Medical University of Graz

Institute for Medical Informatics, Statistics and Documentation,

Modularity Issues in Biomedical Ontologies

Stefan Schulz

Structure of Presentation

- Overview of large ontology projects in the biomedical domain
- Modularization issues
- Study on purpose-driven modularization of SNOMED CT

Structure of Presentation

- Overview of large ontology projects in the biomedical domain
- Modularization issues
- Study on purpose-driven modularization of SNOMED CT

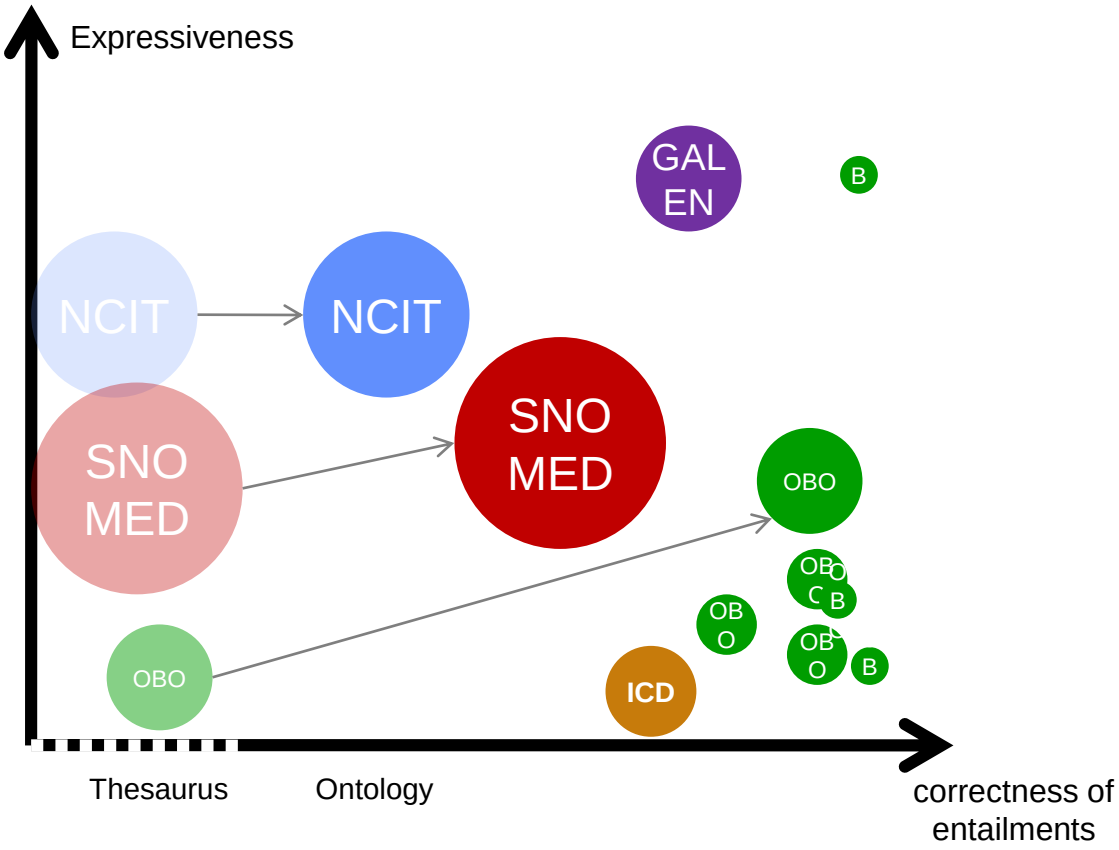
Scoping: Large biomedical ontologies

- Size: > 10^4 representational units
- Domain: life science research and health care
- Use of logic
  - Typical ontologies: OBO Foundry Ontologies (GO, ChEBI, FMA, OBI, …), SNOMED CT, NCIT, GALEN,
  - No ontologies: UMLS meta, MeSH

Large biomedical ontologies are slowly maturing

- Legacy: originally not conceived as ontologies but as systems for classification and semantic annotation
- "Ontologization" of thesaurus-like structures
- Few use cases for logic-based reasoning
- General tendency towards OWL
- Persisting problems
  - Understanding foundations of logic
  - Clarifying ontological commitment

The gradual mutation from thesauri to ontologies



Thesauri and Ontologies represent different things

Informal Thesauri

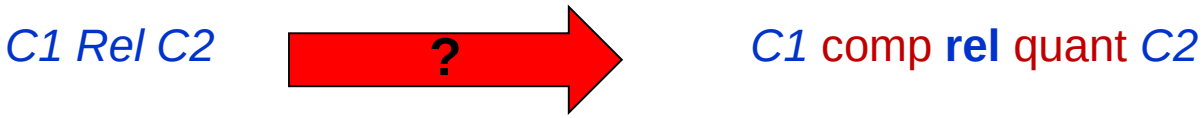
- Examples: MeSH, UMLS Metathesaurus, WordNet
- Describe **terms** of a domain
- **Concepts**: represent the meaning of (quasi-) synonymous terms
- Concepts related by (informal) semantic relations
- Linkage of concepts: *C1 Rel C2*

Formal ontologies

- Examples: openGALEN, OBO, SNOMED
- Describe **entities** of a domain
- **Classes**: collection of entities according to their properties
- Axioms state what is universally true for all members of a class
- Logical expressions: *C1* comp **rel** quant *C2*

Misconceptions of easy and straighforward thesaurus ontologization

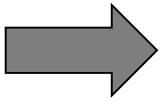
- Upgrading a thesaurus to a formal ontology
- Rationales: use of standards (e.g. OWL-DL), enhanced reasoning, clarification of meaning, internal quality assurance…
- Underestimated:
  - Correctness of quantifications
  - Ontological commitment
  - Thesaurus triples cannot be unambiguously translated into ontology axioms



The translation of triples into DL statements is ambiguous

Translation of triples

C1 Rel C2



C1 subClassOf **rel** some *C2*

or

C1 subClassOf **rel** only *C2*

or

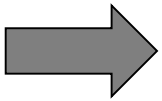
C2 subclassOf inv(**rel**) some *C2*

or…

Translation of groups of triples

C1 Rel C2

C1 Rel C3



C1 subClassOf (**rel** some *C2*) and (**rel** some *C3*)

or

C1 equivalentTo (**rel** some *C2*) and (**rel** some *C3*)

or

C1 equivalentTo (**rel** some ((*C2* or *C3*)))

or …

Ontologies are not exactly made for represent contingent knowledge

- *"Aspirin Treats Headache"*
  "Headache Treated-by Aspirin"
  (seemingly intuitively understandable)
- Translation problem:
  - Not every aspirin tablet treats some headache
  - Not every headache is treated by some aspirin
- Description logics do not allow probabilistic, default, or normative assertions
- Axioms can only state what is true for all members of a class
- Introducing dispositions into ontology possible but not very intuitive ("every aspirin tablet has an inherent disposition which is only realized by treating headache)

Schulz S, Stenzhorn H, Boeker M, Smith B: Strengths and limitations of formal ontologies in the biomedical domain. RECIIS - Electronic Journal in Communication, Information and Innovation in Health, 2009; 3 (1): 31-45:http://dx.doi.org/10.3395/reciis.v3i1.241en
Schulz S, Jansen L: Molecular interactions: On the ambiguity of ordinary statements in biomedical literature. Applied Ontology, 2009; 4 (1): 21-34: http://dx.doi.org/10.3233/AO-2009-0061

Structure of Presentation

- Overview of large ontology projects in the biomedical domain
  - OBO Foundry
  - NCI Thesaurus
  - SNOMED CT
- Study on purpose-driven modularization of SNOMED CT

The OBO Foundry – Modularity by design

- Collaborative, user-driven, bottom up initiative
- Driven by the success of the Gene Ontology
  (… also a thesaurus-like artifact incrementally mutating into an ontology)
- Guided by ontology development principles
- Rooted in upper ontologies (BFO + RO)
- Goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain
- Division of labor amongst domain experts
- Originally using semi-formal OBO syntax, now increasingly OWL-DL

Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L. J.; Eilbeck, K. et al. (2007). "The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration".Nature Biotechnology 25 (11): 1251–1255. doi:10.1038/nbt1346

OBO Foundry: Orthogonality by Upper-level, and Granularity divisions

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)

Linking OBO Foundry ontologies

- OBO Foundry principles
  - Single *is_a* parenthood, further *is_a* relations *inferred*
  - Genus-species definitions
- Reality: most Foundry ontologies have no equivalence axioms at all
- Example:
  - *Calcitonin secreting cell* (Cell Ontology) can be defined as a *Secretory cell* which secretes *Calcitonin* (ChEBI)
  - *Heart development* (Gene Ontology) can be defined as a *Develomental process* which has *Heart* (FMA) as participant
- Conclusion: Sufficient representation of definitional knowledge requires cross-linking OBO modules.

