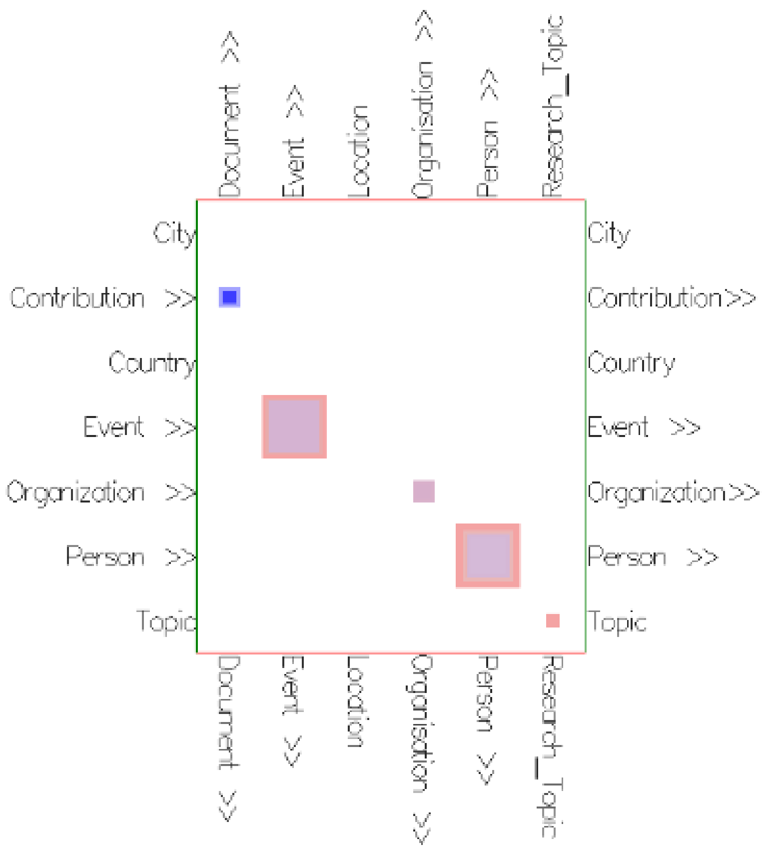
Paired aggregated & detailed side-by-side views:

- Alignment topology
- Concepts network

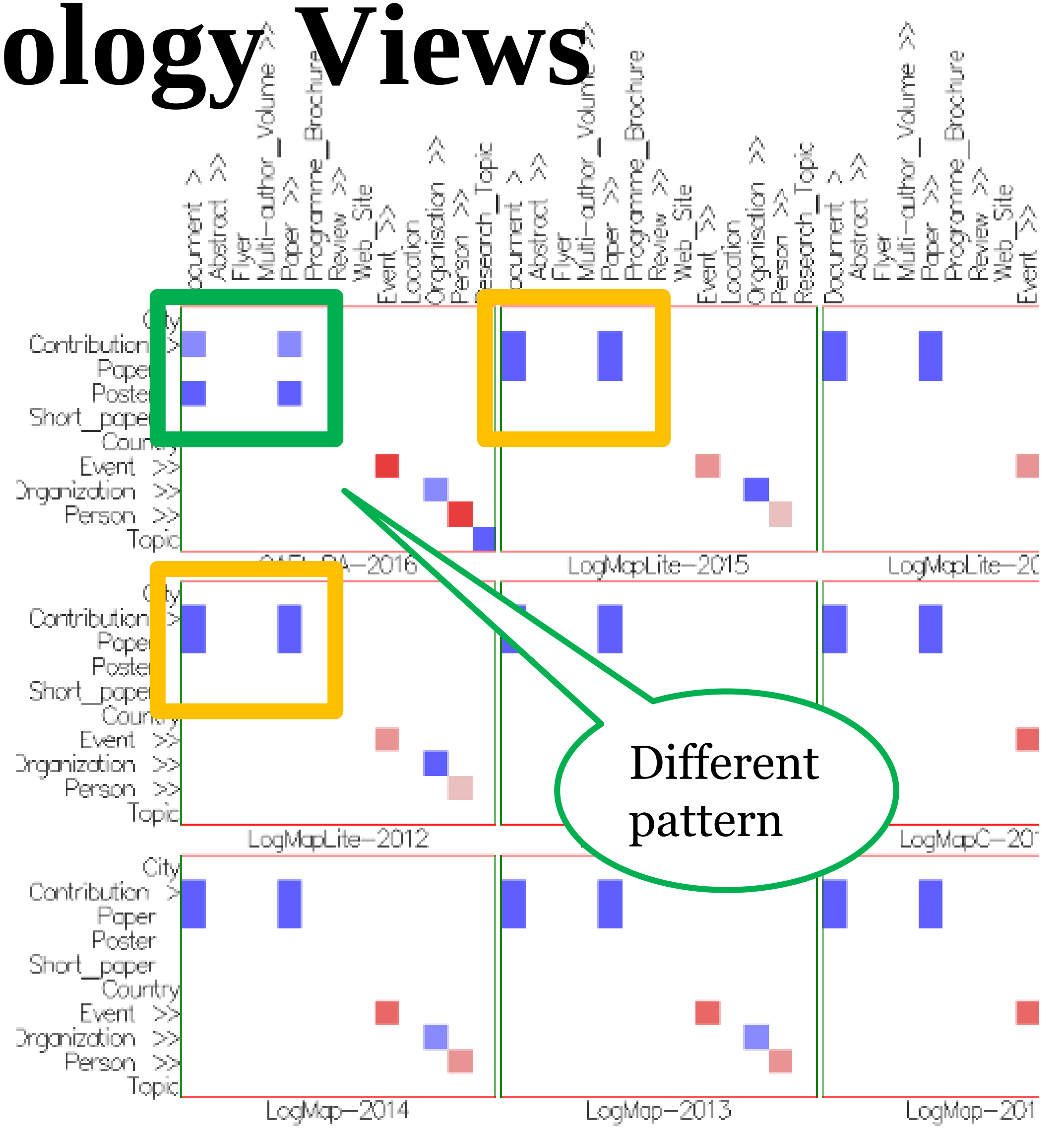


Alignment Topology Views

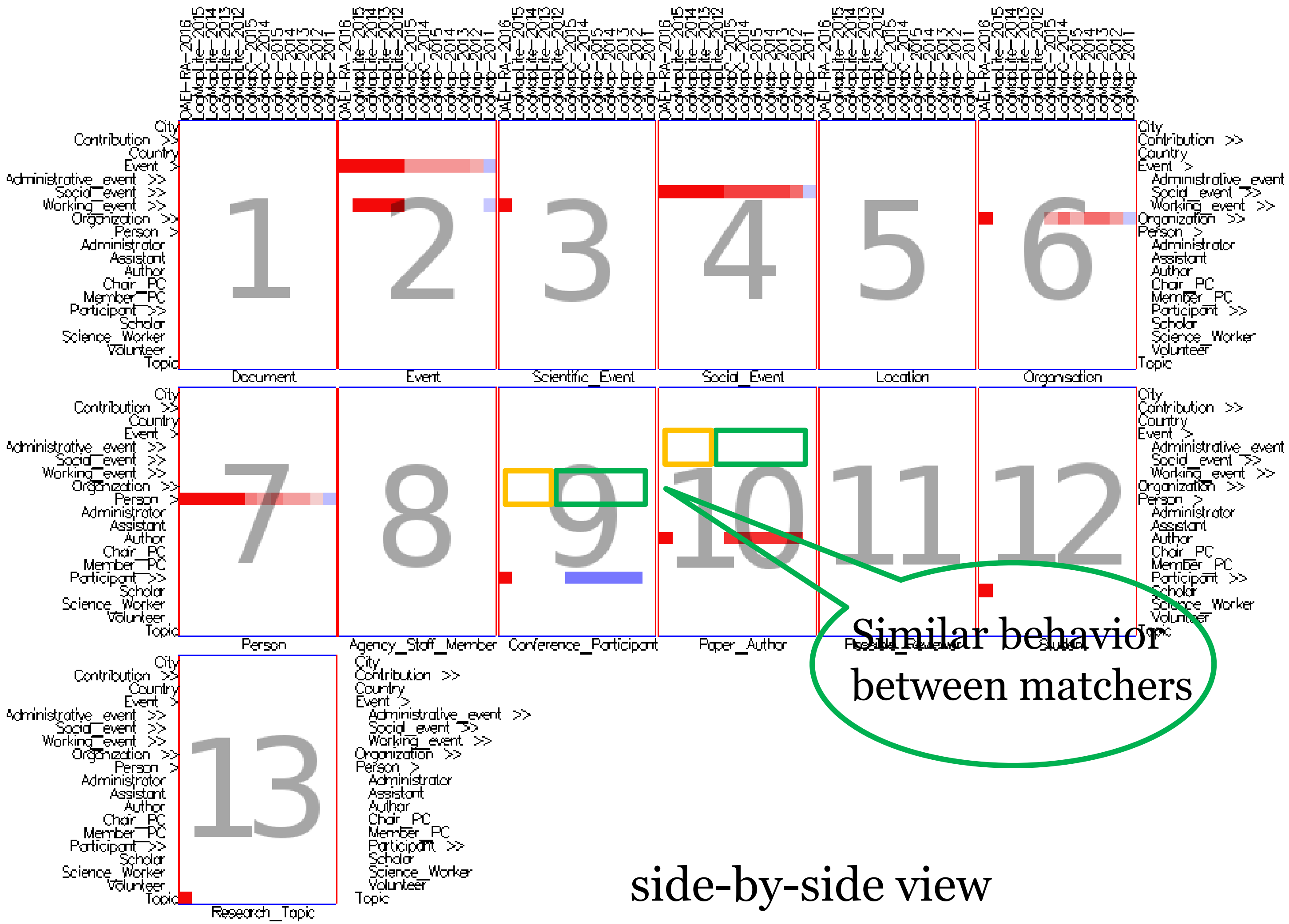
aggregated view



side-by-side view

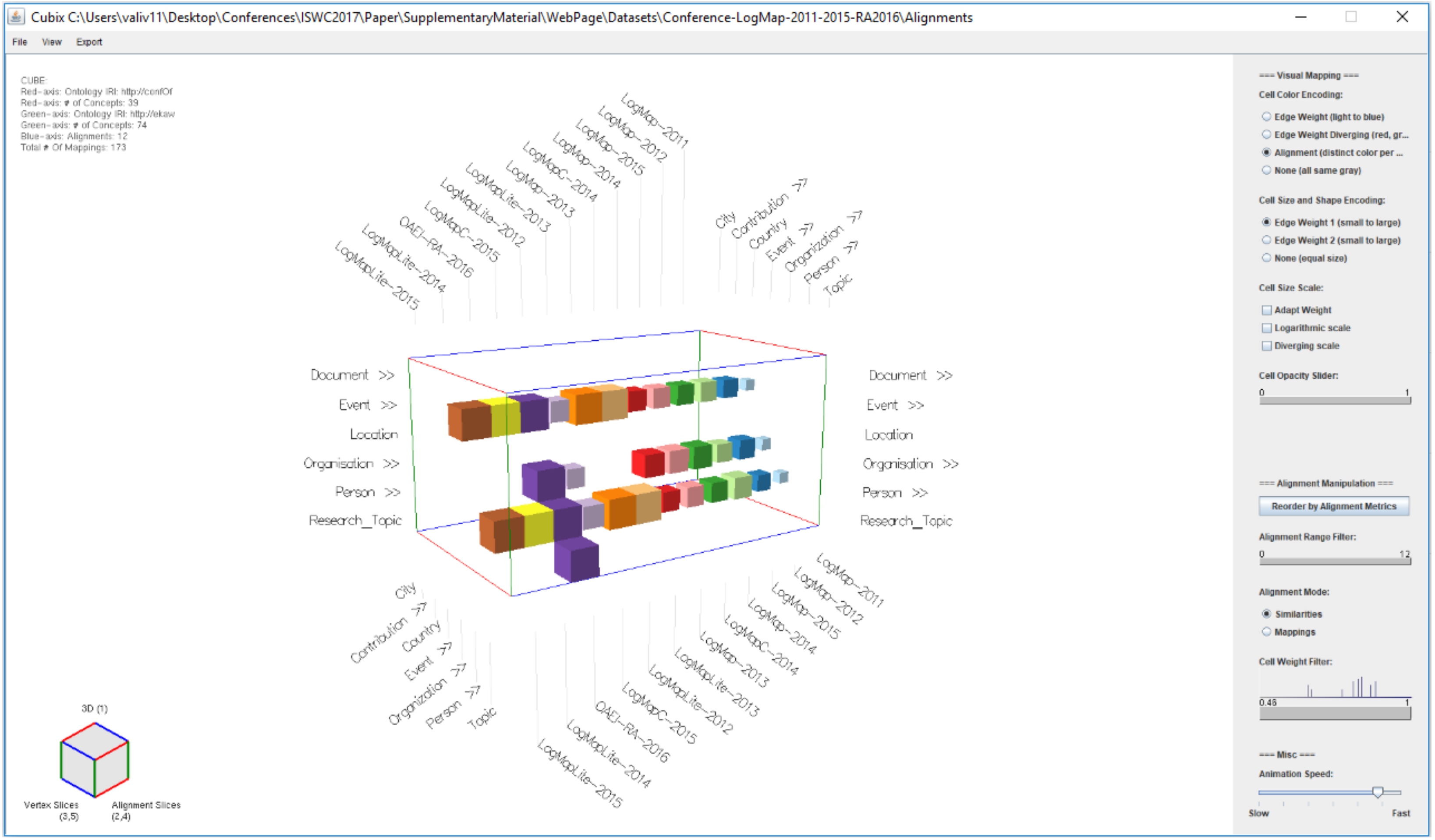


Concepts Network Views



side-by-side view

Alignment Cubes



http://www.ida.liu.se/~patla00/research/AlignmentCubes/

Ontology Alignment

- Ontology alignment
- Ontology alignment strategies
- Evaluation of ontology alignment strategies
- Ontology alignment challenges

Challenges

- Algorithmic approaches
  - Large-scale matching evaluation
  - Efficiency of matching techniques
  - Matching with background knowledge
  - Matcher selection, combination and tuning
- User-engaging approaches
  - User involvement
  - Explanation of matching results
  - Social and collaborative matching
  - Alignment management: infrastructure and support

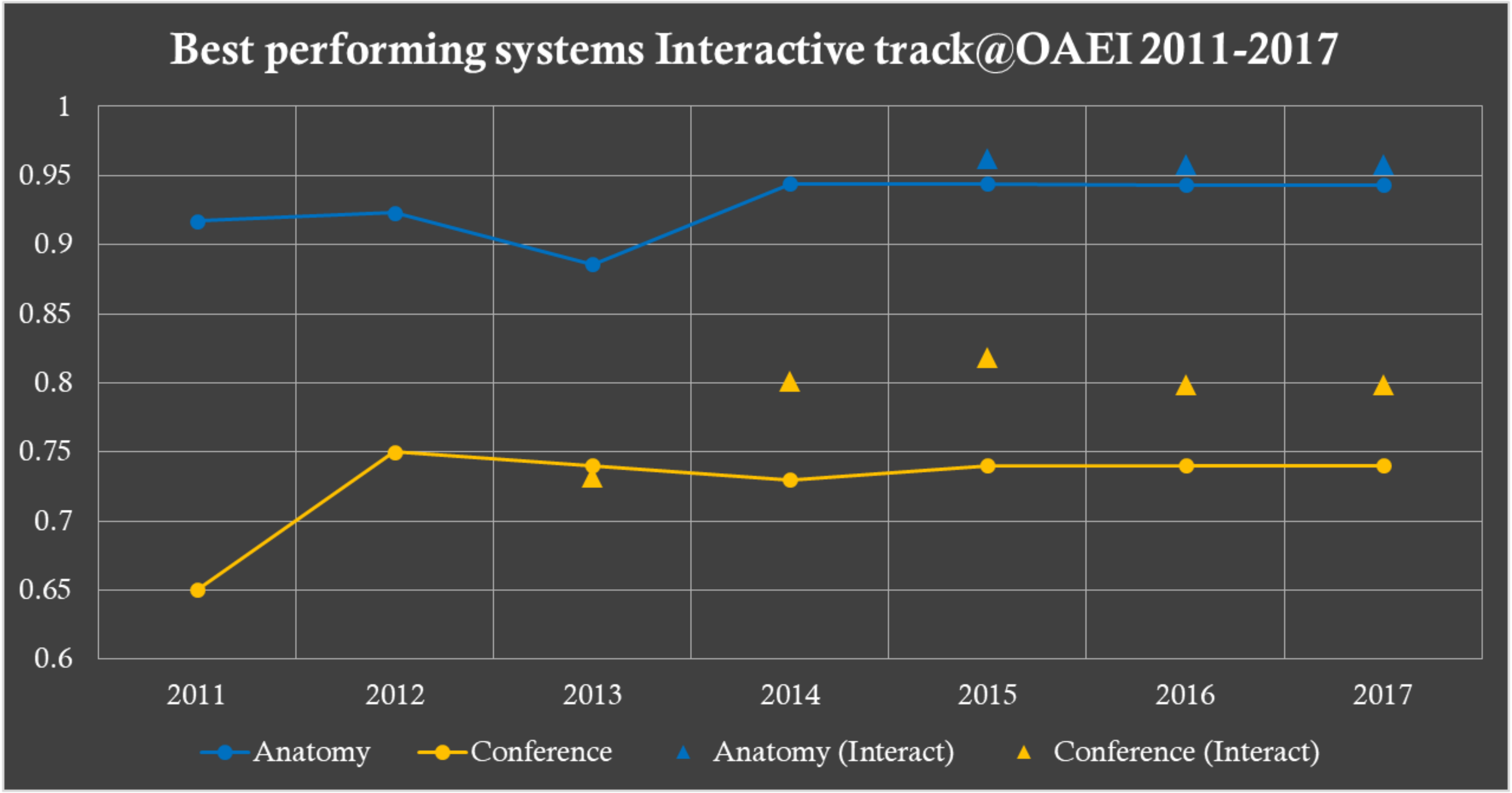
Challenges

- Large-scale matching evaluation
- Efficiency of matching techniques
  - parallellization
  - distribution of computation
  - approximation of matching results (not complete)
  - modularization of ontologies
  - optimization of matching methods

Challenges

- Matching with background knowledge
  - partial alignments
  - reuse of previous matches
  - use of domain-specific corpora
  - use of domain-specific ontologies
- Matcher selection, combination and tuning
  - recommendation of algorithms and settings

Interactive Ontology Alignment



Challenges

- User involvement
  - visualization
  - user feedback
- Explanation of matching results
- Social and collaborative matching
- Alignment management: infrastructure and support

Further Reading

Starting points for further studies

Further Reading Ontology Alignment

- http://www.ontologymatching.org
(plenty of references to articles and systems)

- Ontology alignment evaluation initiative: http://oaei.ontologymatching.org
(home page of the initiative)

- Euzenat, Shvaiko, *Ontology Matching*, Springer, 2007.

.

Further Reading
OAEI – Experience Papers

- Dragisic Z, Ivanova V, Li H, Lambrix P, **Experiences from the Anatomy track in the Ontology Alignment Evaluation Initiative.** *J Biomedical Semantics* 8:56, 2017.
- Shvaiko P, Euzenat J., **Ontology matching: state of the art and future challenges**. *IEEE Trans Knowl Data Eng.* 25(1):158–76, 2013.
- Euzenat J, Meilicke C, Stuckenschmidt H, Shvaiko P, Trojahn C, **Ontology alignment evaluation initiative: Six years of experience.** *J Data Semant.* XV:158–92, 2011.

