The Pitfalls of Thesaurus Ontologization - the Case of the NCI Thesaurus

Stefan Schulz[1,2], Daniel Schober[1],
Ilinca Tudose[1], Holger Stenzhorn[3]

[1]Institute of Medical Biometry und Medical Informatics,
University Medical Center Freiburg, Germany
[2]AVERBIS GmbH, Freiburg, Germany
[3]Paediatric Hematology and Oncology, Saarland University Hospital, Homburg, Germany

averbis
medical language technology

UNIVERSITÄTS KLINIKUM FREIBURG

Typology

Informal Thesauri

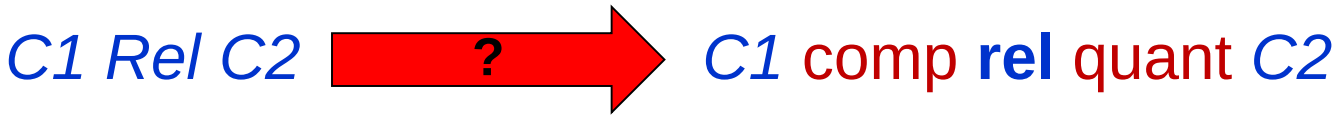
- Examples: MeSH, UMLS Metathesaurus, WordNet
- Describe **terms** of a domain
- **Concepts**: represent the meaning of (quasi-) synonymous terms
- Concepts related by (informal) semantic relations
- Linkage of concepts:
  C1 Rel C2

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*

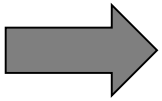
Thesaurus ontologization

- Upgrading a thesaurus to a formal ontology
- Rationales: use of standards (e.g. OWL-DL), enhanced reasoning, clarification of meaning, internal quality assurance…
- Expressiveness of thesauri vs. ontologies:
  - The meaning of thesaurus assertions follows natural language, the meaning of ontology axioms follow mathematical rigor
  - Thesaurus triples cannot be unambiguously translated into ontology axioms



Problem 1: Ambiguity

Translation of triples

C1 Rel C2 

C1 subClassOf **rel** some *C2*

or

C1 subClassOf **rel** only *C2*

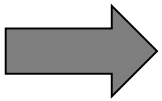
or

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 …

Problem 2: Non-universal statements

- *"Aspirin Treats Headache"*
  "Headache Treated-by Aspirin"
  (seemingly intuitively understandable)
- Translation problem into ontology:
  - 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

Objective of the study

Objective of the study

- Investigate correctness of existentially quantified properties in biomedical ontologies
  - OBO Foundry ontologies
  - OBO Foundry candidates
  - NCIT as an instance of OBO Foundry candidates
- Selection of NCIT
  - Size
  - System in use
  - Importance for generating and communicating standardized meanings in oncology
  - Quality issues already addressed by
    Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine 2005;44(4):498-507.

Assessment Method (I)

- Select a sample of existentially quantified clauses from the NCIT OWL version
- Pattern: *C1* subClassOf **rel** some *C2,* according to description logics semantics :
  *"Every instance of C1 is related to at least one instance of C2 via the relation **rel**"*
- Found: 77 different relation types, used in more than 180,000 existentially qualified clauses
  - Most frequent relation "**Disease_may_have_finding**" (N = 27,653)
  - 15 relation types occurring less than ten times each.
- Sampling: n_i = round $(2 \log_{10}(N_i+1))$ with N_i being the number of existentially qualified restrictions in which r_i was used

Assessment Method (II)

Each sample expression like *C1* subClassOf **Rel** some *C2* was assessed by two experts for correctness

Assessment Criteria:

- Ontological commitment: the NCIT classes extend to real things in the clinical domain
- Focus: to judge whether the ontological dependence of *C1* on *C2* is adequate
- Exact confidence intervals (95%) were computed based on the binomial distribution.
- Also collected: anecdotic evidence of other kinds of errors.