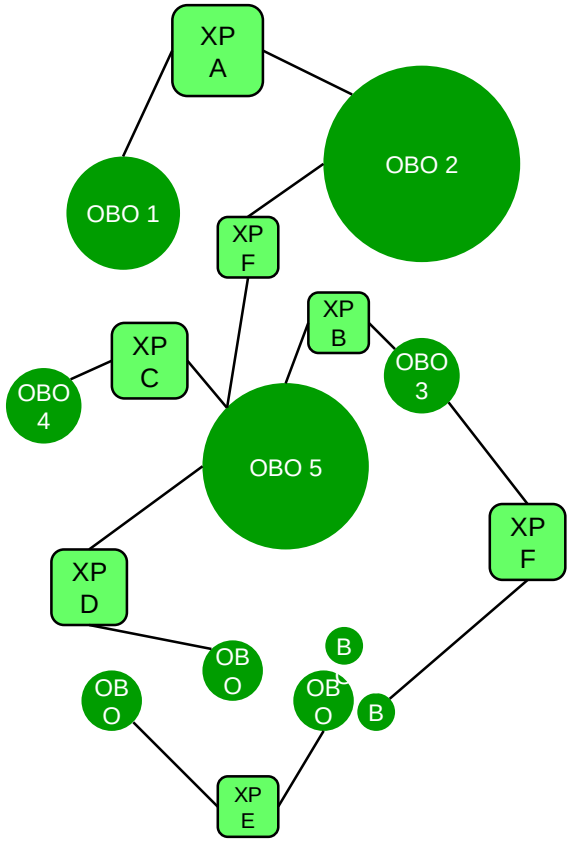
OBO Foundry cross products

		ChEBI	PRO	CC	cell	anat	envo	PATO	MF	BP
ChEBI										
PRO										
GO-CC					16					174
Cell		14		8				70		13
Gross anat										
Envo		19						28		
phen	MP	273	10	?	654	3300		5815	1040	
	WP	115		?	45	432		1450	686	
	TO	127			2	355	19			1
disease										
GO-MF		2407								
GO-BP					511	1085		55		

B. Smith OBO Foundry Update April 2010

OBO Foundry architecture

- Typical axioms in Foundry Modules:
  - A subclassOf B
  - A subclassOf r some C
- Typical axioms in Cross Products
  - A equivalentTo r some D
  - A equivalentTo r some D and s some E



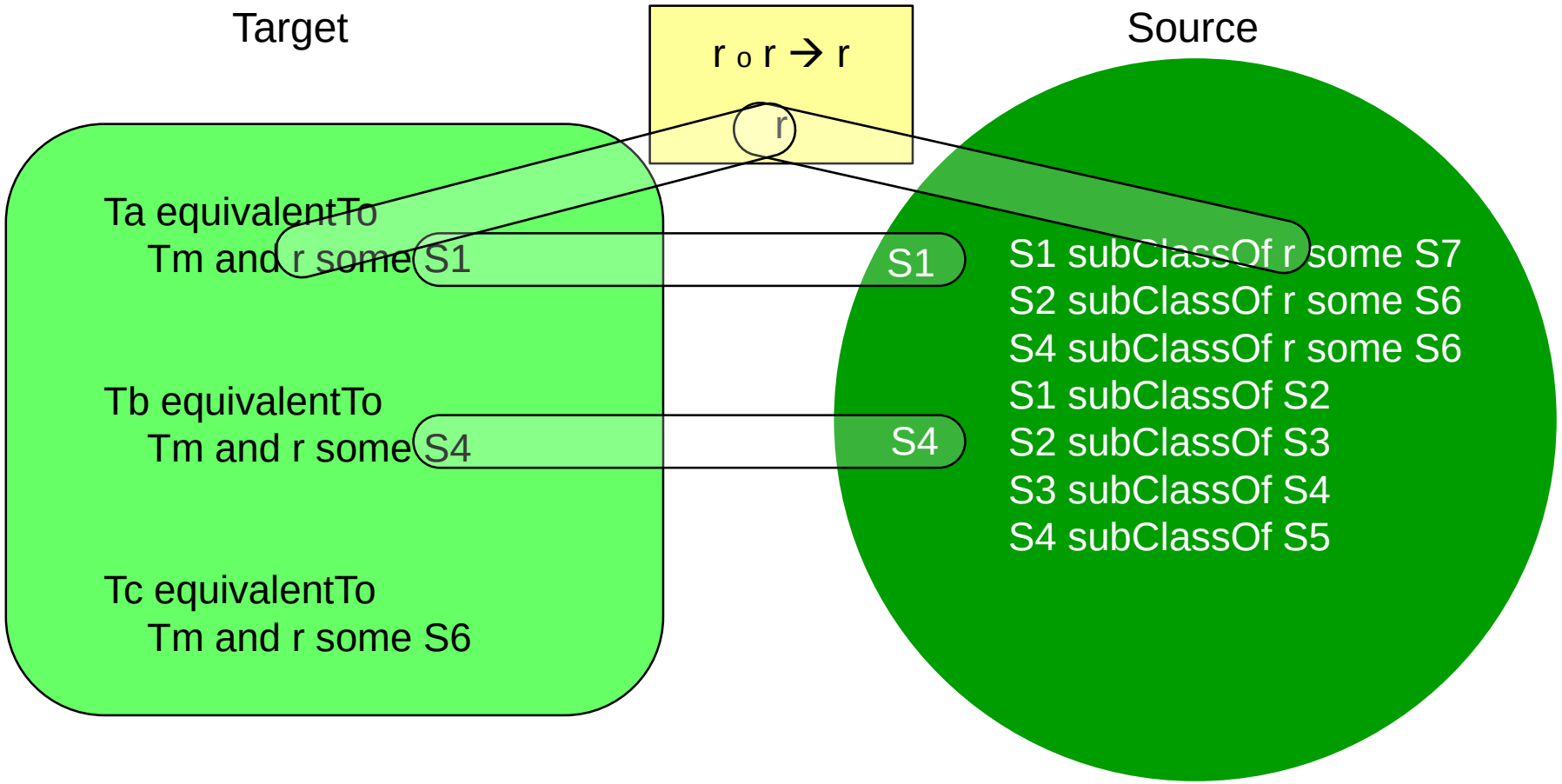
Mélanie Courtot, Frank Gibson, Allyson L. Lister, James Malone, Daniel Schober, Ryan R. Brinkman, and Alan Ruttenberg, MIREOT: The minimum information to reference an external ontology term. Applied Ontology, Vol. 6, Nr. 1 (2011) , p. 23-33

MIREOT: minimum information to reference an external ontology term

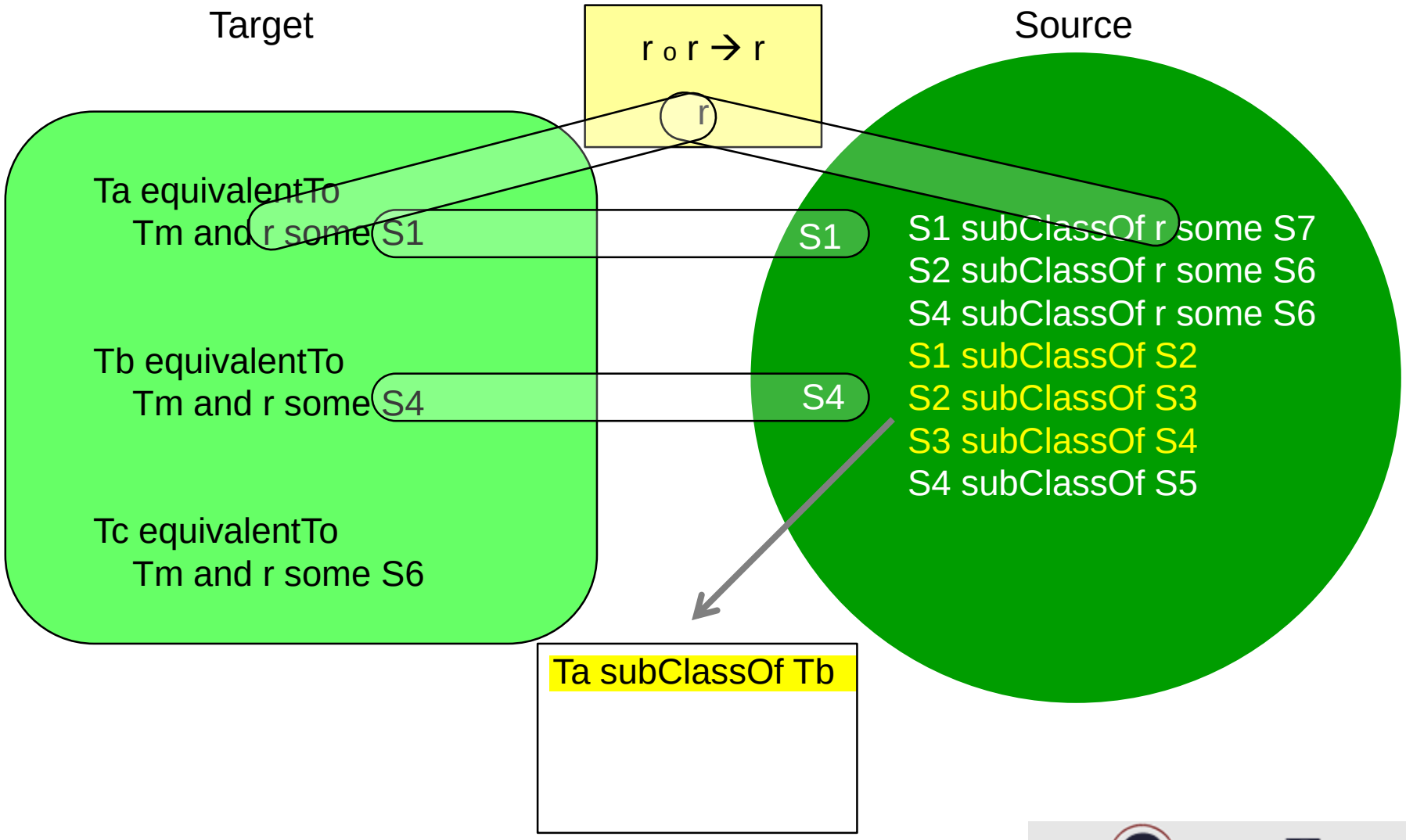
- Guidelines for *manually* importing required terms from an external resource into a target ontology
- Minimal set to unambiguously identify a term: URI of class, source ontology , position in target ontology, superclasses…
- Problem (and possible challenge for WoMo):
  Given a signature (i.e. the set of references to an external ontology): extract axioms in the source ontology that produce new entailments in the target ontology
- Supported by OntoFox (http://ontofox.hegroup.org/)