- Cheatham M, Hitzler P, **String similarity metrics for ontology alignment.** In:*The Semantic Web - ISWC 2013, 12th International Semantic Web Conference, LNCS 8219,* 294–309, 2013.

Further Reading Ontology Alignment

Systems at LiU / IDA / ADIT

- Lambrix, Tan, SAMBO – a system for aligning and merging biomedical ontologies, *Journal of Web Semantics*, 4(3):196-206, 2006.

(description of the SAMBO tool and overview of evaluations of different matchers)

- Lambrix, Tan, A tool for evaluating ontology alignment strategies, *Journal on Data Semantics*, VIII:182-202, 2007.

(description of the KitAMO tool for evaluating matchers)

- Lambrix P, Kaliyaperumal R, A Session-based Ontology Alignment Approach enabling User Involvement, *Semantic Web Journal* 8(2):225-251, 2017.

- Ivanova V, Bach B, Pietriga E, Lambrix P, Alignment Cubes: Towards Interactive Visual Exploration and Evaluation of Multiple Ontology Alignments, 16th International Semantic Web Conference, 400-417, 2017.

Further Reading Ontology Alignment

- Chen, Tan, Lambrix, Structure-based filtering for ontology alignment,*IEEE WETICE workshop on semantic technologies in collaborative applications*, 364-369, 2006.

(double threshold filtering technique)

- Tan, Lambrix, A method for recommending ontology alignment strategies, *International Semantic Web Conference*, 494-507, 2007.

  Ehrig, Staab, Sure, Bootstrapping ontology alignment methods with APFEL, *International Semantic Web Conference*, 186-200, 2005.

  Mochol, Jentzsch, Euzenat, Applying an analytic method for matching approach selection, *International Workshop on Ontology Matching*, 2006.

(recommendation of alignment strategies)

- Lambrix, Liu, Using partial reference alignments to align ontologies, *European Semantic Web Conference*, 188-202, 2009.

(use of partial alignments in ontology alignment)

Ontology Debugging and Completion

Defects in Ontologies

- Syntactic defects
  - E.g. wrong tags or incorrect format
- Semantic defects
  - E.g. unsatisfiable concepts, incoherent and inconsistent ontologies
- Modeling defects
  - E.g. wrong or missing relations

Example: Incoherent Ontology

- Example: DICE ontology
  - Brain ⊑ CentralNervousSystem ⊓ BodyPart ⊓ ∃systempart.NervousSystem ⊓ ∃ region.HeadAndNeck ⊓ ∀region.HeadAndNeck

    A brain is a central nervous system and a body part which has a system part that is a nervous system and that is in the head and neck region.

  - CentralNervousSystem ⊑ NervousSystem

    A central nervous system is a nervous system.

  - BodyPart ⊑¬NervousSystem

    Nothing can be at the same time a body part and a nervous system.

Slide from G. Qi

Example: Inconsistent Ontology

- Example from Foaf:
  - Person(timbl)
  - Homepage(timbl, http://w3.org/)
  - Homepage(w3c, http://w3.org/)
  - Organization(w3c)
  - InverseFunctionalProperty(Homepage)
  - DisjointWith(Organization, Person)
- Example from OpenCyc:
  - ArtifactualFeatureType(PopulatedPlace)
  - ExistingStuffType(PopulatedPlace)
  - DisjointWith(ExistingObjectType,ExistingStuffType)
  - ArtifactualFeatureType $\sqsubseteq$ ExistingObjectType

Slide from G. Qi

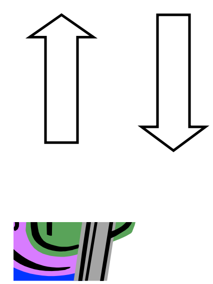
Example: Missing is-a Relations

- In 2008 Ontology Alignment Evaluation Initiative (OAEI) Anatomy track, task 4
  - Ontology MA : Adult Mouse Anatomy Dictionary (2744 concepts)
  - Ontology NCI-A : NCI Thesaurus - anatomy (3304 concepts)
  - 988 mappings between MA and NCI-A
    - 121 missing is-a relations in MA
    - 83 missing is-a relations in NCI-A

Influence of Missing Structure

- Ontology-based querying.

Medical Subject Headings (MeSH)



return 1617 articles

All MeSH Categories

- **Diseases Category**
  - **Eye Diseases**
    - **Scleral Diseases**
      - **Scleritis**

...

Influence of Missing Structure

- Incomplete results from ontology-based queries

Medical Subject Headings (MeSH)

return 1617 articles

return 695 articles

57% results are missed !

All MeSH Categories

- **Diseases Category**
  - **Eye Diseases**
    - **Scleral Diseases**
      - **Scleritis**

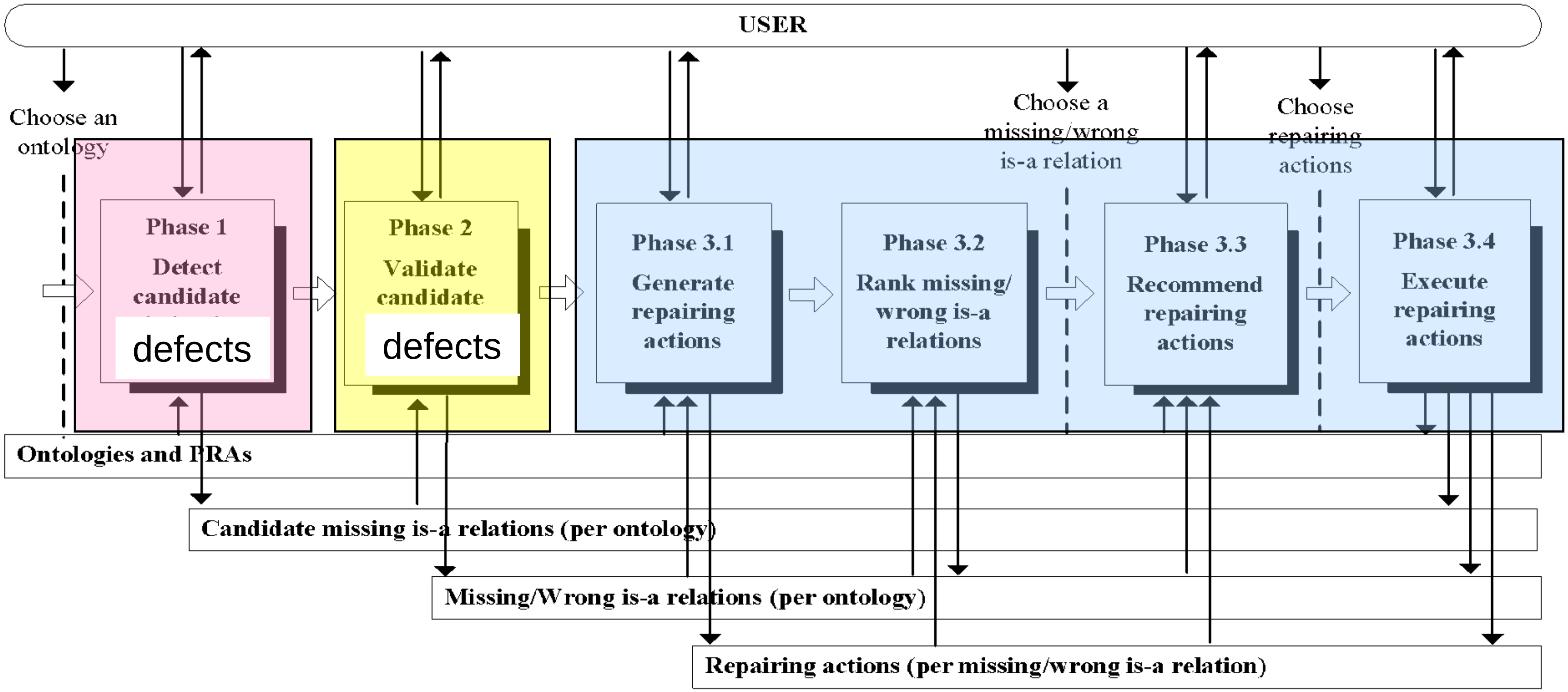
...