Results

NCIT relation type	# occurrences in OWL "someValues From" clause	sam-ple size	# errors in sample	sample error rate	estimated number of errors	95% CI lower bound	95% CI upper bound	95% CI estimate lower bound	95% CI estimate upper bound
Disease_May_Have_Finding	27,652	9	9	1.00	27,652	0.66	1.00	18,353	27,652
Disease_May_Have_Cytogenetic_Abnormality	18,860	9	9	1.00	18,860	0.66	1.00	12,517	18,860
Gene_Product_Plays_Role_In_Biological_Process	15,607	8	8	1.00	15,607	0.63	1.00	9,842	15,607
Gene_Plays_Role_In_Process	14,385	8	8	1.00	14,385	0.63	1.00	9,071	14,385
Chemotherapy_Regimen_Has_Component	10,861	8	0	0.00	0	0.00	0.37	31	4,012
Gene_Product_Encoded_By_Gene	10,754	8	0	0.00	0	0.00	0.37	30	3,973
Disease_May_Have_Molecular_Abnormality	10,687	8	7	0.88	9,351	0.47	1.00	5,060	10,653
Gene_Is_Element_In_Pathway	8,364	8	8	1.00	8,364	0.63	1.00	5,274	8,364
Gene_Product_Is_Element_In_Pathway	8,302	8	8	1.00	8,302	0.63	1.00	5,235	8,302
Gene_Product_Has_Biochemical_Function	7,695	8	0	0.00	0	0.00	0.37	22	2,843
Anatomic_Structure_Is_Physical_Part_Of	6,285	8	1	0.13	786	0.00	0.53	20	3,309
Gene_In_Chromosomal_Location	5,392	7	0	0.00	0	0.00	0.41	0	2,209
Gene_Found_In_Organism	4,086	7	0	0.00	0	0.00	0.41	0	1,674
Disease_May_Have_Associated_Disease	3,353	7	7	1.00	3,353	0.59	1.00	1,980	3,353
EO_Disease_Has_Associated_EO_Anatomy	3,102	7	0	0.00	0	0.00	0.41	0	1,271
Gene_Has_Physical_Location	2,945	7	0	0.00	0	0.00	0.41	0	1,206
Gene_Product_Expressed_In_Tissue	2,476	7	7	1.00	2,476	0.59	1.00	1,462	2,476
Disease_May_Have_Abnormal_Cell	2,442	7	7	1.00	2,442	0.59	1.00	1,442	2,442
Gene_Product_Has_Associated_Anatomy	1,972	7	1	0.14	282	0.00	0.58	7	1,141
Gene_Product_Has_Organism_Source	1,904	7	0	0.00	0	0.00	0.41	0	780
Chemical_Or_Drug_Has_Physiologic_Effect	1,818	7	7	1.00	1,818	0.59	1.00	1,073	1,818
EO_Disease_Maps_To_Human_Disease	1,811	7	7	1.00	1,811	0.59	1.00	1,069	1,811
Gene_Associated_With_Disease	1,581	6	3	0.50	791	0.12	0.88	187	1,394
Gene_Product_Has_Structural_Domain_Or_Motif	1,329	6	0	0.00	0	0.00	0.46	0	610
Chemical_Or_Drug_Has_Mechanism_Of_Action	1,094	6	6	1.00	1,094	0.54	1.00	592	1,094
Gene_Product_Malfunction_Associated_With_Disease	1,049	6	6	1.00	1,049	0.54	1.00	567	1,049
OTHER RELATIONS	6,494	163	67	0.41	2,669	0.34	0.49	2,197	3,168
SUM	**182,300**	**354**	**176**		**121,091**			**76,031**	**145,455**

NCIT relation type	# occurrences in OWL "someValues From" clause	sam-ple size	# errors in sample	sample error rate
Disease_May_Have_Finding	27,652	9	9	1.00
Disease_May_Have_Cytogenetic_Abnormality	18,860	9	9	1.00
Gene_Product_Plays_Role_In_Biological_Process	15,607	8	8	1.00
Gene_Plays_Role_In_Process	14,385	8	8	1.00
Chemotherapy_Regimen_Has_Component	10,861	8	0	0.00
Gene_Product_Encoded_By_Gene	10,754	8	0	0.00
Disease_May_Have_Molecular_Abnormality	10,687	8	7	0.88
Gene_Is_Element_In_Pathway	8,364	8	8	1.00
Gene_Product_Is_Element_In_Pathway	8,302	8	8	1.00

Gene_Product_Has_Associated_Anatomy	1,972	7	1	0.14	282	0.00	0.58	7	1,141
Gene_Product_Has_Organism_Source	1,904	7	0	0.00	0	0.00	0.41	0	780
Chemical_Or_Drug_Has_Physiologic_Effect	1,818	7	7	1.00	1,818	0.59	1.00	1,073	1,818
EO_Disease_Maps_To_Human_Disease	1,811	7	7	1.00	1,811	0.59	1.00	1,069	1,811
Gene_Associated_With_Disease	1,581	6	3	0.50	791	0.12	0.88	187	1,394
Gene_Product_Has_Structural_Domain_Or_Motif	1,329	6	0	0.00	0	0.00	0.46	0	610
Chemical_Or_Drug_Has_Mechanism_Of_Action	1,094	6	6	1.00	1,094	0.54	1.00	592	1,094
Gene_Product_Malfunction_Associated_With_Disease	1,049	6	6	1.00	1,049	0.54	1.00	567	1,049
OTHER RELATIONS	6,494	163	67	0.41	2,669	0.34	0.49	2,197	3,168
SUM	**182,300**	**354**	**176**		**121,091**			**76,031**	**145,455**