Mélanie Courtot, Frank Gibson, Allyson L. Lister, James Malone, Daniel Schober, Ryan R. Brinkman, and Alan Ruttenberg, MIREOT: The minimum information to reference an external ontology term. Applied Ontology, Vol. 6, Nr. 1 (2011) , p. 23-33

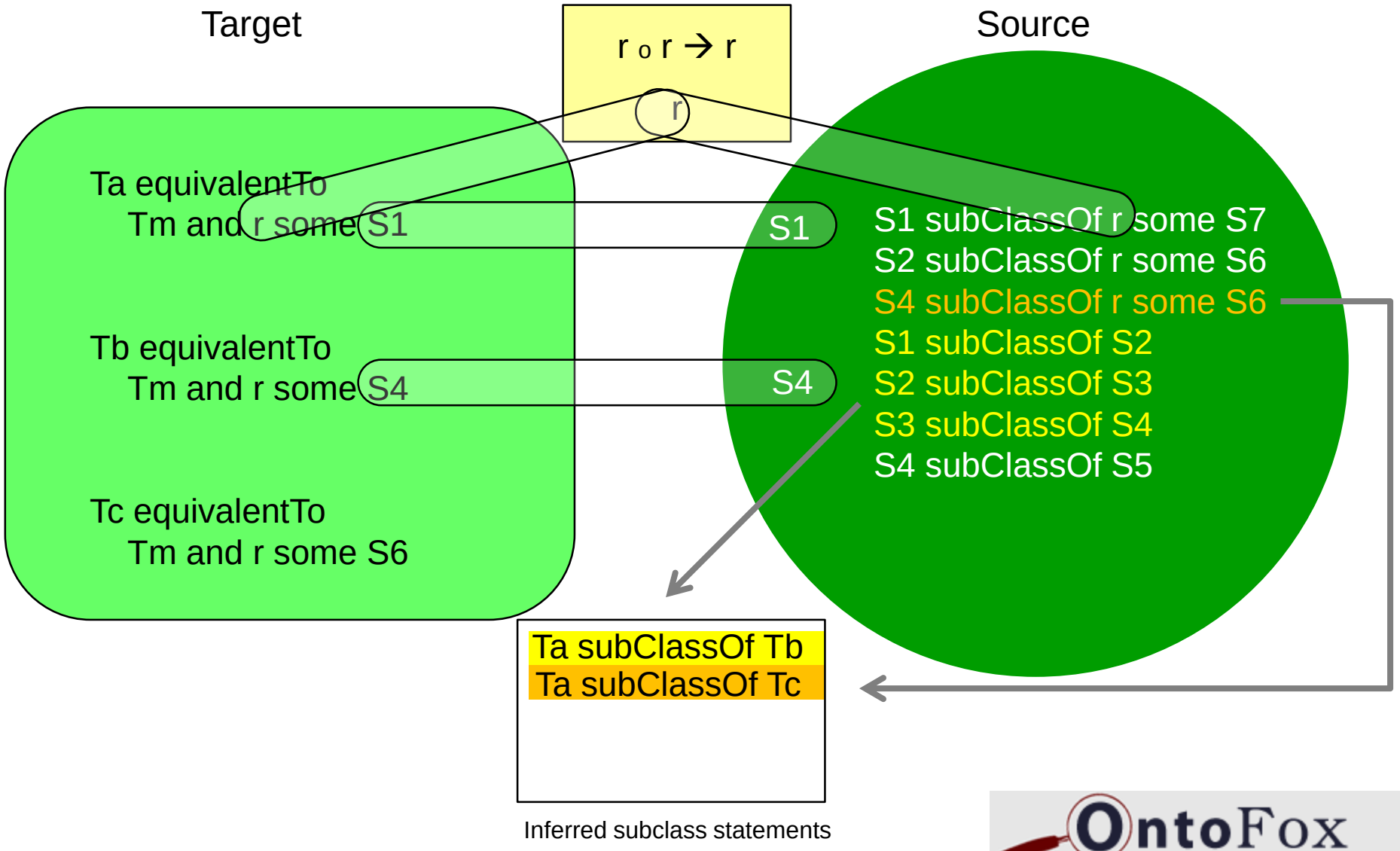
Ontofox implementation of MIREOT: extracting significant axioms from source



Ontofox implementation of MIREOT: extracting significant axioms from source



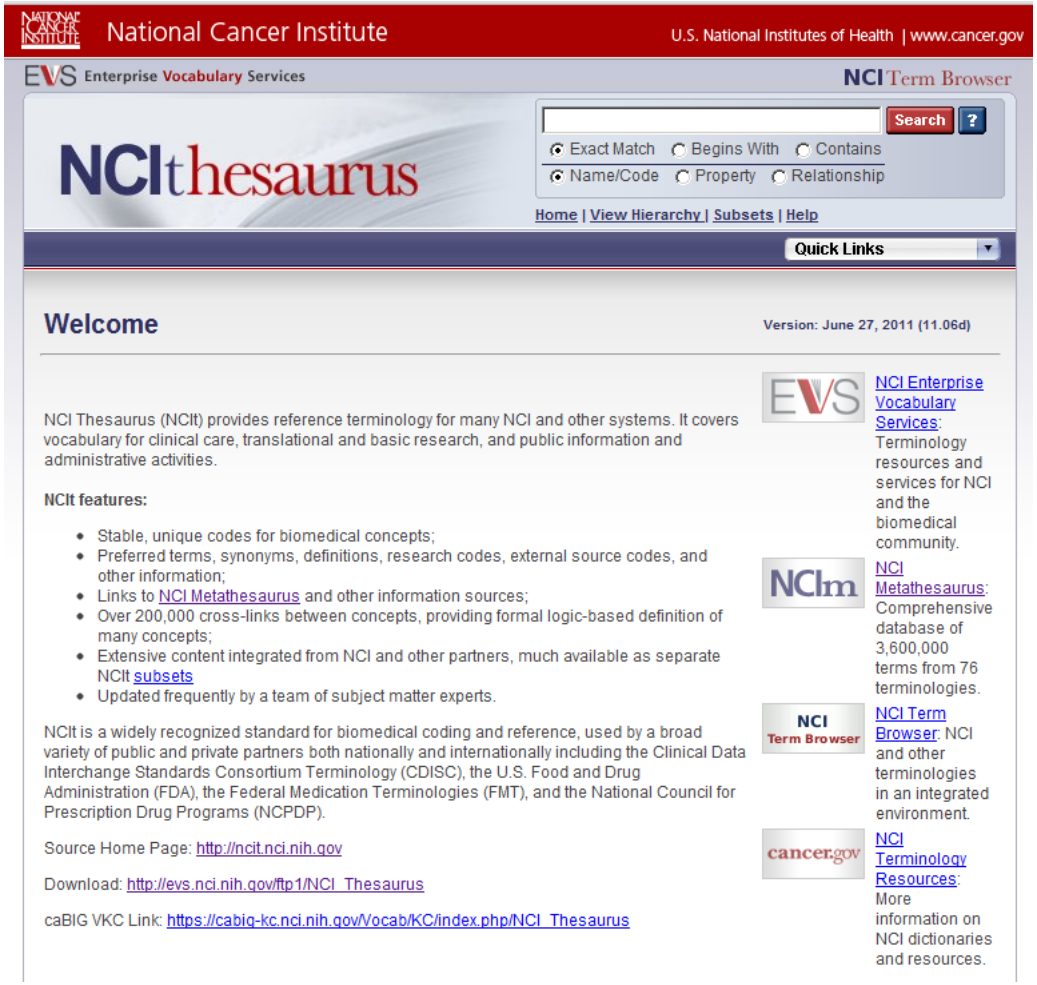
Ontofox implementation of MIREOT: extracting significant axioms from source