Defects in Ontologies and Ontology Networks

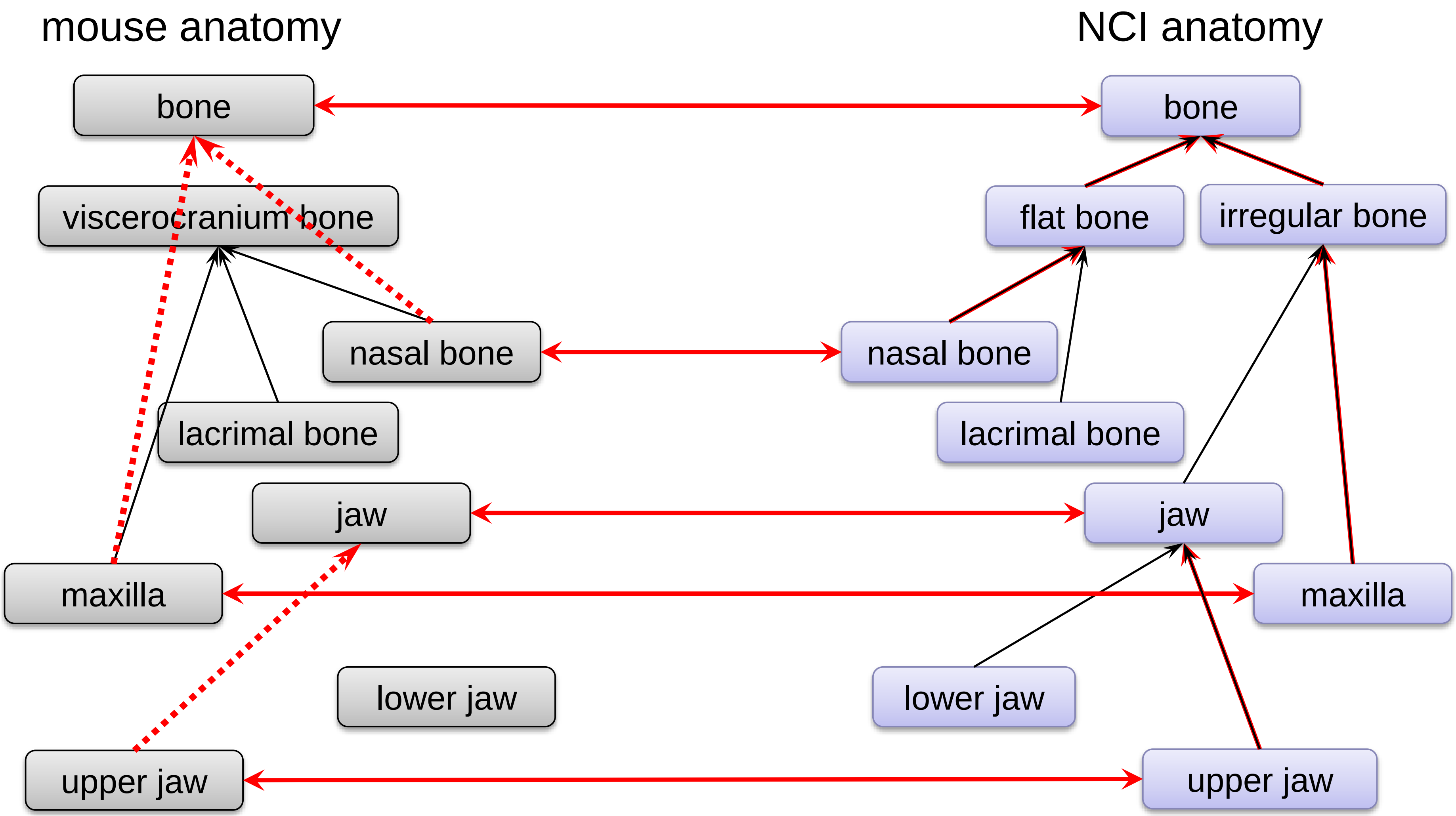
- Ontologies and ontology networks with defects, although often useful, also lead to problems when used in semantically-enabled applications.

→ Wrong conclusions may be derived or valid conclusions may be missed.

Overview of Debugging Approach



Example: Missing Structure



Debugging semantic defects

Example: Incoherent Ontology

Consider the following TBox $\mathcal{T}^*$, where A, B and C are primitive and $A_1, \ldots, A_7$ defined concept names:

$ax_1 : A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$	$ax_2 : A_2 \dot{\sqsubseteq} A \sqcap A_4$
$ax_3 : A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$	$ax_4 : A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$
$ax_5 : A_5 \dot{\sqsubseteq} \exists s.\neg B$	$ax_6 : A_6 \dot{\sqsubseteq} A_1 \sqcup \exists r.(A_3 \sqcap \neg C \sqcap A_4)$
$ax_7 : A_7 \dot{\sqsubseteq} A_4 \sqcap \exists s.\neg B$	

⇓

DL Reasoner ⇒ The ontology is incoherent!
The set of unsatisfiable concepts are : $\{A_1, A_3, A_6, A_7\}$.

What are the root causes of these defects?

Explain the Semantic Defects

- We need to identify the sets of axioms which are necessary for causing the logic contradictions.

$ax_1: A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap A_3$	$ax_2: A_2 \dot{\sqsubseteq} A \sqcap A_4$
$ax_3: A_3 \dot{\sqsubseteq} A_4 \sqcap A_5$	$ax_4: A_4 \dot{\sqsubseteq} \forall s.B \sqcap C$
$ax_5: A_5 \dot{\sqsubseteq} \exists s.\neg B$	$ax_6: A_6 \dot{\sqsubseteq} A_1 \sqcup \exists r.(A_3 \sqcap \neg C \sqcap A_4)$
$ax_7: A_7 \dot{\sqsubseteq} A_4 \sqcap \exists s.\neg B$	

- For example, for the unsatisfiable concept "A_1", there are two sets of axioms.

$ax_1: A_1 \dot{\sqsubseteq} \underline{\neg A} \sqcap A_2 \sqcap A_3$
$ax_2: A_2 \dot{\sqsubseteq} \underline{A} \sqcap A_4$

$ax_1: A_1 \dot{\sqsubseteq} \neg A \sqcap A_2 \sqcap \underline{A_3}$
$ax_3: A_3 \dot{\sqsubseteq} \underline{A_4 \sqcap A_5}$
$ax_4: A_4 \dot{\sqsubseteq} \underline{\forall s.B} \sqcap C$
$ax_5: A_5 \dot{\sqsubseteq} \underline{\exists s.\neg B}$

Minimal Unsatisfiability Preserving Sub-TBoxes (MUPS)

Definition 1 Let A be a concept which is unsatisfiable in a TBox $\mathcal{T}$. A set $\mathcal{T}' \subseteq \mathcal{T}$ is a *minimal unsatisfiability-preserving sub-TBox (MUPS)* of $\mathcal{T}$ if

- A is unsatisfiable in $\mathcal{T}'$, and
- A is satisfiable in every sub-TBox $\mathcal{T}'' \subset \mathcal{T}'$.

We will abbreviate the set of MUPS of $\mathcal{T}$ and A by $mups(\mathcal{T}, A)$.

$$mups(\mathcal{T}^*, A_1) = \{\{ax_1, ax_2\}, \{ax_1, ax_3, ax_4, ax_5\}\}$$

- The MUPS of an unsatisfiable concept imply the solutions for repairing.

  → Remove at least one axiom from each axiom set in the MUPS

Example

$$
\begin{aligned}
mups(\mathcal{T}^*, A_1) &= \{\{\cancel{ax_1}, ax_2\}, \{\cancel{ax_1}, \cancel{ax_3}, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_3) &= \{\{\cancel{ax_3}, \cancel{ax_4}, ax_5\}\} \\
mups(\mathcal{T}^*, A_6) &= \{\{\cancel{ax_1}, ax_2, \cancel{ax_4}, ax_6\}, \\
&\qquad \{\cancel{ax_1}, \cancel{ax_3}, \cancel{ax_4}, ax_5, ax_6\}\} \\
mups(\mathcal{T}^*, A_7) &= \{\{\cancel{ax_4}, ax_7\}\}
\end{aligned}
$$

- Possible ways of repairing all the unsatisfiable concepts in the ontology:

$$\{ax_1, ax_3, ax_4\}$$

How to represent all these possibilities?

Minimal Incoherence Preserving Sub-TBox (MIPS)

Definition 2 Let $\mathcal{T}$ be an incoherent TBox. A TBox $\mathcal{T}' \subseteq \mathcal{T}$ is a *minimal incoherence-preserving sub-TBox (MIPS) of* $\mathcal{T}$ if

- $\mathcal{T}'$ is incoherent, and
- every sub-TBox $\mathcal{T}'' \subset \mathcal{T}'$ is coherent.

$$
\begin{aligned}
mups(\mathcal{T}^*, A_1) &= \{\{ax_1, ax_2\}, \{ax_1, ax_3, ax_4, ax_5\}\} \\
mups(\mathcal{T}^*, A_3) &= \{\{ax_3, ax_4, ax_5\}\} \\
mups(\mathcal{T}^*, A_6) &= \{\{ax_1, ax_2, ax_4, ax_6\}, \\
&\quad \{ax_1, ax_3, ax_4, ax_5, ax_6\}\} \\
mups(\mathcal{T}^*, A_7) &= \{\{ax_4, ax_7\}\}
\end{aligned}
$$

We will abbreviate the set of MIPS of $\mathcal{T}$ by $mips(\mathcal{T})$. For $\mathcal{T}^*$ we get three MIPS:

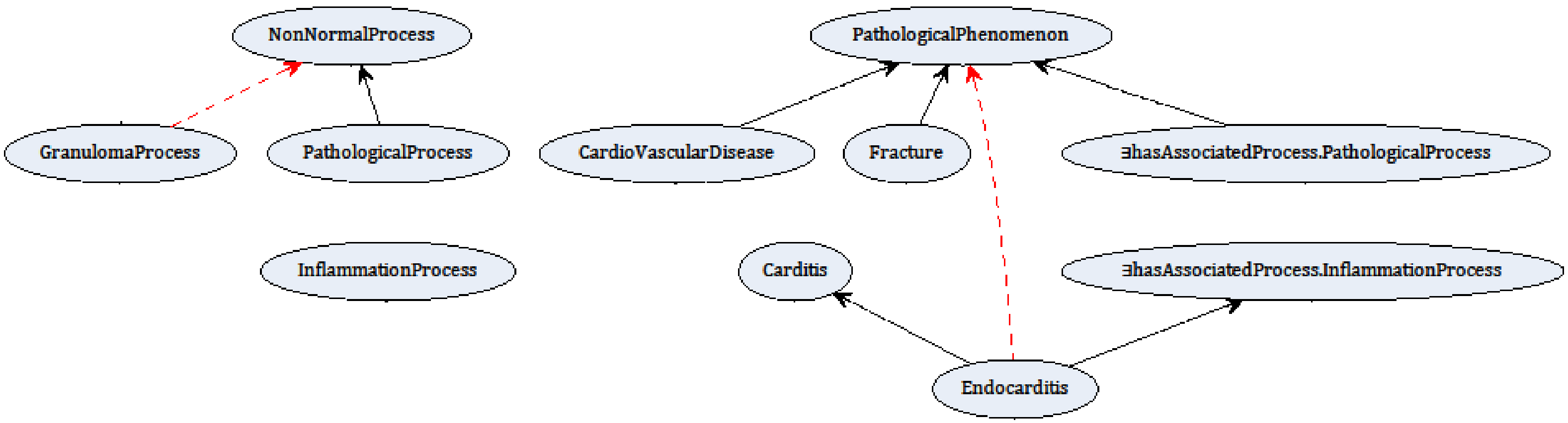
$$mips(\mathcal{T}^*) = \{\{ax_1, ax_2\}, \{ax_3, ax_4, ax_5\}, \{ax_4, ax_7\}\}$$