Results

- Very high rate of ontologically inadequate axioms:
  Half of the sample: n = 176 rated as inadequate
  Estimation 0.5 [0.42 – 0.80] 95%
- inter-rater agreement (Cohen's Kappa):
  0.75 [0.68 – 0.82] 95%
- Typical inadequate statements
  1. relations including "may" (**disease_may_have_finding**)
  2. relations including "role" (**gene_product_plays_role_in_process**)
  3. inverse dependencies (e.g. parts on wholes)
  4. distributive assertions formulated as conjunctions

Why are they rated false?

- *Ureter_Small_Cell_Carcinoma* subclassOf
  Disease_May_Have_Finding some *Pain*
- in plain English: For every member of the class *Ureter_Small_Cell_Carcinoma* there is a relation to at least one member of the class *Pain* (regardless of the nature of the relation)
- Let us abstract the relation **Disease_May_Have_Finding** to the parent relation **Associated_With** (the top of the relation hierarchy):
- With *Ureter_Small_Cell_Carcinoma* subclassOf *Carcinoma*, a query for painless cancer: *Carcinoma* and not **Associated_With** some *Pain* will not retrieve any disease case classified as *Ureter_Small_Cell_Carcinoma*
- A DSS using NCIT-OWL + reasoner could then fatally infer that the absence of pain rules out the diagnosis *Ureter_Small_Cell_Carcinoma*

What is the basic problem?

- Mismatch between
  - the intended meaning of a relation, here the notion of “may” in **Disease_May_Have_Finding**
  - the set-theoretic interpretation of the quantifier “some” in Description Logics
- Problem: DLs have no in-built operator for expressing possibility
- Solution (Workaround ?): dispositions with value restrictions:

Ureter_Small_Cell_Carcinoma subclassOf
Bearer_of some (*Disposition* and
Has_Realization only *Pain*)

Other errors and possible solutions (I)

- Antibody_Producing_Cell *subclassOf*
  Part_Of *some* Lymphoid_Tissue
- Problem: Cells produce antibodies also outside the lymphoid tissue
- Solution: Inversion:
  Lymphoid_Tissue *subclassOf*
  Has_Part *some* Antibody_Producing_Cell

(which is NOT the same as the above axiom)

Other errors and possible solutions (II)

- *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*
- Problem: False encoding of distributive statements
  (a single molecule cannot be located in disjoint locations)
- Solution (but probably not complete…):
  Calcium-Activated_Chloride_Channel-2 subClassOf
  Gene_Product_Expressed_In_Tissue only
  (*Lung_Structure* or
  Mammary_Gland _Structure or
  Trachea_Structure)

Discussion

- Obviously, NCIT-OWL – if strictly interpreted according OWL semantics, abounds of errors
- NCIT curators: *"much more (…) a 'working terminology' than as a pure ontology"*

  de Coronado S et al. The NCI Thesaurus Quality Assurance Life Cycle. Journal of Biomedical Informatics 2009 Jan 22.
- But then why is it disseminated in OWL?
- If interpreted according to OWL semantics, systems using logical inference on NCIT axioms might become unreliable

Conclusion (beyond NCIT)

- Main problem of thesaurus ontologization:
  term / concept representation → reality representation
- Consequences
  - labor-intensive if done manually
  - error-prone if done automatically
- Recommendations
  - don't "OWLize" a thesaurus it if there is no clear use case
  - use other Semantic Web standard, e.g. SKOS
  - in case there is a good reason for transforming to a formal ontology,
    - use a principled ontology engineering approach
    - use categories and relations from an upper-level ontology
    - invest in quality assurance measures

Thanks

Schulz et al.: The Pitfalls of Thesaurus Ontologization - the Case of the NCI Thesaurus

- Contact: steschu@gmail.com

- Funding:
  EC project “DebugIT” (FP7-217139)
- Thanks to reviewers who provided high quality and detailed recommendations