Structure of Presentation

- Overview of large ontology projects in the biomedical domain
  - OBO Foundry
  - NCI Thesaurus
  - SNOMED CT
- Study on purpose-driven modularization of SNOMED CT

The NCIthesaurus



http://ncit.nci.nih.gov/ncitbrowser/

The NCI Thesaurus (NCIT) – a (N)ontology for cancer research

- Reference terminology of the National Cancer Institute (clinical care, translational and basic research, public and administrative information)
- Using Ontylog (Apelon, Inc. Terminology Development Environment (TDE))
- 90k classes, 11k eq axioms, 110k subclass axioms, 23k classes interpreted as individuals ("punning") for enabling non-quantified triplets
- Upper-level partition (disjoint axioms at 1st hierarchical level)
- Domain / range restrictions of relations
- Distributed in OWL (SH(D)) as stated and inferred version
- Reasoning use cases unknown
- Statement: "*NCI Thesaurus has some ontology-like features but NCI Thesaurus is not an ontology and is not designed or intended to one*."

http://ncit.nci.nih.gov/ncitbrowser/
ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/NCI%20Thesaurus%20Semantics.pdf

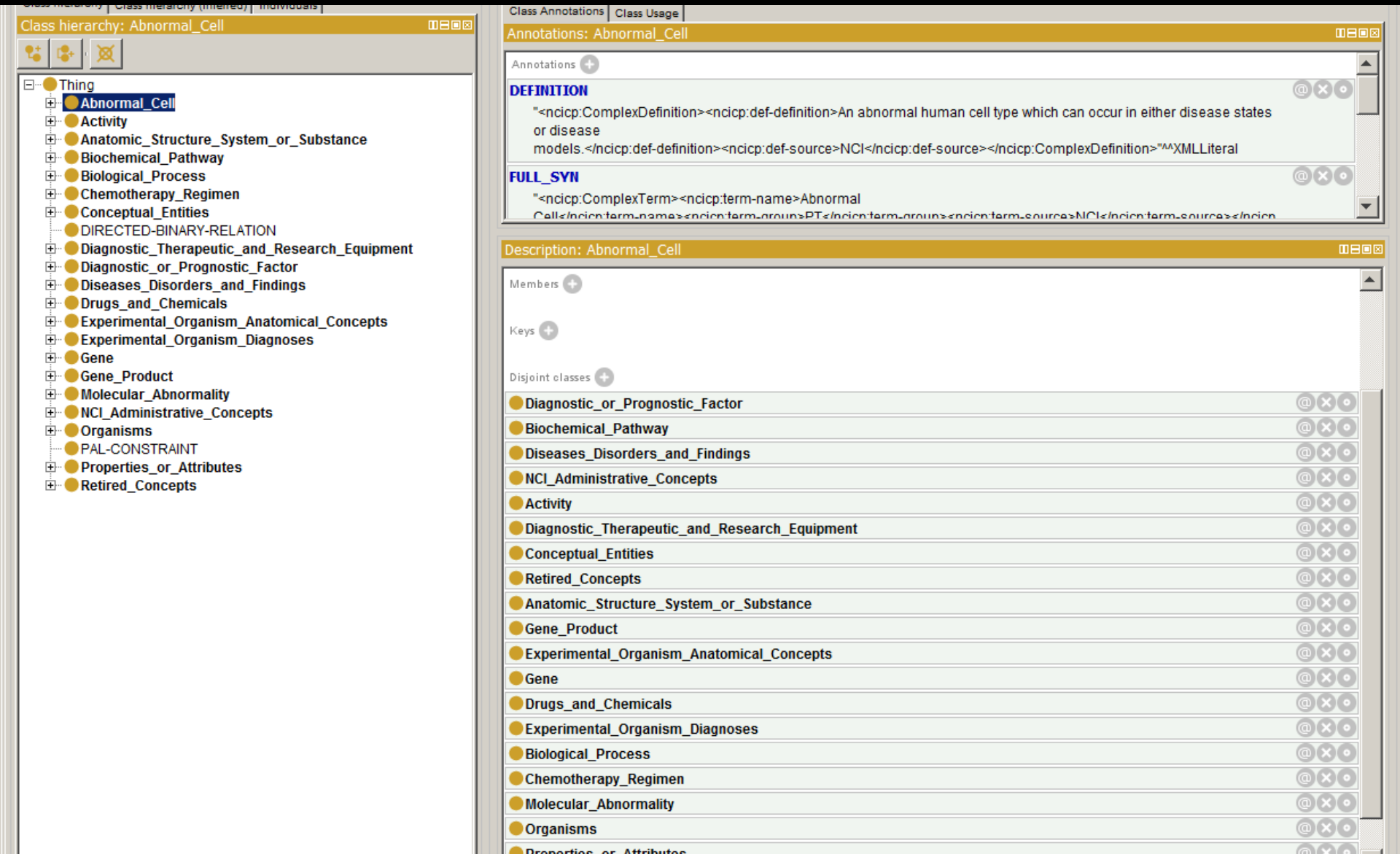
Ontylog language used in the construction of NCIT and SNOMED CT

Constructor	*Syntax*	*Semantics*
Concept name	C	C^I (where $C^I \subseteq \Delta^I$)
Top	$\top$	Δ^I
Bottom	$\bot$	$\varnothing$
Conjunction	$C \sqcap D$	$C^I \cap D^I$
Disjunction	$C \sqcup D$	$C^I \cup D^I$
Universal restriction	$\forall R.C$	$\{x \mid \forall y : R^I(x,y) \rightarrow C^I(y)\}$
Existential restriction	$\exists R.C$	$\{x \mid \exists y : R^I(x,y) \wedge C^I(y)\}$
Modal restriction	$\Diamond R.C$	$\{x \mid \Pr(\exists y : R^I(x,y) \wedge C^I(y)) > 0\}$
Role name	R	R^I (where $R^I \subseteq \Delta^I \times \Delta^I$)

Definitional or Axiomatic Constraint	*Syntax*	*Semantic Constraint*
Concept definition	$C \doteq D$	$C^I \equiv D^I$
Concept subsumption axiom	$C \sqsubseteq D$	$C^I \subseteq D^I$
Role subsumption axiom	$R \sqsubseteq S$	$R^I \subseteq S^I$
Right identity axiom	$R \circ S \doteq R$	$(R \circ S)^I \equiv R^I$

[1] *Note that disjunction, a feature currently under development, can only be used in role values*

The NCIT upper level



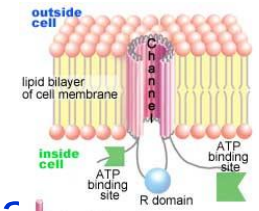
NCIT Content problems

Mismatch between the intended meaning of labels and DL semantics:

- *Ureter_Small_Cell_Carcinoma* subclassOf
  Disease_May_Have_Finding some *Pain*

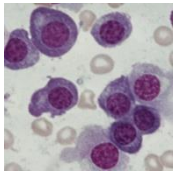
False encoding of distributive statements

- *Calcium-Activated_Chloride_Channel-2* subClassOf
  Gene_Product_Expressed_In_Tissue some *Lung* and
  Gene_Product_Expressed_In_Tissue some *Mammary_Gland* and
  Gene_Product_Expressed_In_Tissue some *Trachea*



Existential quantification over parts instead of wholes

- Antibody_Producing_Cell *subclassOf*
  Part_Of *some* Lymphoid_Tissue

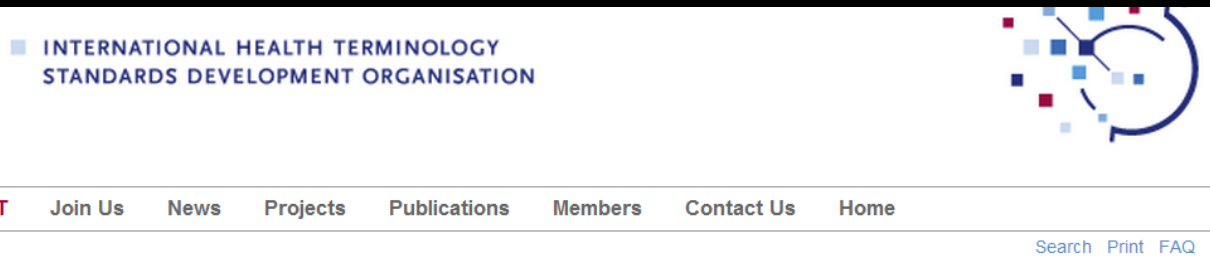


Schulz S, Schober S, Tudose I, Stenzhorn H: The Pitfalls of Thesaurus Ontologization – the Case of the NCI Thesaurus. AMIA Annu Symp Proc, 2010: 727-731: http://proceedings.amia.org/127gtf/1