A possible repairing is $\{ax_i\} \cup \{ax_j\} \cup \{ax_k\}$, where

- $ax_i \in \{ax_1, ax_2\}$
- $ax_j \in \{ax_3, ax_4, ax_5\}$
- $ax_k \in \{ax_4, ax_7\}$

Completing the is-a structure of ontologies

Example



Repairing actions:

$\{\text{Endocarditis} \dot\sqsubseteq \text{PathologicalPhenomenon}, \text{GranulomaProcess} \dot\sqsubseteq \text{NonNormalProcess}\}$

$\{\text{Carditis} \dot\sqsubseteq \text{CardioVascularDisease}, \text{GranulomaProcess} \dot\sqsubseteq \text{PathologicalProcess}\}$

$\{\text{Carditis} \dot\sqsubseteq \text{Fracture}, \text{GranulomaProcess} \dot\sqsubseteq \text{NonNormalProcess}\}$

Description Logic EL

- Concepts

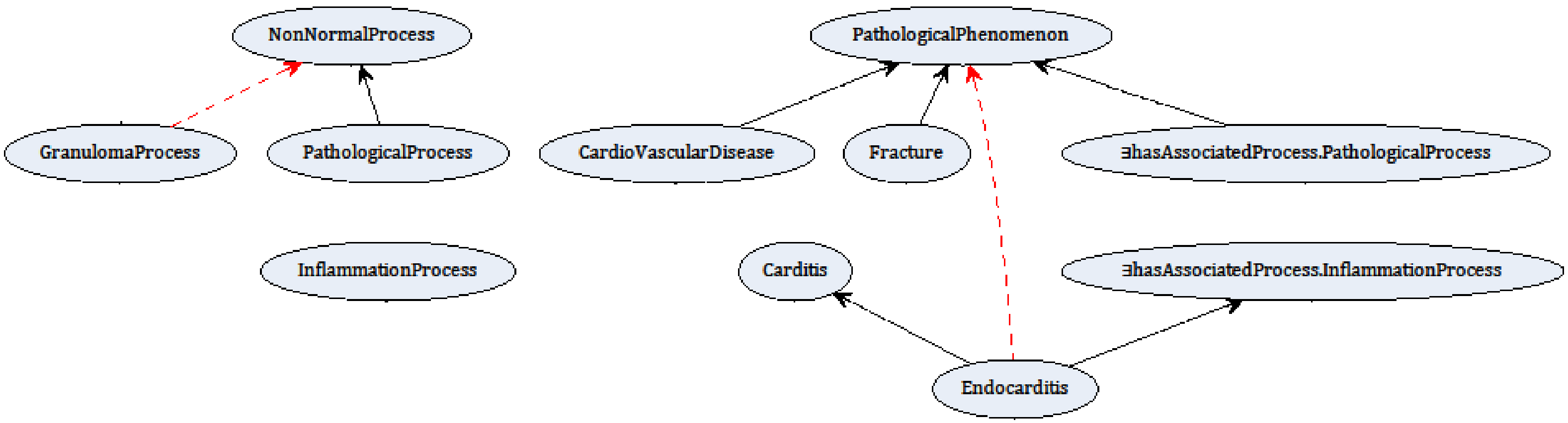
Atomic concept	A
Universal concept	$\top$
Intersection of concepts	$C \sqcap D$
Existential restriction	$\exists r.C$

- Terminological axioms: equivalence and subsumption

Generalized Tbox Abduction Problem – GTAP(T,C,Or,M)

- Given
  - **T**- a Tbox in EL
  - **C**- a set of atomic concepts in **T**
  - M = $\{A_i \sqsubseteq B_i\}_{i=1..n}$ and $\forall$ i:1..n: A_i, $B_i \in$ **C**
  - Or: $\{C_i \sqsubseteq D_i \mid C_i, D_i \in$ **C**$\} \rightarrow$ {true, false}
- Find
  - S = $\{E_i \sqsubseteq F_i\}_{i=1..k}$ such that
    $\forall$ i:1..k: E_i, $F_i \in$ **C** and Or($E_i \sqsubseteq F_i$) = true
    and T U S is consistent and T U S |= M

Example: GTAP



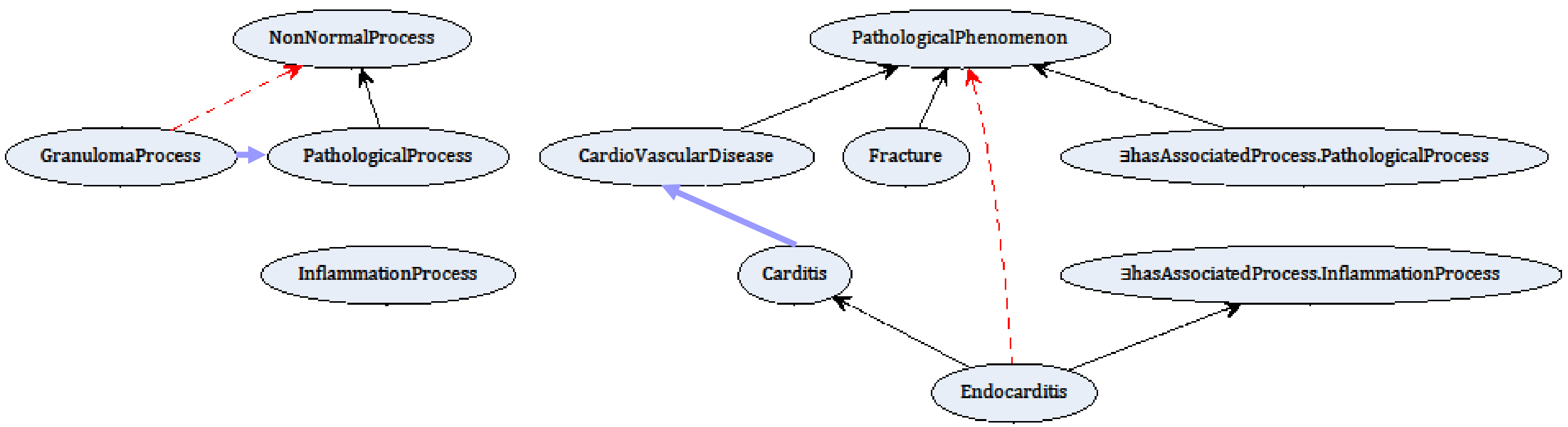
C = { GranulomaProcess, CardioVascularDisease, PathologicalPhenomenon, Fracture, Endocarditis, Carditis, InflammationProcess, PathologicalProcess, NonNormalProcess}

T = { GranulomaProcess $\dot\sqsubseteq$ $\top$, hasAssociatedProcess $\dot\sqsubseteq$ $\top \times \top$, CardioVascularDisease $\dot\sqsubseteq$ PathologicalPhenomenon, Fracture $\dot\sqsubseteq$ PathologicalPhenomenon, ∃hasAssociatedProcess.PathologicalProcess $\dot\sqsubseteq$ PathologicalPhenomenon, Endocarditis $\dot\sqsubseteq$ Carditis, Endocarditis $\dot\sqsubseteq$ ∃hasAssociatedProcess.InflammationProcess, PathologicalProcess $\dot\sqsubseteq$ NonNormalProcess }

M = { Endocarditis $\dot\sqsubseteq$ PathologicalPhenomenon, GranulomaProcess $\dot\sqsubseteq$ NonNormalProcess }

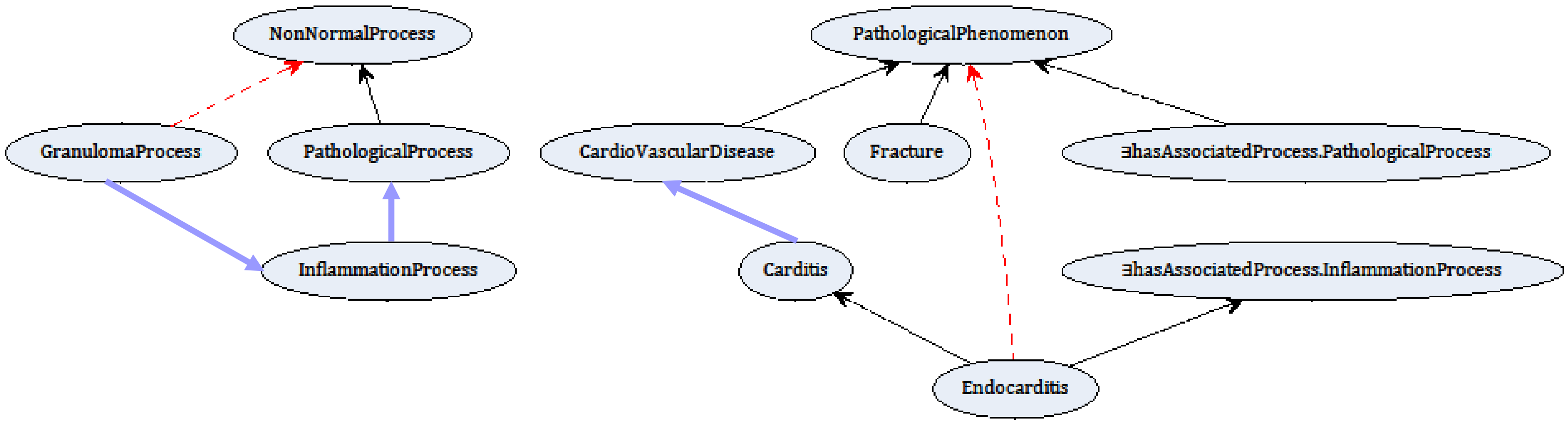
Preference Criteria

- There can be many solutions for GTAP



Preference Criteria

- There can be many solutions for GTAP



Not all are equally interesting.