Structure of Presentation

- Overview of large ontology projects in the biomedical domain
  - OBO Foundry
  - NCI Thesaurus
  - SNOMED CT
- Study on purpose-driven modularization of SNOMED CT

SNOMED CT

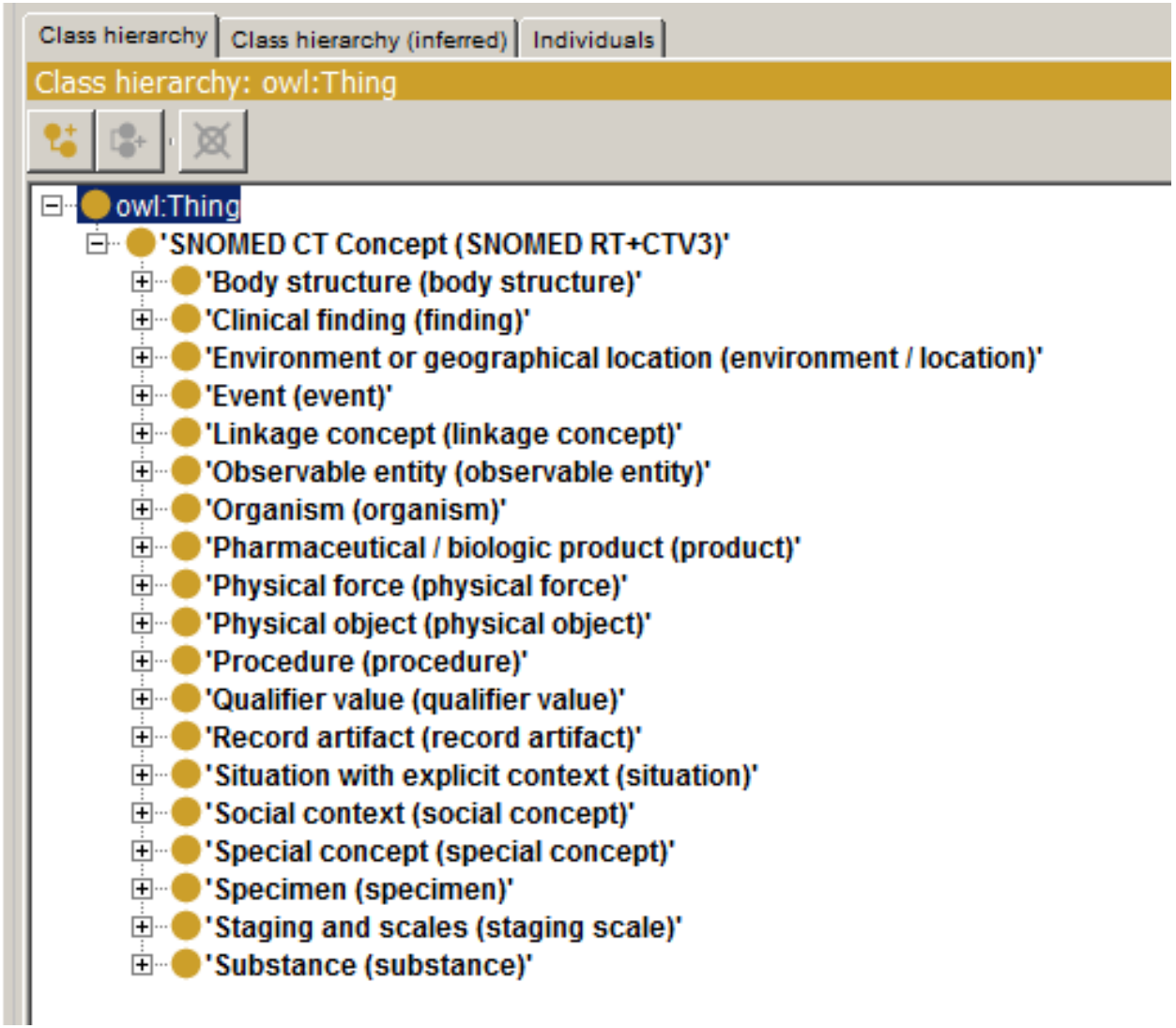


http://www.ihtsdo.org/

SNOMED CT – universal clinical terminology with ontological foundations

- Systematized Nomenclature of Medicine - Clinical Terms
- Terminology for clinical data covering diseases, findings, procedures, organisms, substances etc.
- Using Ontylog (Apelon, Inc. Terminology Development Environment (TDE)), distribution only of the inferred version in tabular format
- Extensive taxonomy with than 311, 000 concepts, connected by 1,360,000 relational expressions
- Script creates OWL version (EL++); no disjointness statements
- OWL version not used in routine

The SNOMED CT upper level



SNOMED CT content problems (frequent)

- Anatomy-related entailments:

  AmputationOfTheFoot equivalentTo *rg* some
  *(***method** some *Amputation and* **procedureSiteDirect** some *FootStructure)*

  AmputationOfToe equivalentTo rg some
  *(***method** *some Amputation* and **procedureSiteDirect** some *ToeStructure)*

  ToeStructure subClassOf *FootStructure*

  AmputationOfToe subclassOf *AmputationOfTheFoot*

- Negations

  ThumbAbsent subClassOf *VenousFinding* and
  hasFindingSite some *ThumbStructure*

- Confusion Epistemology / Ontology

  PresumedViralAgent subClassOf *Virus*

Schulz S, Suntisrivaraporn B, Baader F, Boeker M: SNOMED reaching its adolescence: Ontologists' and logicians' health check. Int J Med Inform, 2009; 78 Suppl. 1: 86-94:http://dx.doi.org/10.1016/j.ijmedinf.2008.06.004

General observations on NCIT and SNOMED CT

- Thesaurus / Ontology hybrids
  - NCIT: "no ontology"
  - SNOMED CT: ontology-aware redesign projects underway, increasing use OWL for prototyping
- Both OWL versions are logically consistent
- Used primarily for semantic annotations
- DL reasoning used in design process, but not really in applications
- Numerous entailments are unreliable, due to plain design errors, but also due to "workarounds" to express negation or probabilistic knowledge
- Browsers can deal with huge amount of terms, reduction in size not a prime desideratum for
- Truth-maintaining modularization not (yet) seen as a requirement

Modularity in design of NCIT / SNOMED CT

- SNOMED CT:
  - Subhierarchies are meant to be disjoint (implicitly)
  - Subhierarchies are separately maintained.
  - Strict control of links between subhierarchies by Domain / Range restriction of relations (in editing tool, not in OWL version)
    - **hasFindingSite** relates *Finding* with *Body Structure*
    - **hasProcedureSite** relates *Procedure* with *Body Structure*
- NCIT:
  - Stated disjointness of top level categories
  - Domain / Range of relations specific for subhierarchies

Modularity in use (SNOMED CT)

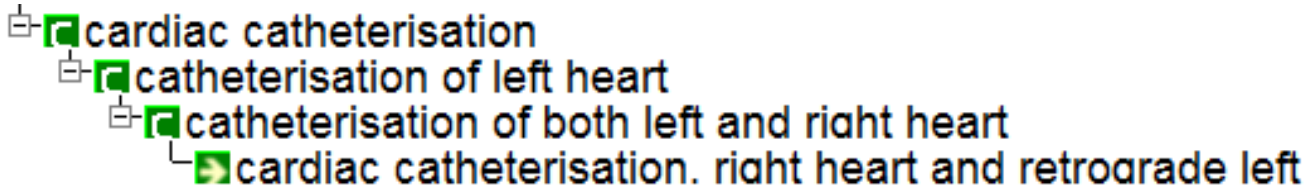
- SNOMED CT: Cross-hierarchy modularity
  - Manually derived problem list subset for summary level clinical documentation:
    - maximize data interoperability among institutions and facilitate the use of SNOMED CT as the primary coding system
    - focus on redesign and QA in content development process
  - SNOMED CT module as ontological basis of the planned version ICD-11 of the International Classification of Diseases
    - manual alignment of SNOMED CT disease classes to ICD classes
    - consistency of taxonomic links
    - inclusion of content referred to in the SNOMED definitions (body parts, qualities, microorganisms)