More Informative

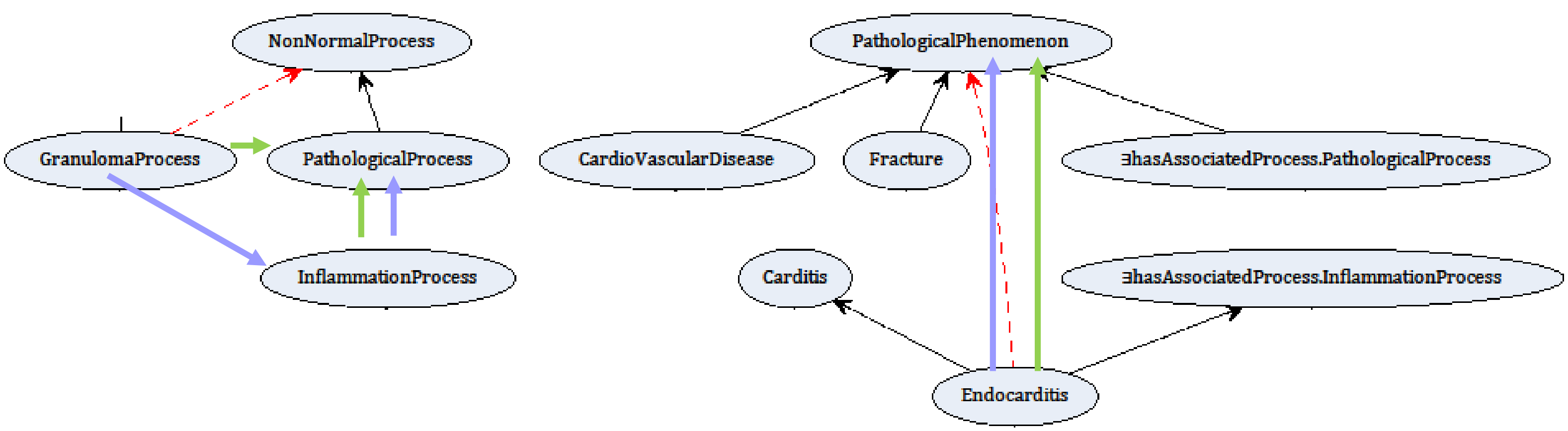
- Let *S* and *S'* be two solutions to GTAP(**T**,**C**,Or,M). Then,

- *S* is more informative than *S'*
iff **T** U *S* |= *S'* but not **T** U *S'* |= *S*

- *S* is equally informative as *S'*
iff **T** U *S* |= *S'* and **T** U *S'* |= *S*

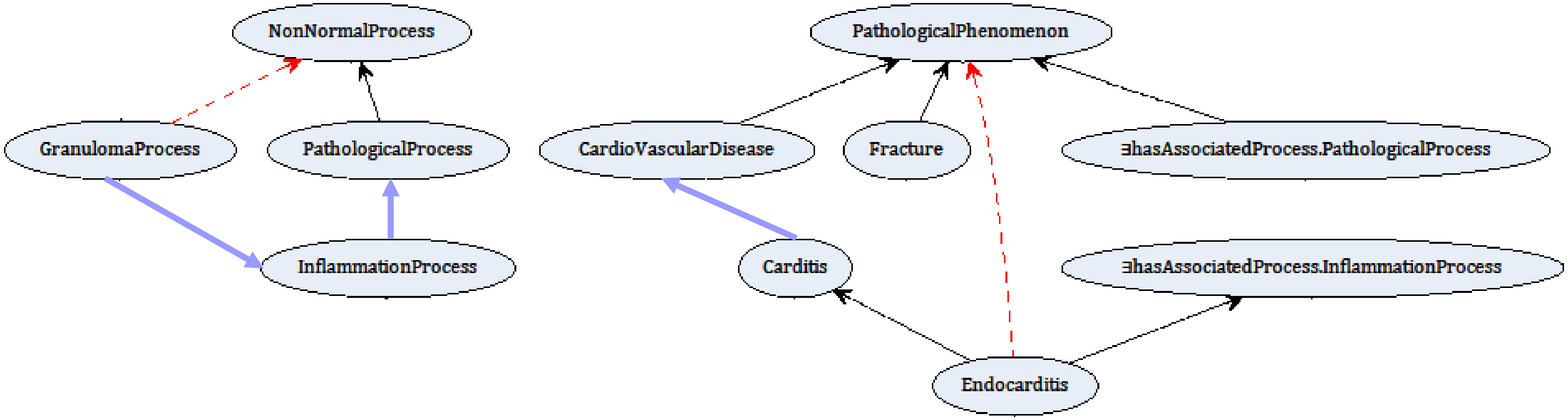
More Informative

- 'Blue' solution is more informative than 'green' solution



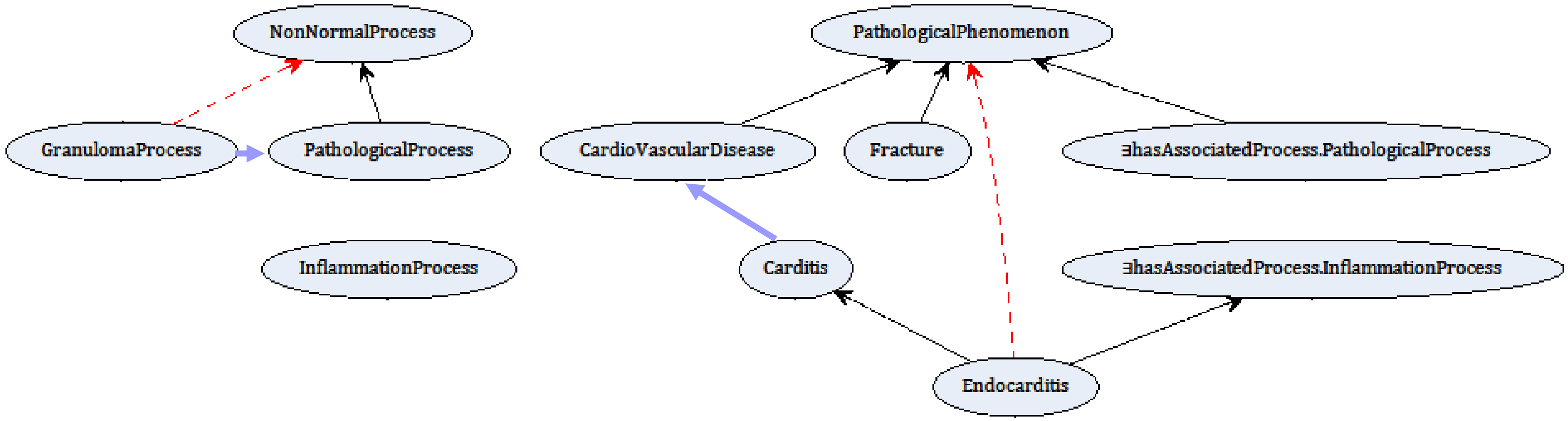
Semantic Maximality

- A solution *S* to GTAP(**T**,**C**,Or,M) is semantically maximal iff there is no solution *S*' which is more informative than *S*.



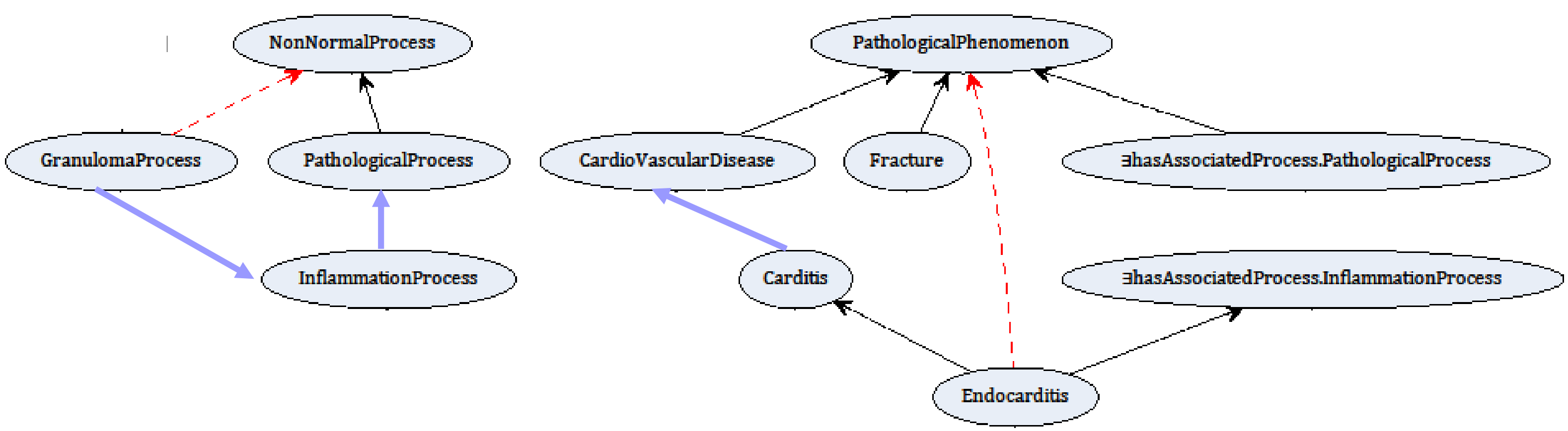
Subset Minimality

- A solution *S* to GTAP(**T**,**C**,Or,M) is subset minimal iff there is no proper subset *S*' of *S* that is a solution.