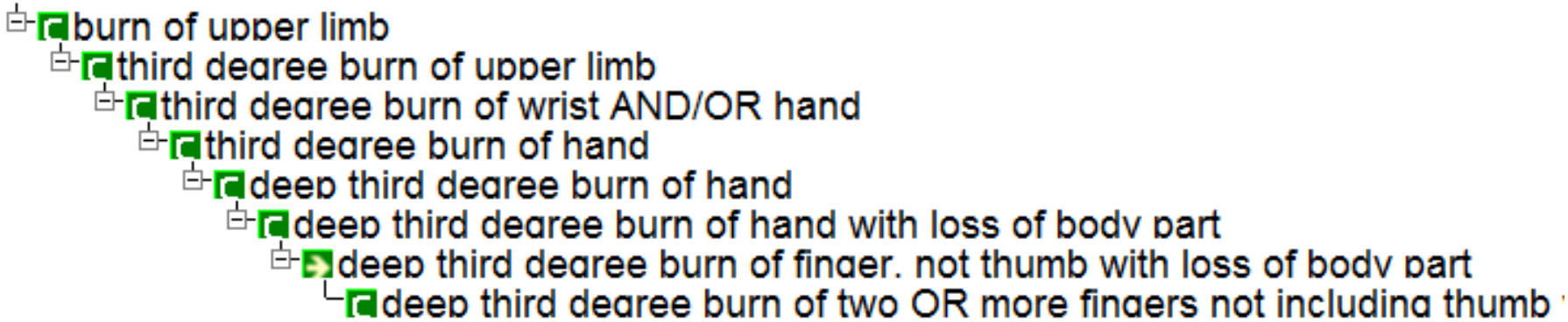
http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html
http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/About_IHTSDO/Harmonization/Final_agreement_2010_WHO_and_IHTSDO.pdf

Structure of Presentation

- Overview of large ontology projects in the biomedical domain
  - OBO Foundry
  - NCI Thesaurus
  - SNOMED CT
- Study on purpose-driven modularization of SNOMED CT

High granularity of SNOMED CT is only partly of interest…





High granularity of SNOMED CT is only partly of interest…

- cardiac catheterisation
  - catheterisation of left heart
    - catheterisation of both left and right heart
      - cardiac catheterisation, right heart and retrograde left ⇦

Cardiologist

- burn of upper limb
  - third degree burn of upper limb
    - third degree burn of wrist AND/OR hand ⇦
      - ~~third degree burn of hand~~
        - ~~deep third degree burn of hand~~
          - ~~deep third degree burn of hand with loss of body part~~
            - ~~deep third degree burn of finger, not thumb with loss of body part~~
              - ~~deep third degree burn of two OR more fingers not including thumb~~

High granularity of SNOMED CT is only partly of interest…

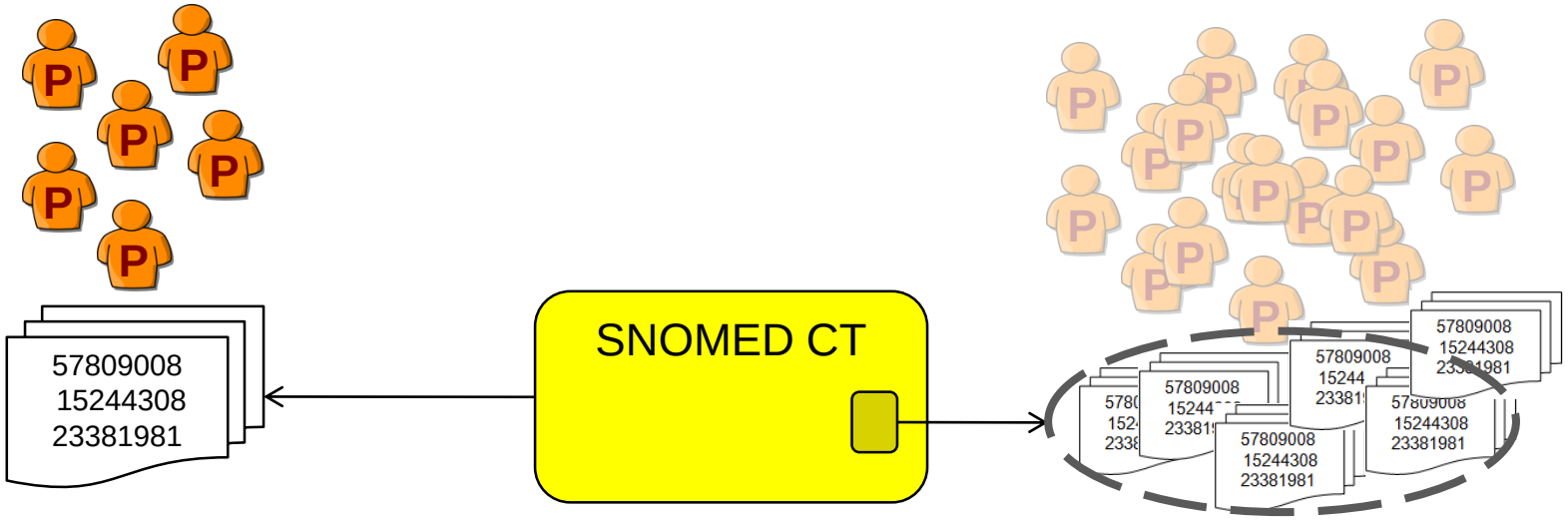
- cardiac catheterisation
  - ~~catheterisation of left heart~~
    - ~~catheterisation of both left and right heart~~
      - ~~cardiac catheterisation, right heart and retrograde left~~

- burn of upper limb
  - third degree burn of upper limb
    - third degree burn of wrist AND/OR hand
      - third degree burn of hand
        - deep third degree burn of hand
          - deep third degree burn of hand with loss of body part
            - deep third degree burn of finger, not thumb with loss of body part
              - deep third degree burn of two OR more fingers not including thumb

Surgeon

Experiment on SNOMED CT modularization

- Given
  - an inpatient population of a specific clinical discipline (cardiology)
  - a collection of fully SNOMED coded medical records
- Create a compact SNOMED module suited to contain address the coverage and granularity requirements of the subdomain



Pablo López-García, Martin Boeker, Arantza Illarramendi, Stefan Schulz. Usability-driven pruning of large ontologies. The case of SNOMED CT. Paper currently under review

Sample fragment of an annotated discharge summary (manually coded)

Original Text	SNOMED CT Term	SNOMED CT Concept ID
Masculino,	Male (finding)	248153007
43 anos,	Aging (finding)	248280005
hipertenso,	Hypertensive disorder, systemic arterial (disorder)	38341003
tabagista,	Tobacco user (finding)	110483000
etilista,	Current drinker of alcohol (finding)	219006
interna	Hospital admission (procedure)	32485007
por infarto agudo do miocárdio	Acute myocardial infarction (disorder)	57054005
sem supradesnivelmaneto de segmento ST.	ST segment elevation (finding)	76388001

Graph traversal heuristics

- SNOMED CT as a graph with
  - taxonomic links — A subClassOf B
  - nontaxonomic links — A subClassOf r some B
    (A subClassOf rg some (r some B))

- Basic approach (Seidenberg & Rector):
  - for each signature concept:
    - builds a set containing the complete hierarchy of the node
    - recursively follow links for every node in the set.
    - property filtering and depth limiting to limit the size of the target ontology

Seidenberg J, Rector A. Web ontology segmentation: analysis, classification and use. Word Wide Web Conference, 2006. 2006:13–22.

Graph traversal heuristics: Four variations

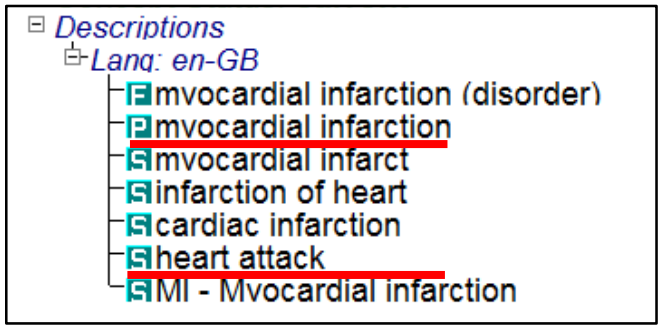
- Four strategies
  - Upwards Segmentation: does not follow links in the subtree of the input concept
  - S-Heuristic: adding all sibling nodes (from all parents) and their *is-a* links.
  - ST-Heuristic: additionally include the complete subtree of each sibling.
  - IL-Heuristic: Each node connected by an incoming link is used as an extra seed.

Seidenberg J, Rector A. Web ontology segmentation: analysis, classification and use. Word Wide Web Conference, 2006. 2006:13–22.

Relevance assessment of each SNOMED CT concepts based on corpus statistics

- Basic principle: score each SNOMED CT concept by relevance according to the occurrences of the attached terms in a reference corpus (MEDLINE)
- Several thresholds

SNOMED CT
Concept 22298006



J Cardiovasc Pharmacol Ther. 2011 Sep;16(3-4):260-6.

Perconditioning and postconditioning: current knov

Vinten-Johansen J, Shi W.

1Division of Cardiothoracic Surgery, The Cardiothoracic Research Laboratory, Ca

Abstract

The broad definition of "conditioning" is the application of a series of alte
ischemic event in ischemic preconditioning, it is applied during the even
heart attack victims to the hospital before percutaneous coronary interve
enthusiastic attention from scientists that have done much to demonstra
including gradual normalization of tissue pH, reduction in generation of r
postconditioning does not exert cardioprotection in experimental models
unknown, as is the interrelationship between the many molecular, cellul
conditioning from another are unanswered questions. Yet, the translatio
developments from industry.

PLoS One. 2011;6(7):e22693. Epub 2011 Jul 28.

Identifying unique neighborhood characteristics to guide health planning for stroke and heart attack: fuzzy cluster and

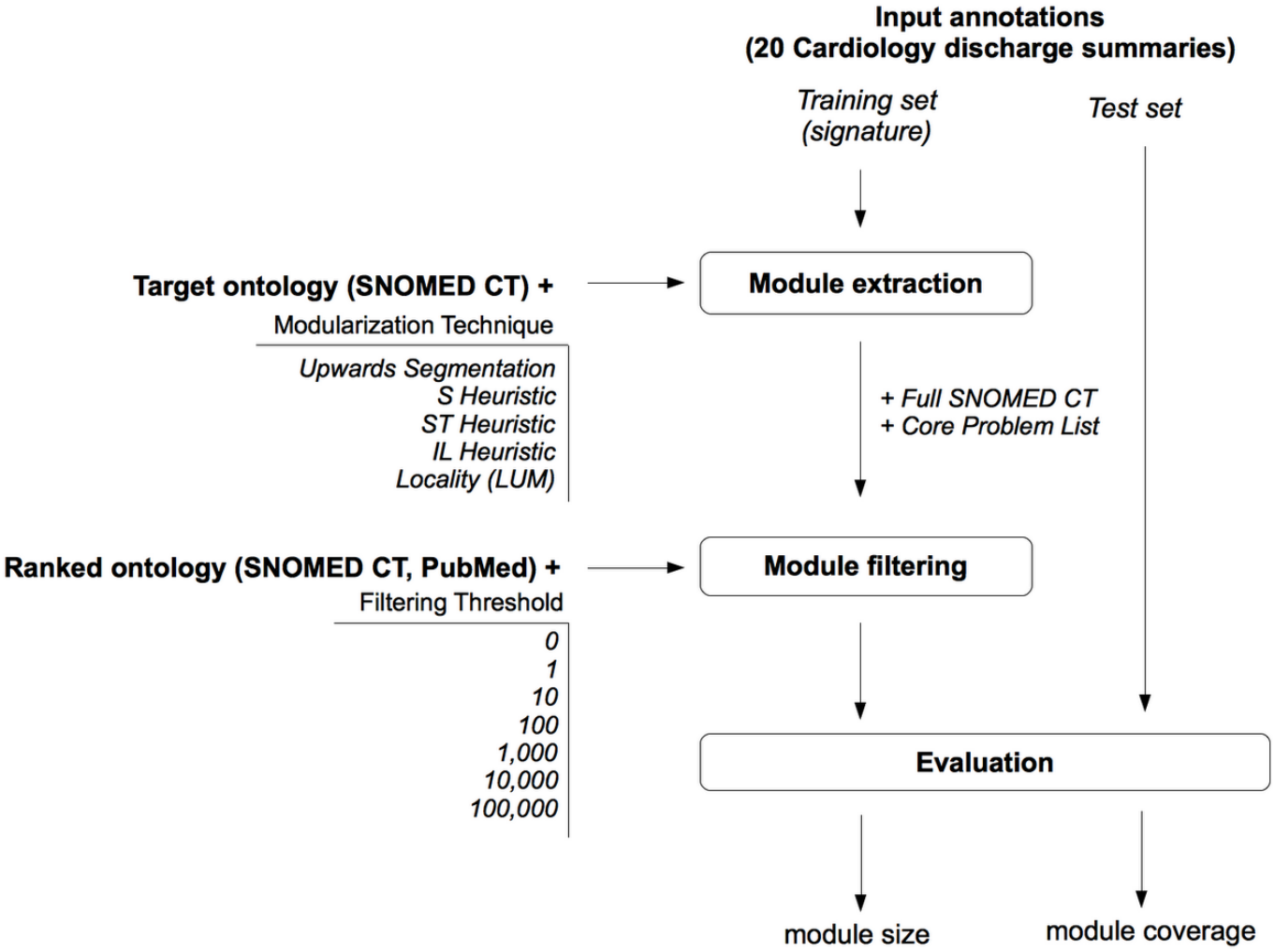
Pedigo A, Seaver W, Odoi A.

Department of Comparative Medicine, The University of Tennessee, Knoxville, Tennessee, United States of America.

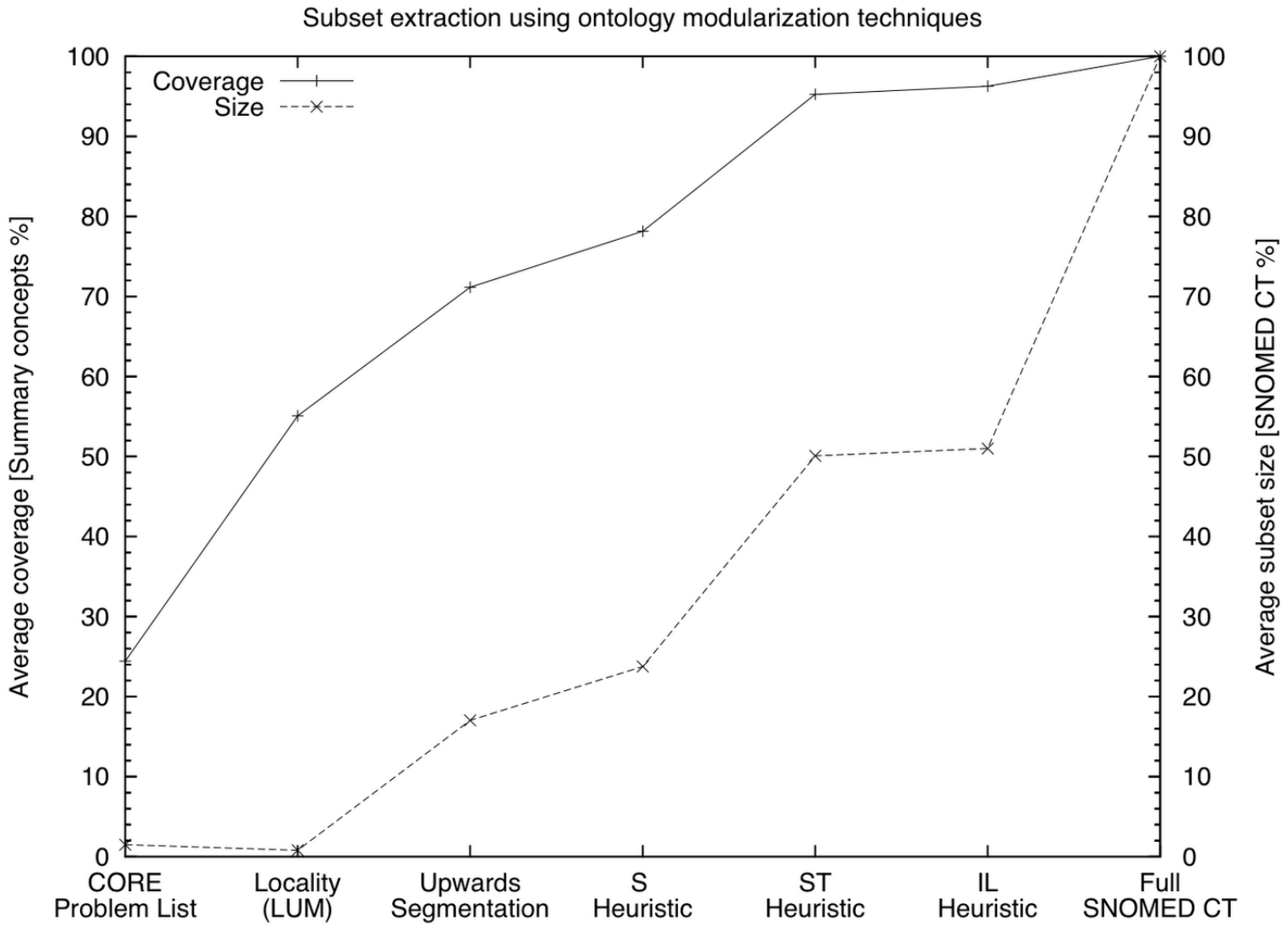
Abstract

BACKGROUND: Socioeconomic, demographic, and geographic factors are known determinants of stroke and myocardial infarction (MI) risk. Clustering of these factor
programs intended to reduce disparities. Given the complex and multidimensional nature of these factors, multivariate methods are needed to identify neighborhood cl