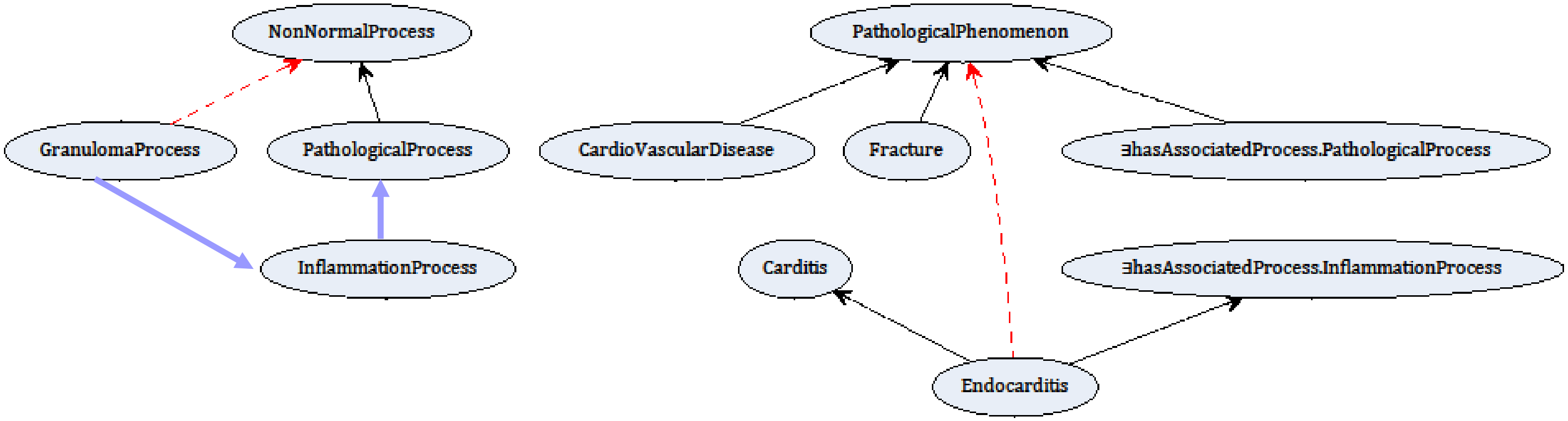
Combining with Priority for Semantic Maximality

- A solution *S* to GTAP(**T**,**C**,Or,M) is maxmin optimal iff *S* is semantically maximal and there is no other semantically maximal solution that is a proper subset of *S*.



Combining with Priority for Subset Minimality

- A solution *S* to GTAP(**T**,**C**,Or,M) is minmax optimal iff *S* is subset minimal and there is no other subset minimal solution that is more informative than *S*.



Combining with Equal Preferences

- A solution *S* to GTAP(**T**,**C**,Or,M) is skyline optimal iff there is no other solution that is a proper subset of *S* and that is equally informative than *S*.
  - All subset minimal, minmax optimal and maxmin optimal solutions are also skyline optimal solutions.
  - Semantically maximal solutions may or may not be skyline optimal.

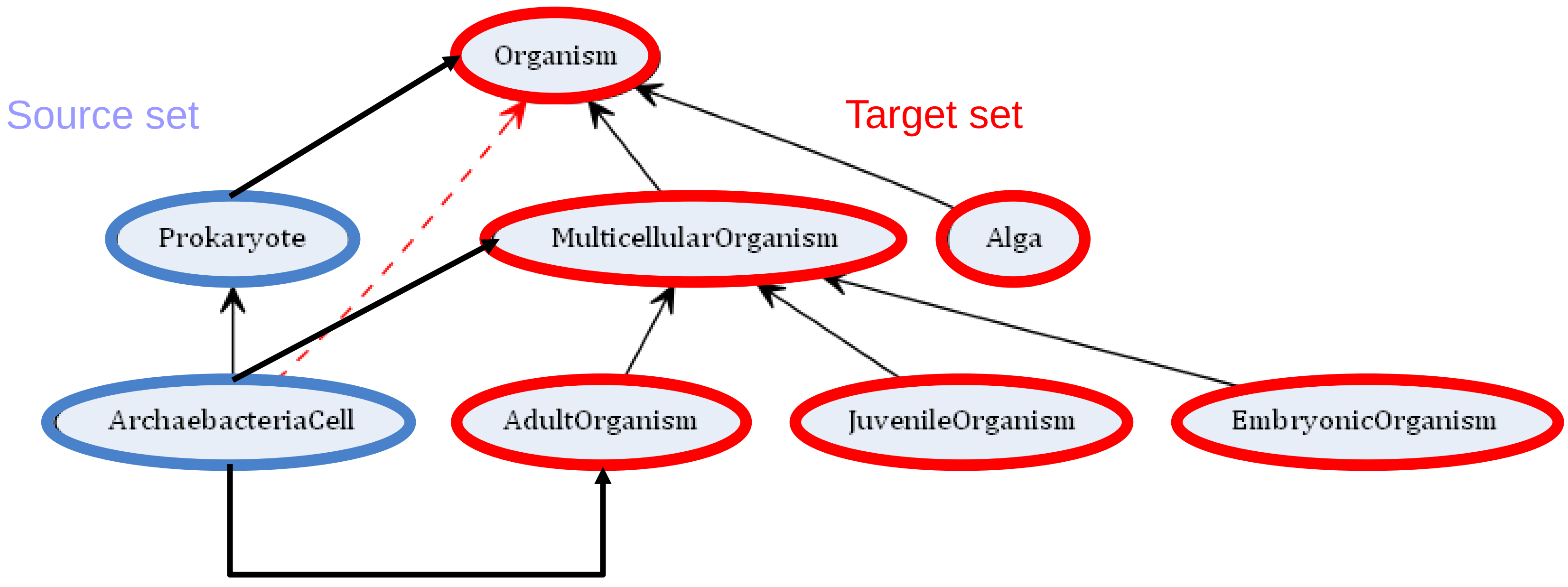
Preference Criteria - Conclusions

- In practice it is not clear how to generate maxmin or semantically maximal solutions (the preferred solutions)
- Skyline optimal solutions are the next best thing and are easy to generate

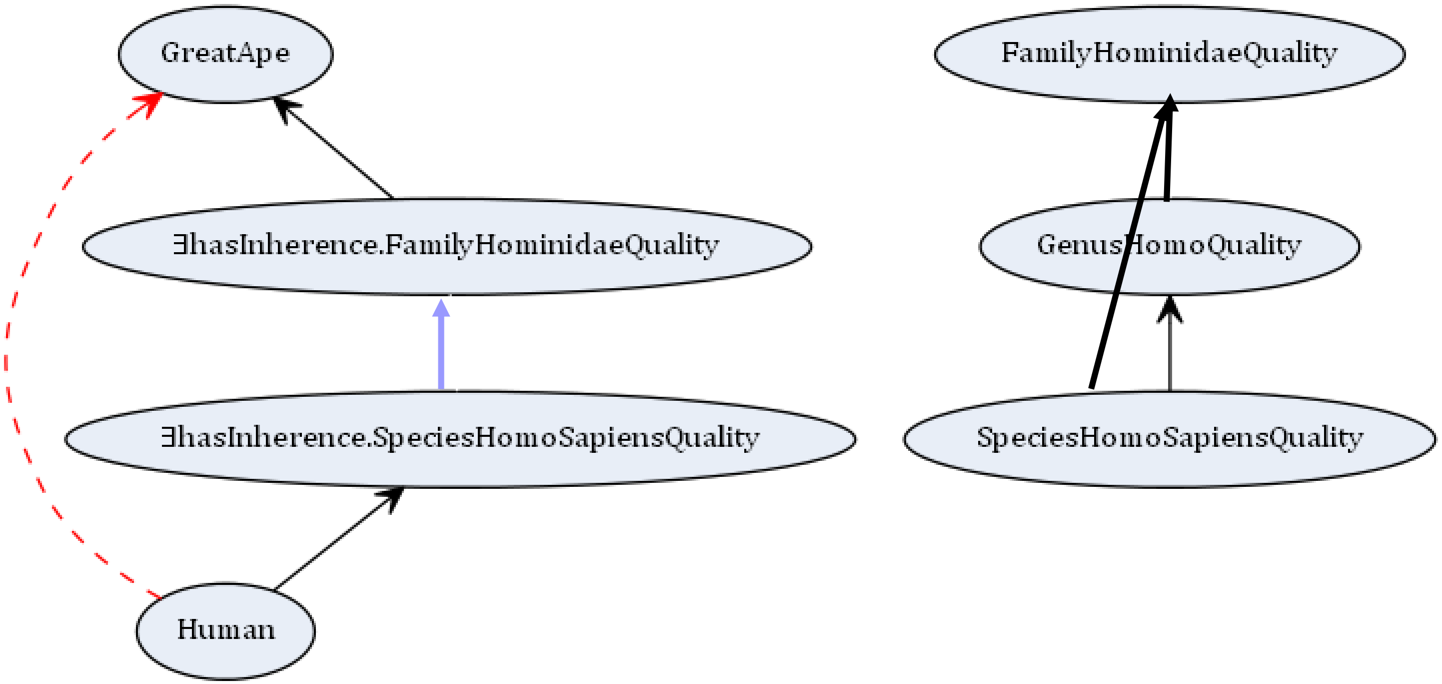
Approach

- Input
  - Normalized EL - TBox
  - Set of missing is-a relations (correct according to the domain)
- Output – a skyline-optimal solution to GTAP
- Iteration of three main steps:
  - Creating solutions for individual missing is-a relations
  - Combining individual solutions
  - Trying to improve the result by finding a solution which introduces additional new knowledge (more informative)