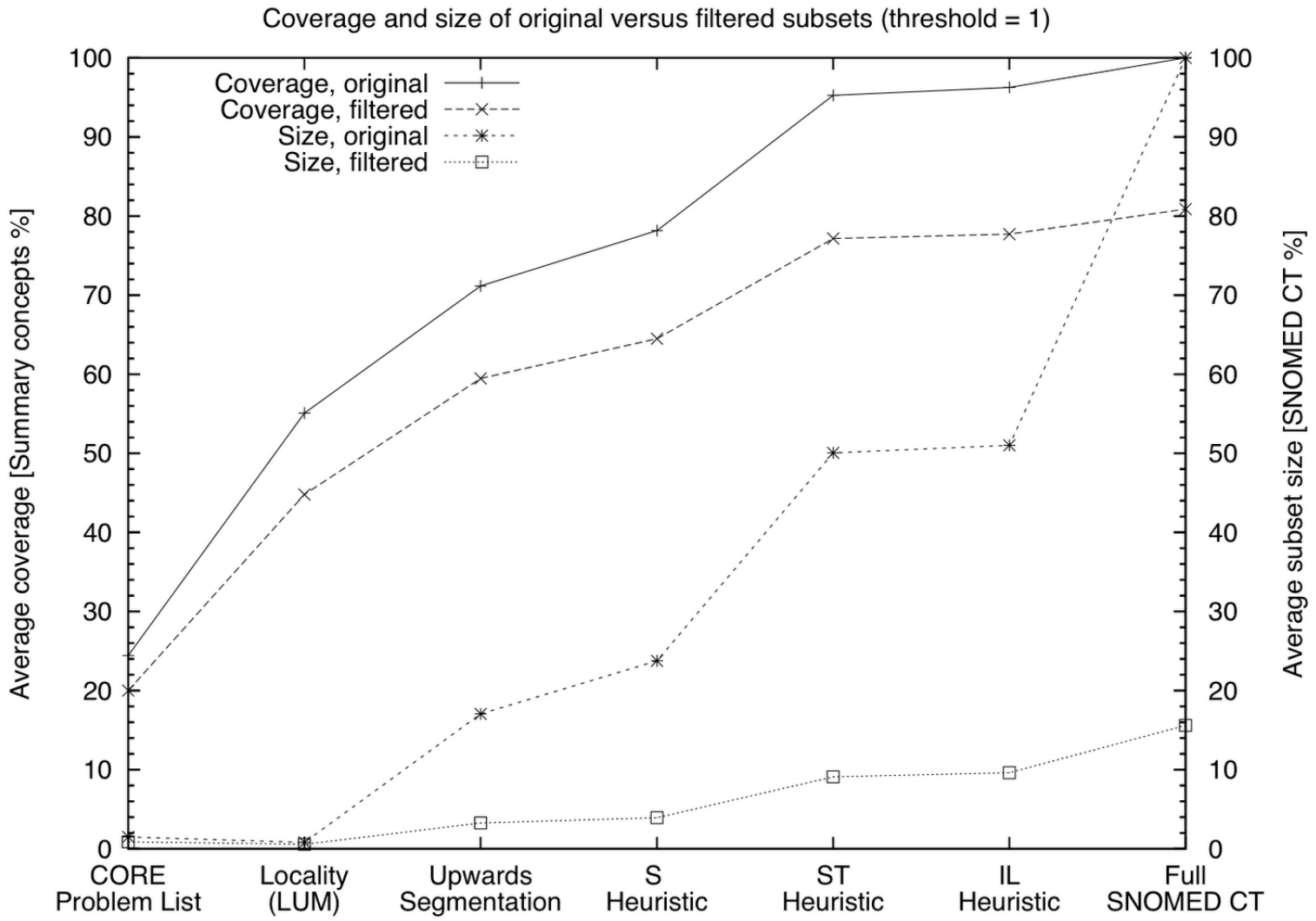
Experimental design, parameters and measured variables

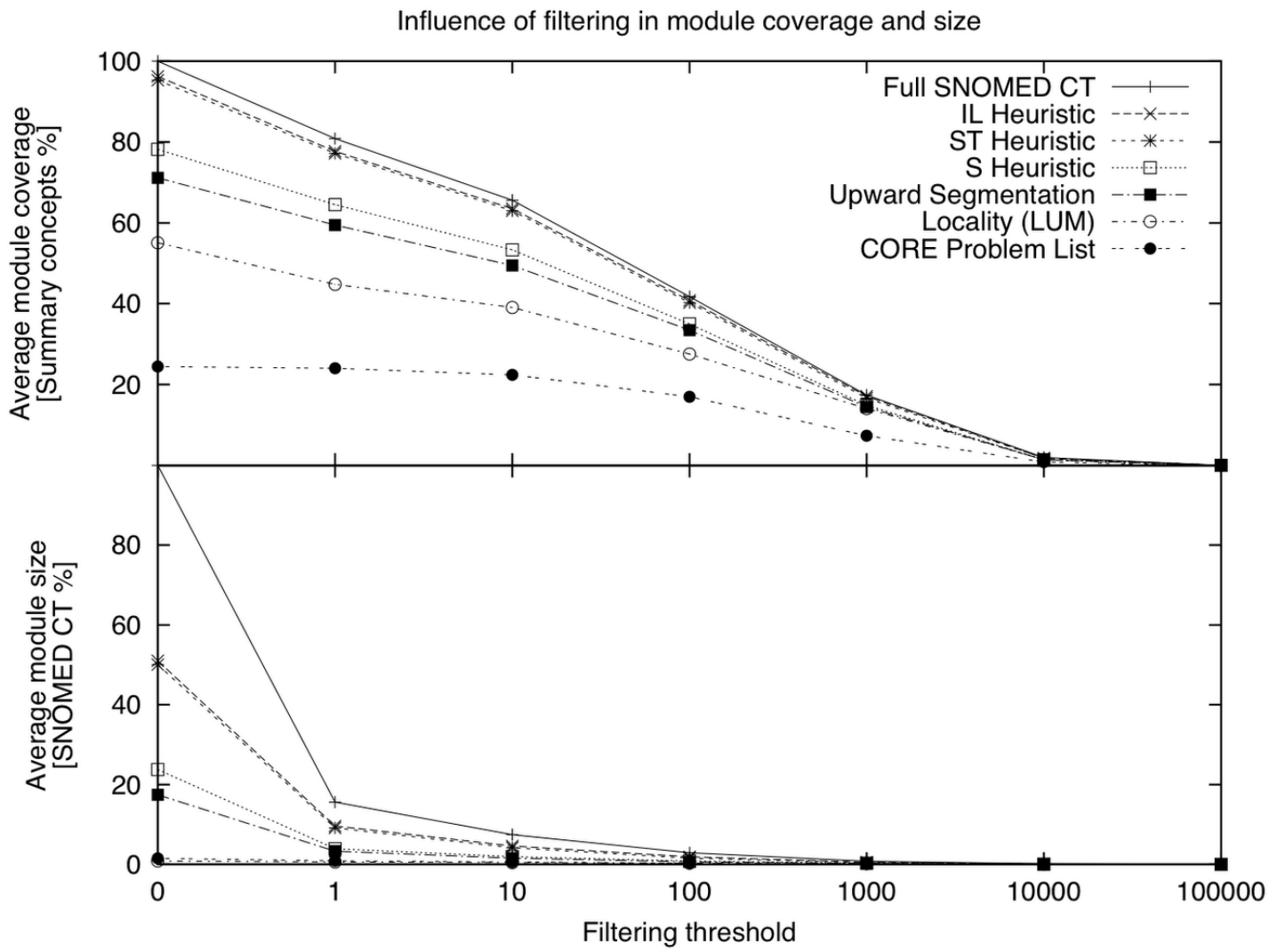


Results: Coverage and Size



Results: Coverage and Size




Influence of filtering in module coverage and size
Average module coverage [Summary concepts %]
Average module size [SNOMED CT %]
Filtering threshold
Full SNOMED CT
IL Heuristic
ST Heuristic
S Heuristic
Upward Segmentation
Locality (LUM)
CORE Problem List
0
20
40
60
80
100
0
1
10
100
1000
10000
100000

Discussion

- Module creation
  - If a coverage of 80% is considered, S-heuristics method of choice
  - Possible improvement: use of different approaches in different SNOMED CT subhierarchies
- Filtering
  - With complete SNOMED CT only 80%
  - Problem: vocabulary mismatch trainings corpus / test corpus. Improvement expected using clinical text and looser string matching
- Work in progress
- Testing / adapting other segmentation methods

Conclusion

- Module creation in large biomedical ontologies:
  - Main interest in domain-specific subset
  - Module should cover must much more than the signature nodes
  - Logical properties of modules are still secondary, since logical entailments of both SNOMED CT and NCIT are not yet reliable
- Module creation in the context of (already modularly structured) OBO Foundry
  - Main interest in enrichment of the axioms in target module by re-using content from source module(s)
  - Finding additional axioms from source to compute new entailments in target

Acknowledgements

Organizers of WoMo 2011

Pablo López

(PhD candidate at University of the Basque Country, Spain, currently at Stanford University)

pablo_lopez@ehu.es