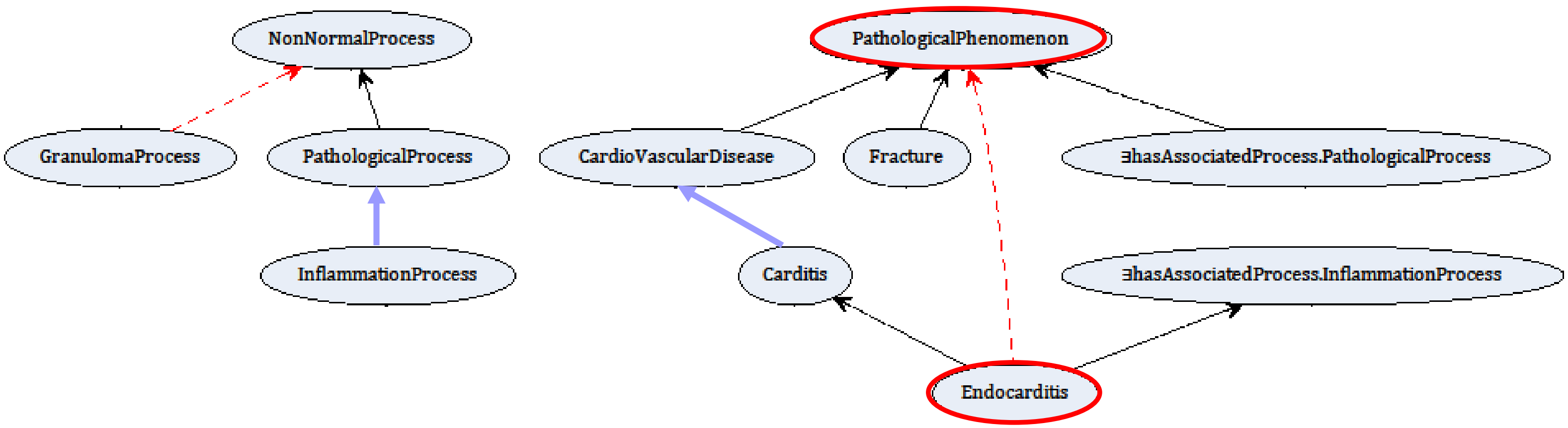
Intuition 1


Organism
Source set
Target set
Prokaryote
MulticellularOrganism
Alga
ArchaebacteriaCell
AdultOrganism
JuvenileOrganism
EmbryonicOrganism

Intuitions 2/3



Example: Repairing Single is–a Relation



~~Endocarditis $\dot{\sqsubseteq}$ PathologicalPhenomenon~~

~~Endocarditis $\dot{\sqsubseteq}$ Fracture~~ false

~~Endocarditis $\dot{\sqsubseteq}$ CardioVascularDisease~~

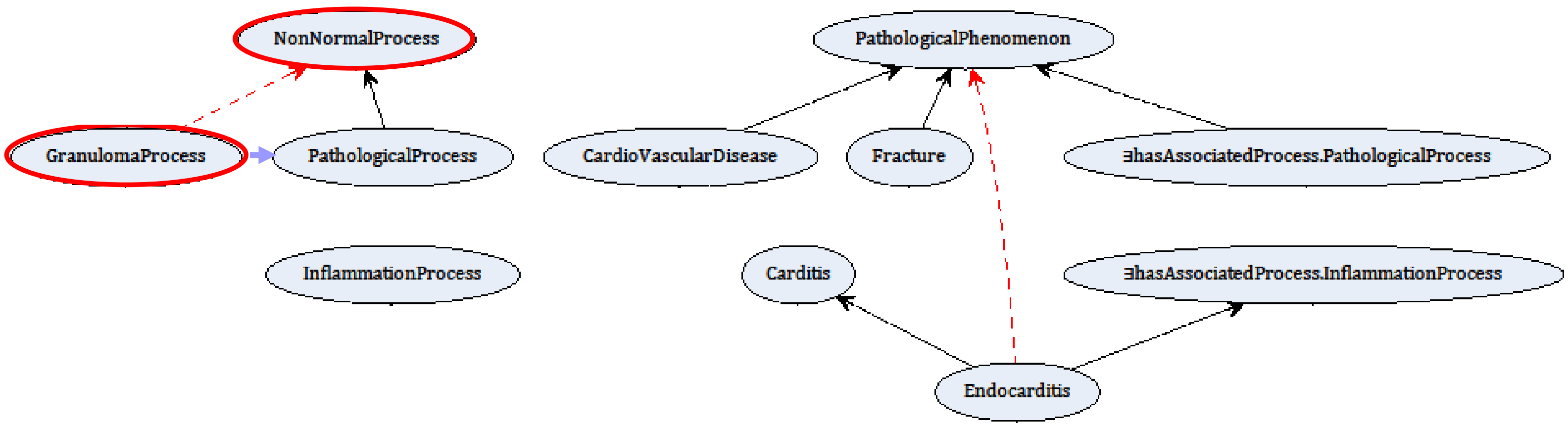
~~Carditis $\dot{\sqsubseteq}$ PathologicalPhenomenon~~

~~Carditis $\dot{\sqsubseteq}$ Fracture~~ false

Carditis $\dot{\sqsubseteq}$ CardioVascularDisease

InflammationProcess $\dot{\sqsubseteq}$ PathologicalProcess

Example: Repairing Single is–a Relation



~~GranulomaProcess $\dot{\sqsubseteq}$ NonNormalProcess~~

GranulomaProcess $\dot{\sqsubseteq}$ PathologicalProcess

Algorithm: Repairing Multiple is-a Relations

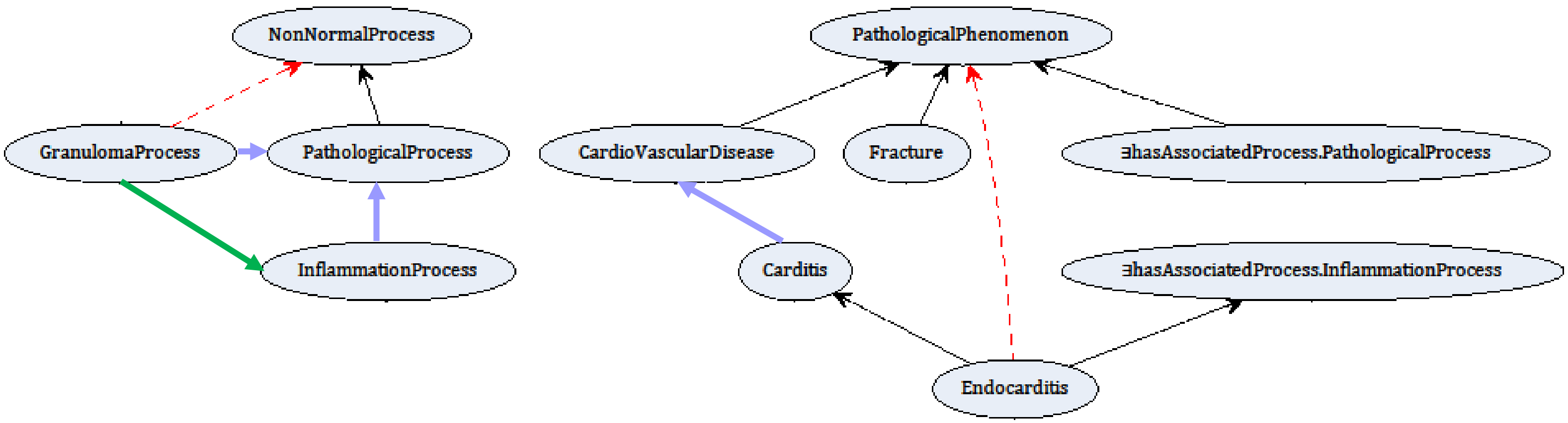
- Combine solutions for individual missing is-a relations
- Remove redundant relations while keeping the same level of informativness
- Resulting solution is a skyline optimal solution

$$\{\text{InflammationProcess} \mathrel{\dot{\sqsubseteq}} \text{PathologicalProcess},$$
$$\text{Carditis} \mathrel{\dot{\sqsubseteq}} \text{CardioVascularDisease},$$
$$\text{GranulomaProcess} \mathrel{\dot{\sqsubseteq}} \text{PathologicalProcess}\}$$

Algorithm: Improving Solution

- Solution S from previous step may contain relations which are not derivable from the ontology.
- These can be seen as new missing is-a relations.
- We can solve a new GTAP problem: GTAP(**T** U *S*, **C**, Or, *S*)

Example: Improving Solutions



$$\text{GranulomaProcess} \mathrel{\dot{\sqsubseteq}} \text{InflammationProcess}$$

$$\{\text{InflammationProcess} \mathrel{\dot{\sqsubseteq}} \text{PathologicalProcess},$$
$$\text{Carditis} \mathrel{\dot{\sqsubseteq}} \text{CardioVascularDisease},$$
$$\text{GranulomaProcess} \mathrel{\dot{\sqsubseteq}} \text{InflammationProcess}\}$$

Algorithm Properties

- Sound
- Skyline optimal solutions

Experiments

Two use-cases

- Case 1: given missing is-a relations
  AMA and a fragment of NCI-A ontology – OAEI 2013
  - AMA (2744 concepts) – 94 missing is-a relations
    → 3 iterations, 101 in repairing (47 additional new knowledge)
  - NCI-A (3304 concepts) – 58 missing is-a relations
    → 3 iterations, 54 in repairing (10 additional new knowledge)

- Case 2: no given missing is-a relations
  Modified BioTop ontology
  - Biotop (280 concepts, 42 object properties)
    randomly choose is-a relations and remove them: 47 'missing'
    → 4 iterations, 41 in repairing (40 additional new knowledge)

Further Reading

Starting points for further studies

Further Reading Ontology Debugging

- http://www.ida.liu.se/~patla/DOOM/

Semantic defects

- Schlobach S, Cornet R. Non-Standard Reasoning Services for the Debugging of Description Logic Terminologies. *18th International Joint Conference on Artificial Intelligence - IJCAI03*, 355-362, 2003.
- Schlobach S. Debugging and Semantic Clarification by Pinpointing. *2nd European Semantic Web Conference - ESWC05*, LNCS 3532, 226-240, 2005.

Further Reading
Ontology Completion

Completing ontologies

- Fang Wei-Kleiner, Zlatan Dragisic, Patrick Lambrix. Abduction Framework for Repairing Incomplete EL Ontologies: Complexity Results and Algorithms. 28th AAAI Conference on Artificial Intelligence - AAAI 2014, 1120-1127, 2014.

- Lambrix P, Ivanova V, A unified approach for debugging is-a structure and mappings in networked taxonomies, *Journal of Biomedical Semantics* 4:10, 2013.

- Lambrix P, Liu Q, Debugging the missing is-a structure within taxonomies networked by partial reference alignments, *Data & Knowledge Engineering* 86:179-205, 2013.

Further Reading Systems

- RepOSE
  http://www.ida.liu.se/~patla00/research/RepOSE/
  video:
  http://videolectures.net/semantic_ivanova_taxonomies/

- OOPS!
  http://oops.linkeddata.es/